RECONSTRUCTING THE MIND: REPLICABILITY IN RESEARCH ON HUMAN DEVELOPMENT

RECONSTRUCTING THE MIND: REPLICABILITY IN RESEARCH ON HUMAN DEVELOPMENT

by René van der Veer,
Marinus van IJzendoorn,
and Jaan Valsiner (Eds.)

ABLEX PUBLISHING CORPORATION
NORWOOD, NEW JERSEY

Copyright © 1994 by Ablex Publishing Corporation

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, microfilming, recording, or otherwise, without permission of the publisher.

Printed in the United States of America.

Library of Congress Cataloging-in-Publication Data

Reconstructing the mind : replicability in research on human development / by René van der Veer, Marinus van IJzendoorn, and Jaan Valsiner (eds.).
p. cm.
Includes bibliographical references and index.
ISBN 0-89391-871-7 (cloth)
1. Developmental psychology—Research—Methodology.
2. Replication (Experimental design) I. Veer, René van der, 1952– . II. IJzendoorn, Marinus H. van, 1952– . III. Valsiner, Jaan.
BF713.R434 1994
155'.072—dc20 94-8872
CIP

Ablex Publishing Corporation
355 Chestnut Street
Norwood, New Jersey 07648

Contents

Preface		vii
General introduction		1
Part I. The notion of replicability: Historical and theoretical issues		**11**
Chapter 1.	The varieties of replication: A historical introduction *Kurt Danziger & P. Shermer*	17
Chapter 2.	Replication, restudy, and the nature of anthropological fieldwork *Peter Kloos*	37
Chapter 3.	A process model of replication studies: On the relation between different types of replication *Marinus H. van IJzendoorn*	57
Chapter 4.	Constructivism and pragmatism: How to solve the problematic relation between methodology and epistemology in the debate about replication *Siebren Miedema & Gert J.J. Biesta*	71
Part II. The methods of replication		**93**
Chapter 5.	Do our methods work for or against us? *Donald W. Fiske*	97
Chapter 6.	Replication and meta-analysis: The story of a meta-analysis of parents' socialization practices *Hugh Lytton*	117

Chapter 7. From virtual to systematic replications 151
Jurg Siegfried

Part III. Replication in practice: The interdependence of theories and empirical context 171

Chapter 8. Replicability in context: The problem of generalization 173
Jaan Valsiner
Chapter 9. Situational context features and early memory development: Insights from replications of Istomina's experiment 183
Wolfgang Schneider & Marcus Hasselhorn
Chapter 10. Replication: What's the angle? 207
Trisha H. Folds-Bennett
Chapter 11. The forbidden colors game: An argument in favor of internalization? 233
René van der Veer
Chapter 12. Does maternal responsiveness increase infant crying? 255
Frans O.A. Hubbard & Marinus H. van IJzendoorn

Epilogue 271

Author Index 285

Subject Index 293

Preface

This book concentrates on issues of replicability in research on human development from an international perspective. Psychologists, anthropologists, educators, and methodologists from North America and Europe reflect on their experiences with different kinds of replication studies. The developmental perspective emphasizes change instead of stability and, consequently, calls into doubt the traditional idea of replication as an exact repetition of process and outcome of previous studies. In research on human development replication studies may nevertheless play a crucial role in stimulating scientific discourse about the cultural and historical context of human behavior.

Preparation of this book was made possible by a Pioneer grant of the Netherlands Foundation of Scientific Research (NWO) to Marinus H. van IJzendoorn and by a traveling grant of the same foundation to Jaan Valsiner. We are also indebted to the Department of Education (Vakgroep Algemene Pedagogiek) of the University of Leiden for its continuous support of this project.

René van der Veer
Marinus H. van IJzendoorn
Jaan Valsiner
Leiden and Chapel Hill,
December 1990

General Introduction

In Milan Kundera's *The Unbearable Lightness of Being* the author reflects on the idea of eternal return and concludes that the normal, transitory nature of things constitutes a mitigating circumstance. We cannot condemn things that are ephemeral, in transit. In the sunset of dissolution everything is illuminated by the aura of nostalgia, even the guillotine. Kundera concludes that our attitude toward things past reveals the profound moral perversity of a world that rests essentially on the nonexistence of return, for in this world everything is pardoned in advance and therefore everything cynically permitted.

Kundera's analysis reflects a basically Western understanding of time as a linear process that admits of no exact repetitions (Gale, 1968). Time is thought of as an arrow rather than a boomerang. However, even accepting the noncircular nature of time and the nonexistence of exact return that follows from it, most people would disagree with the pessimistic conclusion that, consequently, everything is permitted. Scientists, in particular, have always had problems with the idea that "anything goes" (Feyerabend, 1975, 1978) and have formulated all sorts of methodological rules for decision making. Paradoxically, they have advanced the possibility of return or replication as the test of reliability or validity of statements and decisions.

It seems then that we are faced with an unsolvable problem: Accepting the transient nature of things leads to the impossibility of exact replications and, therefore, to the impossibility of establishing the reliability (or, accepting Kundera's argument, the morality) of decisions and results. This would lead to the relativistic position that the adequacy or morality of conclusions cannot be adequately judged by persons living in another historical period. Similar arguments have been advanced with respect to the concept of culture

(Kohlberg, 1981). Many investigators would be unwilling to accept such a relativistic position. An obvious way out is, of course, to simply assume that time is irrelevant for the subject of study and this is, indeed, an assumption often made in the natural and social sciences. This assumption may be warranted for the natural sciences—after all, the nature of nuclear particles does not seem to be dependent on the historical period in which they are studied—but it is very problematic for the social sciences and humanities. There, a more adequate way out of the problem might be to accept the basic developmental and transhistorical (Larsen, 1990) nature of things and to study them exactly in the process of development (Valsiner, 1987, 1989).

CULTURAL BASIS FOR THE NEED OF REPLICATION

The idea that decisions or results should be repeatable is, of course, very generally accepted in our culture. Convicts bring an appeal against the judgment in the hope that another court and another jury will come to a different decision. When the judgment remains the same our belief in the justness of the decision is strengthened. Apparently, we conclude, the judgment was not merely based on the personal opinions of this specific court and this specific jury. To some extent it is instructive to compare the scientific enterprise with a lawsuit: Like scientists, judges and juries will try to weigh the evidence in a detached way and come to objectively valid and reliable—that is, repeatable—conclusions.

The basic grounds for bringing an appeal have to do with human fallibility. Judges and juries may make mistakes in interpreting the evidence. There are many reasons this could be so. They might fail to look at certain aspects of the material, selectively forget parts of it, be prejudiced against the defendant, make methodological mistakes, and so forth (Wagenaar, 1989). Virtually the same factors play a role in scientific reasoning and below some of them will be mentioned.

In several well-documented cases it is clear that the outcomes of classic scientific investigations were erroneous because of sloppy methodological procedures or even fraudulent behavior. One example is that of the well-known psychological experiment on conditioned emotional reactions reported by Watson and Rayner (1920). The authors' experiment was an attempt to condition an emotional reaction to a specific stimulus, and then to see whether that emotion would reappear in the presence of other similar stimuli and persist through time. Their one and only subject was a little boy named Albert. As a baseline, Watson and Rayner established that Little Albert showed no fear when confronted with such stimuli as a white rat, a dog, a monkey, and burning newspapers. They then struck a steel bar behind Little Albert each time the child touched a white rat. Not surprisingly, their subject

did not like this and after several repetitions he became fearful of the rat. The experimenters then established that this fear generalized to other similar stimuli such as a fur coat, a rabbit, and Watson's hair. They concluded that the emotional response did generalize to similar stimuli and that the effect persisted for a longer period than a month. Unfortunately, later researchers (Harris, 1979; Hilgard & Marquis, 1940) were unable to replicate Watson's findings, and it has been pointed out (Cornwell & Hobbs, 1976; Paul & Blumenthal, 1989; Prytula, Oster, & Davis, 1977; Samelson, 1980) that Watson's study was methodologically flawed. Among other things, Watson and Rayner had to "freshen" the conditioned reaction after 10 days, because Albert's fear of the rat began to fade. Also, emotional reactions typically did not occur unless the experimenters removed Albert's thumb from his mouth (Paul & Blumenthal, 1989).

Another (in)famous case is that of Sir Cyril Burt's intelligence research. It has now been well documented (Gould, 1981; Hearnshaw, 1979; Kamin, 1974; Lewontin, Rose, & Kamin, 1984) that the evidence for Burt's claim that IQ is genetically determined was nonexistent. The collapse of Burt's claims within the scientific community began when attention was drawn to some numerical implausibilities in his published papers. In three consecutive studies of separated identical twins with differing sample size ($n = 21$; $n = 30$; $n = 53$) Burt found a IQ correlation of .771. The IQ correlations for other relatives also remained identical to the third decimal place in these studies. This wholly unlikely result was first observed by Kamin (1974) and led him and other researchers to make a careful study of Burt's alleged experiments. It is now established beyond reasonable doubt that Burt's *idée fixe* of the innateness of intelligence led him to fake his results. The irony of this case of scientific fraud will be clear: Fellow researchers started doubting Burt's data for the very reason that he reported strict replicability of results in three consecutive investigations.

These examples of classic experimental findings may not be typical for the whole field of psychology, but they do show that at least several of them do not stand up to careful scrutiny (see also Chapter 3). They also demonstrate that important research findings may serve a sociological, ideological, and even psychological function for the persons concerned. This circumstance—whether it is seen as a factor contaminating research outcomes that can be avoided or as an inevitable aspect of doing scientific research—may clearly hinder or prevent the replication of results. This leads us to the factors that might actually make replication of research outcomes difficult.

DIFFERENT KINDS OF REPLICATION

Let us first remark that researchers may strive for replication studies that copy the experimental conditions of the original study as faithfully as possible

or for studies in which one or more of the experimental conditions are deliberately changed. In the first case we may speak of strict, precise, (virtually) exact, or carbon-copy replications. It is clear that strict replications can never exactly copy the original study, if only for the reason mentioned by Kundera. In the second case we may speak of conceptual, imprecise, or inexact replications. It has been remarked by various researchers that the two types of replication studies serve a different goal: Strict replications bolster confidence in the original data by confirming their reliability (Amir & Sharon, 1990; Hendrick, 1990; Rosenthal, 1990); conceptual replications yield novelty by showing that the original results can be found under other circumstances and, thus, demonstrate their external validity or generalizability. Recently researchers have suggested implementing strict and conceptual replications in the design of the original study as a matter of routine. A "multitrait–multimethod" model, in which different potentially important factors are systematically varied, is proposed (Hendrick, 1990; Rosenthal, 1990).

In order to see what is exactly meant by exact or conceptual replications it is useful to distinguish several of the factors that may actually influence the possibility of replicating research outcomes. Hendrick (1990; see also Amir & Sharon, 1990, for a similar list) has listed the following classes of factors that affect the total research reality:

1. **Subject characteristics.** *Such variables as age, sex, and so on may* vary from study to study and affect the results of a given research project. It is well known, for example, that most textbook findings rest on research carried out under the population of American middle-class students and that, consequently, the story they tell may be highly biased. Replication of the research results in different samples may be problematic.
2. **Specific research histories of subjects.** The prior treatment of subjects may affect performance in a current study. This factor is crucial under the current practice of doing empirical research where most psychology students are asked to participate in research more than once and may develop all sorts of expectations concerning the desired research outcome. Generally speaking this factor may undermine the common assumption of stability of the phenomenon under study;
3. **Social-historical-cultural context.** No doubt, psychology is a highly ethnocentric affair (Valsiner, 1989) and the fact that subjects in different cultures and historical periods will react differently to the "same" stimuli makes psychology's findings as presented in textbooks of limited value.

4. **General physical setting of the research.** This class includes such factors as lighting, ventilation, acoustics, and so on, that may affect the mood of subjects.
5. **Control agent.** The person running an experiment interacts with the subjects in a unique way that may affect the behavior of subjects. In addition he or she may have certain expectations as to the outcomes. Rosenthal (1966) has convincingly shown how these experimenter effects may play a contaminating role.
6. **Specific task variables.** This class refers to specific minute material circumstances such as the nature of research booklets, the style of typing, and so on. They are generally assumed to be theoretically irrelevant, but this remains to be seen (Hendrick, 1990).
7. **Primary information focus.** This refers to the set of instructions given to the subject that should guarantee that the subjects construe the research reality as intended by the experimenter. In practice, of course, subjects may define or structure the situation in a way that is guided by personal, cultural, and historical factors.
8. **Modes of data reduction and presentation.** Different means of analyzing and presenting the data will lead to different conclusions as to what the data show. Recently advanced methods, such as meta-analysis, may to a limited extent be helpful in reducing this source of variation.

This list of factors that may vary across empirical investigations once again shows that strict replication is impossible in principle. For even in the improbable case that the same researcher would carry out the same study with the same material and the same subjects in the same circumstances, factors such as the subject characteristics and the sociocultural-historical context might have changed in a relevant way. Really strict replication is not even possible in the case where one would arbitrarily divide the sample in two equal parts and then compare the results of the analysis computed for each half separately. Apart from the fact that such a borderline case of "split-half replicability" stretches our understanding of the notion of replicability rather too far, it is clear that the subjects' characteristics or their research history would differ in potentially relevant ways. It is clear, then, that the idea of strict replication should be thought of as a regulative idea and that strict and conceptual replications do not form a dichotomy. However, these two extremes on the similarity continuum do serve a somewhat different goal and it is significant, for instance, that Rosenthal (1990)—arguing in favor of conceptual replication—has recently suggested that a replication of an experiment that obtains similar results is maximally convincing if it is maximally seperated from the first experiment along such dimensions as

time, personal attributes of the experimenters, experimenters' degree of personal contact with each other, and so forth.

REPLICABILITY AND PUBLISHABILITY

It would seem that the fact that very few replication studies are published in social scientific journals partly reflects an understanding of replication that is biased toward strict replication. As these replication studies do not yield novelty, but rather check the reliability of the original results, they are less valued in a community where (limited) originality is highly valued. Rather surprisingly, in natural scientific journals the number of replication studies published is much higher (Bornstein, 1990b). We surmise, however—relying on the sociological literature on this issue (Chapter 3)—that these are almost exclusively conceptual replications. That social scientific journals discourage replication studies has been observed more than once (e.g., Bornstein, 1990a; Mahoney, 1985). There is now growing empirical evidence for this suggestion. First, the number of replication studies published in social scientific journals has been found to be consistently low in several studies. In fact, in none of the studies does the percentage of published replication experiments reached 1% of the sample size (Bozarth & Roberts, 1972; Greenwald, 1975; Reid, Soley, & Wimmer, 1981). This, of course, may reflect the fact that researchers simply do not conduct replication studies regardless of editorial policies. But this seems to us an improbable explanation of the low publication numbers. First, informal observation suggests that many students are required to carry out a replication study as part of their training, or "rite of passage." This would imply that many more replication studies are actually carried out than is generally thought and reflected by the publication numbers. The fact that *The Journal of Social Behavior and Personality* (also published as the *Handbook of Replication Research in the Behavioral and Social Sciences*) recently gathered some 50 replication studies by sending out notices to academic departments indeed suggests that there may be quite a few replications available (Neuliep & Crandall, 1990). We surmise, therefore, that despite discouraging editorial policies replication studies are still carried out on a regular basis. Recently, Neuliep and Crandall (1990) found strong evidence in favor of the hypothesis of editorial bias. The authors mailed a questionnaire concerning the issue of replication to a sample of editors. The great majority of editors (72%) found replications less important than studies that demonstrate a new effect. Rather surprisingly, a majority of the editors felt that the burden of proof lies with the replicator when a replication study fails to replicate an earlier finding. This assumed asymmetry of responsibilities clearly favors the publication of new results and discourages replication studies.

If leading journals discourage replication studies and researchers continue doing them then a "file drawer problem" will inevitably develop. It would imply that the studies published in the social sciences are a biased sample of the studies that are actually carried out. For if journals discourage the publication of nonsignificant results—and replication studies that fail to replicate earlier findings are often of the nonsignificant type—then the file drawers in the laboratory will be gradually filled with studies showing nonsignificant results. This would mean that journals are simply filled with studies showing Type I errors. Rosenthal (1990) has recently suggested some quantitative measures to cope with this problem.

Several researchers have suggested that the exclusive concentration on significance scores is in itself misleading. Moreover, researchers tend to think of the success of a replication study in a dichotomous fashion (e.g., Jegerski, 1990; Rosenthal, 1990). Rosenthal (1990) has argued that it is far more informative to concentrate on the effect size as the relevant summary statistic of studies. In comparing original study and replication the focus should be on effect size and the evaluation of whether the replication has been successful should be made in a continuous fashion. Rosenthal (1990) has suggested several metrics of the success of replication. These technical recommendations all rest on the assumption that the variables and outcomes of different studies can indeed be meaningfully compared. Some of the contributions in this book seem to cast doubt on this assumption (see Chapters 1 and 2, respectively). It seems that, although such metrics may be very useful, the researcher comparing the results of various studies—as in meta-analysis—cannot avoid conceptual issues; nor can he or she avoid many difficult decisions that cannot be based on purely quantative grounds (see also Chapter 6).

Earlier we saw that published replication studies are far more frequent in natural scientific journals than in social scientific ones. Does this circumstance perhaps reflect some fundamental differences between the natural and social sciences? One thing that has been suggested (Chapter 6) is that the description of materials and procedures can be more accurate in the natural sciences as they developed a generally accepted vocabulary. As a result replicators in the natural sciences would have less difficulty replicating the original results. Although it remains to be seen whether the situation in the natural sciences is indeed much better—such objective factors as lack of space and the wish to conceal novel procedures may lead researchers to be rather brief in the description of their material and procedures—it is true that social studies are often difficult to replicate on the sole basis of the description found in the journal publication. Amir and Sharon (1990), who carried out six replication studies in social psychology, report that "in most cases the information published in the articles was not sufficient for an adequate replication. At times, even additional information, directly supplied by the original re-

searcher, was not detailed enough to ensure a precise replication." Likewise, Van IJzendoorn (Chapter 3) observes that the use of some research techniques—such as the adult attachment interview—requires sometimes extensive training under the supervision of the original experimenter. One wonders what in the light of these circumstances one is to make of Rosenthal's claim that replications are less convincing the more the original experiment is similar to the replication along such dimensions as experimenters' degree of personal contact with each other, and so forth (see earlier).

Another factor that may make replication studies more difficult in the social sciences is the lack of a standardized vocabulary. As Lytton (Chapter 6) points out, a mother giving a young child help with a cognitive task can be viewed as showing support or warmth, but she can also be thought of as intrusive. Such different interpretations and labels of behavior are guided by the researcher's cultural and historical context.

Rather than giving an exhaustive list of all the factors that might make studies in the social sciences more difficult to replicate than those in the natural sciences we refer back to the classes of potentially relevant variables given before (Hendrick, 1990). It is clear that such classes as subject characteristics, specific research histories of subjects, social-historical context, specific task variables, and primary information focus do not play a role in the natural sciences. Part of these differences are perhaps best expressed by saying that, whereas natural scientists work on the assumption of the uniformity of nature, social scientists also and inevitably have to face the variety of culture. Cross-cultural psychology and anthropology has made us aware of the fact that universals in human nature and culture are hard to find. This is also true for social sciences that are generally thought to be more objective than psychology, such as medicine. The views of what constitutes a disease and what are its possible cures vary immensely even across Western cultures (Payer, 1990). Neither do we easily find behaviors nor cultural constructs that are not subject to historical change.

We may conclude, then, that there are objective factors that make exact and conceptual replications less feasible in the social sciences. It is, therefore, part of the goal of this book to indicate the goals, usefulness, but also the boundaries of replication in the social sciences—boundaries that are determined by the fundamental developmental nature of human phenomena and that were described by Kundera in his reflection on the nonexistence of return.

REFERENCES

Amir, Y., & Sharon, I. (1990). Replication research: A "must" for the scientific advancement of psychology. *Journal of Social Behavior and Personality, 5* (4), 51–70.

Bornstein, R. F. (1990a). Manuscript review in psychology: An alternative model. *American Psychologist, 45*, 672–673.

Bornstein, R. F. (1990b). Publication politics, experimenter bias and the replication process in social science research. *Journal of Social Behavior and Personality, 5* (4), 71–81.

Bozarth, J. D., & Roberts, R. R. (1972). Signifying significant significance. *American Psychologist, 27*, 774–775.

Cornwell, D., & Hobbs, S. (1976, March 18). The strange saga of Little Albert. *New Society*, pp. 602–604.

Feyerabend, P. (1975). *Against Method.* London: NLB.

Feyerabend, P. (1978). *Der wissenschaftstheoretische Realismus und die Autorität der Wissenschaften.* Braunschweig: Vieweg.

Gale, R. M. (Ed.). (1968). *The philosophy of time.* New Jersey: Humanities Press.

Gould, S. J. (1981). *The Mismeasure of Man.* Harmondsworth, UK: Penguin.

Greenwald, A. G. (1975). Consequences of prejudice against the null hypothesis. *Psychological Bulletin, 82*, 1–20.

Harris, B. (1979). What ever happened to little Albert? *American Psychologist, 34*, 151–160.

Hearnshaw, L. S. (1979). *Cyril Burt: Psychologist.* London: Hodder & Stoughton.

Hendrick, C. (1990). Replications, strict replications, and conceptual replications. *Journal of Social Behavior and Personality, 5* (4), 41–49.

Hilgard, E. R., & Marquis, D. G. (1940). *Conditioning and Learning.* New York: Appleton-Century.

Jegerski, J. A. (1990). Replication in behavioral research. *Journal of Social Behavior and Personality, 5* (4), 37–39.

Kamin, L. (1974). *The Science and Politics of IQ.* Potomac, MD: Erlbaum.

Kohlberg, L. (1981). *The philosophy of moral development.* San Francisco: Harper & Row.

Larsen, K. S. (1990). The Asch conformity experiment: Replication and transhistorical comparison. *Journal of Social Behavior and Personality, 5* (4), 163–168.

Lewontin, R. C., Rose, S., & Kamin, L. J. (1984). *Not in our genes.* New York: Pantheon Books.

Mahoney, M. J. (1985). Open exchange and the epistemic process. *American Psychologist, 40*, 20–39.

Neuliep, J. W., & Crandall, R. (1990). Editorial bias against replication research. *Journal of Social Behavior and Personality, 5* (4), 85–90.

Paul, D. B., & Blumenthal, A. L. (1989). On the trail of Little Albert. *The Psychological Record, 39*, 547–553.

Payer, L. (1990). *Medicine and Culture.* London: Victor Gollancz Ltd.

Prytula, R. E., Oster, G. D., & Davis, S. F. (1977). The "rat-rabbit" problem: What did John B. Watson really do? *Teaching of Psychology, 4*, 44–46.

Reid, L. N., Soley, L. C., & Wimmer, R. D. (1981). Replication in advertising research. *Journal of Advertising, 10*, 3–13.

Rosenthal, R. (1966). *Experimenter effects in behavioral research.* New York: Appleton-Century Crofts.

Rosenthal, R. (1990). Replication in behavioral research. *Journal of Social Behavior and Personality, 5* (4), 1–30.

Samelson, F. (1980). J. B. Watson's Little Albert, Cyril Burt's twins and the need for a critical science. *American Psychologist, 35*, 619–625.

Valsiner, J. (1987). *Culture and the development of children's action.* Chichester, UK: John Wiley & Sons.

Valsiner, J. (1989). *Human development and culture. The social nature of personality and its study.* Lexington, MA: D.C. Health.

Wagenaar, W. A. (1989). *Identifying Ivan: A case study in legal psychology.* Brighton: Harvester Press.

Watson, J. B., & Rayner, R. (1920). Conditioned emotional reactions. *Journal of Experimental Psychology, 3*, 1–14.

PART I

The Notion of Replicability: Historical and Theoretical Issues

The philosophy of science is replete with methodological precepts that were ostensibly distilled from actual scientific practice. Analyzing the work of successful natural scientists, philosophers of science have formulated rules that supposedly brook no appeal. However, it has become clear that it is exceedingly easy to garner indubitable specimens of successful—even famous—investigations in various strands of science that explicitly violate the methodological canon of philosophers. The quirks and vagaries of actual scientific practice apparently ensure a picture that is incomparably richer than simple methodological rules would suggest. This is not to gainsay that human curiosity is somehow rule bound, or that more and less fruitful approaches in scientific research exist. But we do surmise that the latitudes of scientific practice are broader than is commonly thought and that many a methodological precept has to be thought over and tested against actual scientific practice in the type of sociological research that several of the first chapters of this section refer to.

One of the most well-known rules of scientific methodology states that researchers should endeavor after repeatable or replicable experiments and results. The first chapters of this book deal with questions concerning the nature and function of replication. What do we mean by the word *replication*? Can different forms of replication be distinguished? Does successful replication of experimental results yield an argument in favor of their truth? Do replication studies actually occur with any frequency in scientific practice? Is replication theoretically possible and fruitful? These are some of the questions of the theory of replication that the first chapters address.

Before we give a succinct overview of the content of these chapters and

their interrelatedness we do well to distinguish between various uses of the word *replication* as they occur in these first chapters. By replication we may mean the replication of the *setup* or experimental conditions of an investigation. Such a type of replication is made possible by providing as detailed as feasible a description of the circumstances under which the original experiment took place, the role of the experimenter, the exact text of the instruction or questions, and so on. This use of the word *replication* is probably the most prevalent. One may also speak of replicating the *results* of the original experiment found either with an identical experimental setup or a different one. In this sense of the word replication comes close to concepts like *confirmation* or *corroboration*. Finally, one may argue that the *conclusions*—the interpretations of experimental results or research outcomes—to which an investigation leads replicate conclusions reached earlier. Thus, to give a perverse example, it is possible to invent an experimental setup that makes subjects nervous (e.g., playing very loud music while asking the subjects to perform some extremely boring tasks). The details of this experimental setup—if properly described—can be copied (replicated) with reasonable success. Likewise, in a second study one can hope to replicate the original experimental results, say, that male subjects are generally more nervous than female ones, with the same setup, or another one. Thus, another researcher might copy the original experimental setup or, for example, ask subjects to give a public talk and find similar or different results. Finally, when different or virtually identical setups have led to different or virtually identical results the scientific community will tend to (dis)agree about the interpretation of these results. It might be argued, for example, that stress in the first experimental setup is caused by underlying mechanisms that are very different from those operative in the second experimental setup. Quantitatively identical experimental results may hide very different underlying mechanisms, and researchers consequently might argue that—theoretically speaking—the two experiments cannot be meaningfully compared and that, therefore, the results of the second study in no way replicate the results of the first. This somewhat far-fetched example brings out some of the problems (another problem might be whether one type of "nervous" behavior—e.g., picking one's clothes—confirms or replicates another type of behavior—e.g., excessive sweating, or outbursts of anger) of conceptualizing replication and may help to understand part of the discussion in the following chapters.

In Chapter 1, Danziger and Shermer defend a historical view of the phenomenon of replication in psychology arguing that there are no universally valid and ahistorical rules about what constitutes proper replication. Whereas the founder of psychology Wundt, for example, meant by replication the observation of identical phenomena under identical circumstances, for the Wurzbürgians it meant the observation of common features under different circumstances. Whereas Baldwin selected naive adults, or children

as subjects, Titchener worked with well-trained, normal adults. What one defines as the proper subject matter of psychology determines one's concept of replication. The authors describe the shift from replication with respect to individual subject properties to population properties and the distinction between true and accidental variation that went with it. In the modern view replication has come to mean the reproducibility of particular statistical patterns among the scores of a group of individuals. Danziger and Shermer conclude their chapter by observing that, despite the disagreements as to the nature of true replications between different schools in psychology, there is still unanimous agreement among the schools about two fundamental assumptions. First, it is generally thought possible to eliminate history, that is, what psychology studies are thought to be processes and functions not subject to historical change. Second, it is assumed that the methods of investigation are independent from whatever they were supposed to investigate.

It is exactly these assumptions that Kloos challenges in Chapter 2 on restudy in anthropological research. By its very nature anthropology has to do with the "soft" aspects of human culture that tend to vary much more than the "hard" quasiphysiological properties that some psychologists claim to investigate. That is, while some psychologists might identify with the anatomist who opens one or two bodies and then sees no need of further replicating his experiment (see Chapter 1 for this example), this is unacceptable for anthropologists who have learned that societies tend to be in a continual flux and that the instrument itself (that is, the investigator) is necessarily a co-constructor of the cultural reality he encounters. No two interviewers will sketch the same picture of a person and neither will the same person give the same answers when interviewed in different periods of his life. Likewise, no two anthropologists will give the same account of a society as no two anthropologists will investigate the same (aspect of the) society for the reasons mentioned. The sex of the anatomist will not determine the results of an obduction, whereas the sex of the anthropologist may be crucial. These circumstances have forced anthropologists—perhaps more so than psychologists—to appreciate the fact that the knower in anthropology is changing his subject matter by his knowing activities. Thus, anthropological research has led to a more dialectical view of what it is to gather knowledge about a phenomenon. Accepting this situation as an unavoidable fact Kloos argues that understanding is best furthered by the reflective analysis of contrasting accounts of one and the "same" society (as exact or even approximate replications are impossible anyhow). He suggests even to make the most of the situation and promote different views by asking researchers with different theoretical perspectives to give their views of a society in order to learn from the contrasts.

Drawing on the growing sociological literature about the practice of replication in actual scientific research Van IJzendoorn in Chapter 3 proposes a

process model of replication studies. In this view virtually exact replications are rare in the natural sciences and are only resorted to when systematic variation of experimental setups or less costly variants of replication studies do not yield the theoretically expected results. The author proposes a hierarchical model in which different forms of replication studies all find their place. Van IJzendoorn's emphasis on the role of the scientific forum in judging the value of deliberate variations of experimental setups and attempts at exact replication is somewhat reminiscent of Kloos's view. Both authors seem to encourage nonexact replications (in Kloos's case preferably nonexact replications by researchers with a different theoretical outlook) whose results will be evaluated by the scientific community. The difference seems to be that Van IJzendoorn finds a limited role for virtually exact replication studies as a last resort, whereas Kloos thinks that these are impossible in principle and useless as he sees great value in contrasting views. At the background of both views, however, there seems to be the fundamental conviction that we learn by extending our cognitive schemas and not by literally repeating them. Simple isomorphic copies of research designs seem indeed logically impossible and theoretically of little value. Instead, in their search for knowledge, both researchers and children seem to follow the epistemological pattern of nonisomorphic imitation or varied replication first described by Baldwin and Piaget (see Valsiner & van der Veer, 1988; van der Veer & Valsiner, 1988).

In Chapter 4, Miedema and Biesta address the thorny problem of whether successful replication of research findings demonstrates simply reliability or repeatibility of results, or whether it proves their ontological truth. Drawing on the same literature as Van IJzendoorn they first claim that exact replication is the exception rather than the rule. The topic of scientific debate seems not so much to be the reliability or repeatibility of results, but the validity or acceptability of the conclusions (research outcomes) seen against the accepted background knowledge. Even in the rare case of virtually exact replication, different researchers may disagree about the outcomes of this replication and consensus must be reached through a process of negotiation. Like Kloos, the authors emphasize the value of the scientific discussion about replication results and deny that replication can contribute to our knowledge of some independently given reality. Rather, with Dewey they view knowledge as a transactional process in which knowledge is the result of specific actions upon the objective environment and their feedback effects upon the subject. Knowledge, then, is a relationship that exists between the mutually dependent subject and object. This conclusion brings us nicely back to Danziger and Shermer's conclusion that the attempt to separate the conditions and the products of psychological research was never successful as knowledge is a relationship with an object and not a thing in itself.

These first theoretical chapters then will enrich our notion of the function, nature, and possibility of replication. A thicket of different interpretations of

what consitutes replication exists, and different research schools and disciplines adhere to different notions of replication. It seems impossible to conflate these various notions into one simple model. Although researchers agree with laymen that it is always useful to ask for a second opinion, they typically disagree about the way this should be done.

REFERENCES

Valsiner, J., & Van der Veer, R. (1988). On the social nature of human cognition. *Journal for the theory of the Behavioral Sciences, 18,* 117–135.

van der Veer, R., & Valsiner, J. C. (1988). Lev Vygotsky and Pierre Janet. On the origin of the concept of sociogenesis. *Developmental Review, 8,* 52–65.

CHAPTER 1

The Varieties of Replication: A Historical Introduction

Kurt Danziger and P. Shermer
Department of Psychology
York University
Ontario, Canada

> Generally, similarity, and with it repetition, always presuppose the adoption of a *point of view*: some similarities or repetitions will strike us if we are interested in one problem, and others if we are interested in another problem. But if similarity and repetition presuppose the adoption of a point of view, or an interest, or an expectation, it is logically necessary that points of view, or interests, or expectations, are logically prior, as well as temporally (or causally or psychologically) prior, to repetition. (Popper, 1959, pp. 421–422)

Although the requirement of replicability is undoubtedly a prominent part of the lore of science, it can be given an unambiguous and self-evident meaning only as long as we adopt a suitably restricted perspective. As soon as we go beyond the conventions that have come to prevail in a particular scientific community, we discover a variety of conceptions of what replication means in practice and why it is considered important. We also find a corresponding variety of actual practices that have been considered to satisfy the requirement of replicability at different times and in different scientific communities. An examination of this variety can provide valuable insights into the fundamental "points of view" or "interests" (to use Popper's terms) that are taken for granted in everyday scientific practice but that underlie its basic direction.

It is a curious paradox of scientific practice that, although everyone pays lip service to the ideal of replication, "no one ever repeats an experiment" (Hacking, 1983). That is not only because of the ancient problem of stepping in the same river twice, but also because experimenters, as Hacking and others have noted, are not passive observers of nature but creators of phenomena, and in particular, phenomena that are characterized by regu-

larity and stability. Successful replication helps to demonstrate the feasibility of this enterprise in specific cases. But replication can only play this role if there is agreement on the nature of the phenomena in question. Otherwise, one would not know whether to classify the phenomena created in the first and in the second experiment as "similar." Agreement on the nature of the phenomena, however, includes agreement on the relevance and significance of a multitude of factors that are associated with their creation. As no two experiments can ever be identical, an acceptable replication will be one characterized by an *acceptable degree of similarity* in the conditions under which phenomena are created. Obviously, what is acceptable to a particular group of investigators at one time may not be acceptable to another group of investigators at a different time.

Thus, there are no abstract and universally valid rules about what constitutes a proper replication. Such rules never have more than a local and historically limited significance. As scientists change their views about the nature of the phenomena they are dealing with, so their standards for acceptable replication change too. In well established fields there may be long periods of basic agreement about the nature of the phenomena, and then there may be little or no need for discussion about what represents a replication and what does not. But in newly emerging fields, like modern psychology in its earlier days, or in fields where scientists with different points of view try to cooperate, the question of what constitutes an adequate replication becomes part and parcel of the debate about what one is or should be studying. There is an inescapable social-historical context to the issue of replication. In recent times sociologists of science have been able to investigate the kind of "negotiation" that surrounds claims to have achieved a valid replication (Collins, 1985) as well as the social resolution of such negotiations.

In following the history of conflicting interpretations of the meaning of replication in a field like psychology we are precluded from using the interview method of the sociologist, but we have a wealth of published material to draw on. Some of this material appeared in the course of public discussion about what it meant to replicate findings in psychology. In other cases new concepts of replication simply made their appearance as a by-product of new points of view about the nature of the phenomena that psychology ought to be interested in. We first consider the major alternative conceptions that arose during the earlier history of modern psychology and then trace the emergence of conceptions that eventually came to dominate the discipline.

FIRST CONTROVERSY: WUNDT VS. THE WÜRZBURG SCHOOL

We have to begin with Wundt, not only because his views towered over the discipline in its earliest formative years, but also because they were the

product of systematic reflection about the nature of scientific method (Wundt, 1883). From the first years of his psychological laboratory Wundt took it for granted that replication was an essential part of scientific method. But for him replication meant the repetition of identical observations under identical conditions. Astronomical observation seems to have functioned as an exemplar. One has to repeat observations because they are subject to random sources of error that must be allowed for in order to arrive at an estimate of the "true" value one is interested in. Psychophysical threshold determinations illustrate this procedure.

Wundt's conception of what constituted a genuine repetition of an observation was extremely restrictive. This eventually brought him into conflict with a new generation of experimentalists who saw a much expanded role for introspection in the production of psychological phenomena. The most important representatives of the new trend were to be found in the Würzburg School; although others, notably Titchener, had a similarly optimistic view of the possibilities of introspective analysis (Danziger, 1980). In the Würzburg experiments introspective evidence was used to establish the nature of thought processes as well as volitional processes. That was anathema to Wundt, and he did not hesitate to condemn these "pseudo-experiments," as he called them (Wundt, 1907). Part of the ensuing controversy revolved around the nature of replication in psychology and provides a telling illustration of how different conceptions of replication are associated with different conceptions of the subject matter.

One of the main reasons Wundt rejected the findings of the systematic introspectionists was that they were not based on repeated observations under identical conditions. In psychophysical and perception experiments one could repeatedly present the same stimulus to the same experimental subject and obtain a distribution of responses that allowed for the elimination of random errors of observation. In the introspective observation of thought processes this was not possible, partly because of the temporally extended, flowing nature of these processes and partly because any attempt at repetition would be nullified by the appearance of memory effects. Accordingly, Wundt considered the scope of experimental psychology to be limited to phenomena that involved a close and immediate tie of subjective events to external stimuli, or at least to manifest external expressions. In practice, this virtually restricted the field to the study of sensation and perception, immediate memory, reaction times and the physiological expression of emotion. Subjective processes beyond these areas were not amenable to experimental investigation—they could not be transformed into the kinds of phenomena that natural science dealt with. Such processes would have to be studied indirectly, by examining their products, for example, language and myth, with methods borrowed from the *Geisteswissenschaften*, the human sciences.

The introspectionists whom Wundt attacked had a much more liberal view

of what constituted an experimentally established psychological phenomenon. Unlike Wundt, they were prepared to accept as data the content of retrospective introspective reports on conscious states that lacked any direct tie to stimulus events or overt responses. The stimulus settings that they used were in the nature of complex tasks (for a good English language account, see Humphrey, 1950), and not the measurable physical stimuli that Wundt preferred to work with. The phenomena they investigated were defined as thought processes rather than perceptual events. As long as one accepted the phenomena created in the Würzburg laboratory (and elsewhere) as legitimate objects of scientific investigation one could be content with a relatively loose fit between an experiment and its repetition and still feel one had achieved replication. Those whom Wundt criticized did not accept his criticism that their experiments were not replicable (Bühler, 1908). A number of subjects could be presented with a series of problem situations and certain common features in their introspective reports could be observed. Wundt's very narrow criteria of similarity were simply inappropriate for replicating observations on thought processes, a fact he seemed to recognize when he declared such investigations out of bounds for experimental psychology.

THE BALDWIN-TITCHENER CONTROVERSY

The Wundt–Bühler confrontation represents one of the first striking illustrations of the interdependence of conceptions of replication and definitions of legitimate objects of investigation in the history of psychology. But there were some fundamental beliefs about the proper subject matter of experimental psychology that Wundt and the Würzburgers shared. For instance, they all accepted that the subject matter to be investigated by psychology was mental events, even though they differed about the kinds of mental events that were legitimate candidates for such investigation. They also agreed that the mental events that were of primary interest to psychology were those commonly encountered in "normal," adult minds, not those in the minds of children, of disturbed people, or of "primitives." This was a legacy of academic psychology's historical tie to philosophy and the still strong belief, in its early European days, that it could justify its existence by its contribution to the solution of general philosophical problems. Opposed to this there was a growing tendency to legitimize psychological research by its potential contribution to practical problems. Those kinds of problems, however, presented themselves as the problems of particular subsections of humanity, not of humanity in general. If one wanted to contribute to an understanding of such relatively local problems one would have to adopt a different definition of the subject matter of psychology and hence a different conception of replication.

The first historical manifestation of this issue occurred very early, though it

took several decades for all the implications to become apparent. In 1895/96, E. B. Titchener and James Mark Baldwin, both soon to become major figures in the new science of psychology, engaged in a sharp controversy about the explanation of differences in reaction times (Baldwin, 1895, 1896; Titchener, 1895, 1896). As the controversy developed, a question of great interest for our topic of replication began to assume fundamental significance; it was the question of the selection of research subjects.

As we have noted, early laboratory psychology took the general features of conscious processes as its subject matter. That meant that its experimental subjects functioned in a dual role. On the one hand, their conscious processes provided the *objects* of research, but on the other hand, the same experimental subjects also functioned as the *observers* of these objects. The selection of such subjects therefore had to be undertaken in terms of two criteria. First of all, because attention was focused on the *general* features of conscious processes, one had to eliminate individuals in whom the relevant processes might be affected by mental or physical deficits or even by special kinds of background experience. Secondly, one had to select individuals whose observations could be relied on. Just as some individuals had no aptitude for making accurate observations of physical events, so there were some who had no aptitude for making accurate observations of mental events. These would have to be eliminated from the category of potential experimental subjects. Moreover, like other scientific observers, psychological observers would have to have some practice before their observations could be relied on. So naive, unpracticed individuals were not suitable as experimental subjects.

Once one had selected one's subjects on these criteria each individual going through the same experimental routine represented a replication of the experiment. The situation was in some ways analogous to the common practice in physiological experimentation or anatomical dissection. There too, if one was not a pathologist or embryologist, one's attention would be focused on general features to be found in normal, adult organisms. One would therefore eliminate diseased animals or those with detectable malformations and abnormalities. Because some abnormalities might not be readily apparent, and because of the chance that one's first animal might happen to be an abnormal specimen, one would do well to repeat one's dissection or physiological investigation on a few animals. But if one obtained consistent results in a few cases one would not consider it necessary to multiply the number of one's specimens unduly. In an analogous way experiments on the general features of conscious processes were usually conducted on a mere handful of subjects, though these had of course been selected on the criteria already mentioned. The investigators who engaged in this practice clearly regarded each additional subject exposed to constant experimental procedures as a replication of the original experiment. Sometimes such replications

were in fact dispensed with and the published experimental results were limited to a single subject (e.g., Mehner, 1885).

Among the phenomena that Wundt and some of his students had demonstrated by the use of such procedures was the *sensory motor difference*. In reaction time experiments there appeared to be a general tendency for response times to be relatively longer when subjects focused their attention on the sensory input than when they concentrated on the motor component. Working in North America, Baldwin failed to replicate this result, some of his subjects showing a longer response time when they concentrated on the motor side, others showing the longer sensory reaction times that had been reported from Wundt's laboratory.

Titchener, having recently obtained his doctorate under Wundt's supervision, proceeded to criticize what he clearly regarded as Baldwin's sloppy methods. Baldwin had used naive subjects with no experience of making psychological observations. Of course, their reports were scientifically worthless. Baldwin's study did not constitute an acceptable replication of the work carried out in Germany.

But Baldwin launched a counterattack in which he questioned the legitimacy of the subject selection procedures that were standard practice in Wundt's laboratory. If one could eliminate potential subjects until one had a group that gave consistent results one was begging the question of whether the phenomena were truly generalizable to the normal adult mind as such. Baldwin was inclined to think that sensory motor differences in reaction times were traceable to typological differences among individuals. Persons with a constitutional preference for motor imagery and memory would have shorter motor reaction times and persons with a constitutional preference for visual or auditory imagery and memory would have shorter sensory reaction times if those sensory modalities were used in the experiment.

What is important about this controversy is not the question of who had the right explanation for sensory motor differences in reaction times. In fact, Angell and Moore (1896) soon showed that there might be some truth in both the universalistic and the typological explanation. What the controversy reveals about the role of replication in different kinds of scientific practice is far more significant.

As more recent commentators (Böhme, 1977; Krantz, 1969; Schultz, 1970) have noted, the Baldwin–Titchener controversy is still interesting because the issues of scientific practice that it raised are so fundamental. It was the first direct clash of two entirely different conceptions of research in the history of psychology. One way of characterizing this difference is in terms of the models of proper research used by the two sides. For Wundt and Titchener the model was clearly derived from the experimental sciences, especially physics and physiology. In those sciences experimentation involved the creation of artificial conditions, nowhere to be found in the world outside the

laboratory. It also involved a high degree of idealization that isolated certain features in a purity that was foreign to the world of everyday phenomena. There was a price to be paid for the extension of that kind of experimentation to psychology. As we have seen, it entailed severe restrictions on the scope of experimental psychology, with virtually all practical or everyday psychological questions lying well outside that scope. It is hardly surprising that most psychologists were not prepared to pay this price, especially as the advantages of Wundtian scientific asceticism proved to be somewhat elusive.

The alternative was to begin with the everyday world, to work with psychological material that was hopelessly "impure," contaminated by countless disturbing factors of limited general interest. That meant working with naive experimental subjects and even with children or with psychological deviant adults. Baldwin was clearly moving in this direction, though others were later to go much further. It was a way of doing psychology that had more in common with field studies—the kind that had been lent enormous prestige by the work of Darwin—than it had with physical experimentation.

It is clear that the question of replication had to take different forms in these divergent styles of scientific practice. For the physics-oriented style the criterion of valid replication primarily involved questions of experimental control. Had potentially contaminating factors been controlled as well or better than in the original study, so that the phenomenon of interest could emerge in desired purity? Obviously, Baldwin's studies, with their naive subjects, must fail this test and could not be accepted by members of Wundt's experimental community as a replication of their own work.

Conversely, Baldwin rejected the work of the Wundtians, but his reasons were different. He intimated that they had created their empirical phenomenon by a process of subject *selection* and therefore had misinterpreted its nature. Indeed, if one's research practice is modeled on the *collection* of samples from the given, everyday world, rather than on the *isolation* of pure systems, one has to assign great importance to the process of selection. Effective replication then depends crucially on the comparability of selected samples. Would all samples of naive subjects anywhere be comparable to the sample that Baldwin happened to chance upon in his study? If not, what exactly is the significance of his results? To raise such questions is to run ahead of historical developments. But the kind of research practice of which Baldwin's study represents an early example would eventually have to come to grips with such issues. Its ability to do so, however, depended on developments in another quarter.

THE GALTONIAN TURN

Although 19th-century physical science was providing ever more impressive demonstrations of how one could lay bare the hidden nature of things by

studying them under highly controlled laboratory conditions, an altogether different method of achieving the same purpose was being tried out in the nascent social sciences. That method ultimately came down to the simple expedient of counting heads.

The emergence of modern societies with a rationalized industry, agriculture, and military system had been accompanied by great improvements in the efficiency of techniques of administration that were highly dependent on precise and extensive record keeping. That meant that much more complete statistical information was now available on a range of human behavior that included marriage, divorce, and various categories of crime, like suicide and homicide. When such data were examined, it emerged that the rates and fluctuations of these social acts were remarkably regular and varied with identifiable factors (Porter, 1986; Quetelet, 1969). It seemed that the categorizing and counting of individuals could lead to the determination of factors in human conduct that might remain hidden if one only studied individuals.

However, the analysis of social statistics had little or no direct influence on psychology because its findings were generally attributed to an "average individual," rather than to real persons. In contemplating the distribution of individuals on a parameter the social statisticians focused on the mean value and effectively treated deviations from the mean as "error." That made sense in terms of their special interest in general social trends; it made no sense to Francis Galton who was obsessively interested in the inherent differences among people (Fancher, 1984). He thought that such differences, rather than environmental factors, were responsible for existing social distinctions and for social progress. Quite consistently, he advocated a eugenic program of improving the human condition by selective breeding rather than by social reform. From this perspective, what was interesting about human statistics were not the averages but rather the distribution of individuals (Cowan, 1972). So far from treating this distribution as "error," he made it the main focus of interest. By means of an analysis of interindividual differences he demonstrated, to his own satisfaction, that social achievement was indeed a matter of heredity.

Galton's statistical analysis of individual differences provided the foundation for a new type of psychological investigation that was utterly different from the laboratory psychology pioneered by Wundt and his followers. It led directly to the invention of the technique of mental testing as well as to the development of relevant statistical techniques of data analysis by the biometric school of Karl Pearson and others (MacKenzie, 1981). Although most of the important early figures in these developments were, like Galton, motivated by eugenicist concerns, a second consideration gradually became more significant for the psychologists. The statistical analysis of interindividual differences offered the possibility of founding a scientific psychology

on the collection of measures from a relatively large number of naive subjects rather than on the intensive study of a very few practiced subjects under restrictive laboratory conditions. This answered to a widespread desire, especially among American psychologists, that psychology should provide data of immediate relevance to practical social concerns. The new approach also sanctioned the use of "quick and dirty" methods of data collection, like questionnaires, and made it possible to replace cumbersome laboratory controls with statistical controls.

The issue of replication now presented itself in a different form. In the traditional laboratory approach one attempted to replicate findings on individual psychological or physiological exemplars. This could no longer be the aim, because it was precisely the differences between these individual exemplars that had now become the essential data base. What one now had to ask was not whether a particular pattern of results was stable from one individual to another, but whether a particular pattern of interindividual differences remained stable from one occasion to another. The direct object of investigation was a statistical distribution, a "collective object" as Fechner (1897) called it, not the parameters of an individual mind or person.

It should be noted that in redefining the question of replication the new approach also sidestepped it. For in constructing a distribution of individual differences the assumption is made that the points on this distribution all represent measures of the same thing, that is, that the responses obtained from different individuals are truly comparable, that there is a continuous underlying variable. But it may be that responses near one end of a distribution are psychologically different from responses near the other end or near the middle, that the underlying processes are qualitatively different. The comparability of individual measures is achieved by fiat, by giving everyone the same tasks, by forcing responses into a standard pattern, and by avoiding any inquiry into how the responses were arrived at. If the traditional laboratory psychology was too glib in assuming the generalizability of its findings, the statistical approach to individual differences was equally glib in its assumption of a continuous variation of psychological functions across individuals.

Nevertheless, Galton's example led to early attempts at developing mental tests (Cattell, 1890) for grading and selecting individuals. The key statistical tool that Galton had developed for the analysis of interindividual differences of performance measures was based on the technique of correlation. Once Pearson had refined and codified this technique in the form of the correlation coefficient, its large-scale adoption by psychologists interested in individual differences was only a matter of time.

The first major figure in the development of correlational techniques for specifically psychological purposes was Charles Spearman at University Col-

lege, London. Strongly influenced by Galton (Spearman, 1930), whose eugenicist concerns he shared (Norton, 1979), Spearman developed a methodology in which the technique of correlation played a crucial role.

Quite early in his work Spearman had to come to terms with the fact that if one correlated two sets of psychological measurements on one occasion and then repeated the exercise on another occasion one would be most unlikely to get the same result. Typically, statistical coefficients were subject to wide variation, too much variation for the taste of many whose model of science was based on the ideal of highly controlled laboratory experimentation. In the face of such skepticism psychometricians like Spearman attempted to make a distinction between acceptable and unacceptable degrees of variation. Obviously, any scientific coefficient would have some unreliability, no matter how slight. The problem was to distinguish this kind of unreliability from the gross and more serious kind. For that one needed a statistical measure of unreliability.

In his development of such a measure Spearman (1904, 1907, 1910) distinguished between two sources of "fallacy" or of variation in the value of the obtained correlation. One source was the "accidental deviations" of the observed correlation:

> For, though the correlation between two series of data is an absolute mathematical fact, yet its whole real value lies in our being able to assume a likelihood of further cases taking a similar direction; we want to consider our result as a truly representative sample. Any one at all accustomed to original investigation must be aware how frequently phenomena will group themselves in such a manner as to convincingly suggest the existence of some law—when more prolonged experiment reveals that the observed uniformity was due to pure hazard and has no tendency whatever to further repeat itself. (Spearman, 1904, p. 75)

Spearman proposed that an observed correlation possessed an "evidential value" in proportion to the likelihood that repeated investigations would yield similar correlation values. But he did not suggest how to estimate this likelihood directly. Instead, Spearman defined "evidential value" negatively. Following the practice of the biometric school, he estimated the probability of obtaining the observed correlation by chance alone. This value is defined by the ratio of the observed correlation to its probable error. The greater the value of this ratio, the smaller is the probability of obtaining the observed correlation by "pure hazard," and the greater its "evidential value."

The second source of error in the value of an observed correlation was due to "systematic deviations." These were the constant, noncompensating variations, either increasing or decreasing the value of an observed correlation from its "true value." Usually, a "systematic deviation" would be due to the

presence of biasing factors considered irrelevant to the question being asked in the investigation.

In effect, Spearman assumed, without much justification, that the reality underlying his empirical findings conformed to a particular statistical model. This could provide an excuse for various manipulations of the data to push them more into line with theoretical preconceptions. Even when this resulted in correlation coefficients above one, Spearman does not seem to have questioned his fundamental assumptions (Fancher, 1985).

The question of reconciling theoretical preconceptions with empirically gathered data was in fact crucial during the emergence of the statistical reliability concept in psychology. Later on, purely pragmatic considerations regarding predictability came to assume more importance, but the early importance of unquestioned theoretical assumptions is striking.

In psychology the statistical reliability concept received explicit treatment almost exclusively within the context of mental testing. In their early days such tests were developed for the express purpose of providing empirical evidence for the existence of unchanging mental abilities for which there existed stable individual differences of degree. Empirically obtained scores on psychological tests, however, never showed anything like the stability that the theory of unchanging individual traits seemed to demand. The notion of test reliability was a first attempt to deal with this discrepancy. It was based on the idea that an actual test score was only a more or less reliable estimate of an underlying "true" score. As the domain to be dealt with was in fact one of interindividual differences in scores one would then make a fundamental distinction between "true variance" and "error variance." The reliability of individual difference measures would be a function of the proportion of true variance.

The estimation of this ratio required an application of the principle of replication. If the qualities assessed by mental tests were truly inert one could establish reliability by repeated administration of the same test to the same group of individuals. This would enable one to estimate unsystematic errors that fluctuated from one occasion to another. But, of course, the qualities of interest are far from inert. People become familiar with the test and their reaction to it changes. So one had to work with "equivalent" or "parallel" forms of the same test. The criterion of similarity was however statistical—the correlation of performance measures on the different forms—and not psychological. High statistical reliabilities might be due to various kinds of psychological factors, including relatively constant differences in motivation, experience, and/or ability. However, in the early days of mental testing there was a strong tendency to jump to psychological conclusions on the basis of highly ambiguous statistical data.

Moreover, estimates of reliability, being based on sets of variances, were highly sensitive to changes in these sets. For instance, increasing the heter-

ogeneity of the population from which performance data were obtained would produce higher reliability coefficients, a fact sometimes exploited by producers and marketers of tests. Different kinds of comparison (e.g., retesting on the same or on "alternate" forms of the test, or comparing two halves of the same test) would also lead to different reliability estimates. In general, the kind of replicability that is expressed in the concept of statistical reliability is no less tied to highly specific conditions than is replicability of laboratory experiments.

It took several decades for this insight to become widely accepted. Especially during the period before World War II there was a strong tendency to believe that the establishment of statistical reliabilities could somehow substitute for the rigors of experimental replication and the uncertainties of psychological theorizing. In a historically important contribution, Thurstone summed up the situation in 1931:

> In this country the reliability formulae have become a sort of fetish rather than a tool. It is almost as though "busy-work" becomes science as soon as it can be made to sprout correlation coefficients. Test investigators seem frequently to be more concerned that all their figures have probable error attachments than that their studies be motivated with ideas or hypotheses to be examined. (Thurstone, 1931, p. 2)

A NEW KIND OF EXPERIMENTATION

Thurstone was looking back on a period during which much psychological research was either based on correlational methods or else continued in the classical pattern of measuring the responses of individual subjects under highly controlled conditions. However, a third type of research practice, one that combined elements of both the classical Wundtian and the Galtonian approach, was definitely on the rise and was soon to provide the dominant model for psychological research. In this type of research one compared group data, but the groups in question had been subjected to different experimental conditions. The question to be answered was whether this difference in conditions could reasonably be held responsible for any observed difference in group performance measures.

Although psychologists were relatively late in appreciating the advantages of comparative experimentation, its use in practical situations that excluded the possibility of laboratory controls can be traced as far back as the 18th century. At that time the rationalization of agriculture was being enthusiastically pursued in England and Holland. In order to maximize yields, men like Arthur Young (Cochran, 1976) began to experiment with systematic field trials that attempted to study the effect of various factors, like soil and

location. No satisfactory method of evaluating the results of such experiments was, however, available.

Comparative experimentation was not unknown in 19th-century medicine (Lilienfeld, 1982). Here, as in the agricultural setting, one tried to experiment under conditions of great ignorance about the nature of the processes responsible for the effects one was interested in. Moreover, the possibilities for isolating and controlling relevant factors were extremely limited. Nevertheless, by replicating observations on sets of individuals exposed to different conditions very useful practical information could sometimes be extracted.

The early users of comparative experimentation wanted their experiments to provide a rational basis for decisions concerning alternate courses of practical action in complex situations. They were interested in the overall relative effectiveness of various practical measures and therefore conducted their experiments in the settings to which their conclusions were to be applied. But these were settings in which adequate experimental control was impossible. Research in such conditions faced difficulties in differentiating the effects of relevant and interesting factors from haphazard and uncontrollable factors.

Comparative experimentation was devised to deal with constraints on experimental control through a process of systematic replication (Sidman, 1960). *In comparative experimentation the experimenter attempts to carry* out a set of experiments (replications) under the "same" conditions. Within such a set, each experiment is considered to differ from the others in a "systematic" way and in an "accidental" way. Ideally, systematic differences are defined as those due to the factors of interest to the experimenter, whereas all other differences are accidental. For example, the results of one experiment, in a set of two, would be assumed to have been produced by a combination of accidental factors (such as variations in soil fertility and climate) plus a systematic factor (say, the traditional method of seeding). The results of the other experiment in the set would be assumed to have been produced by a similar combination of accidental factors (soil and climate) plus another systematic factor (the new method of seeding). In this way, when the two experiments (replications) are compared, the effect of the accidental factors on the difference between the results of the two experiments is minimized, although the effect of the systematic factors on this difference is maximized. However, because it cannot be assumed that the accidental factors have exactly the same effect in the two experiments, they are still likely to have an effect on the difference between the results of the two experiments. Therefore, the experimenter still has the problem of deciding whether the obtained difference in results is large enough to be attributed to the effect of systematic factors over and above the effect of the accidental factors. Statistical tests were eventually proposed as a solution to this problem. They relied on hypothetical distributions of statistical constants obtained

in the idealized conditions of infinite repeated sampling under identical conditions.

In line with earlier developments in other fields, it was an eminently practical motive that prompted the extension of comparative experimentation to education, from which it quickly spread to psychology. The pioneer in this field was a London school inspector, W. H. Winch (Boring, 1954; Rusk, 1932; Solomon, 1949), who was a member of a group of progressive educators interested in developing a more rational educational practice (Hearnshaw, 1964). "The believer in experimental pedagogy," wrote Winch (1914, p. vii), "holds that, until an educational proposal has been submitted to definite tests under rigorously scientific conditions in the schools themselves, there can be no adequate ground for recommending it for general adoption." In his own research, which was always conducted in a classroom, Winch phrased his experimental problems in terms of the practical questions faced by a school administrator or a classroom teacher, though he saw these questions as implying more general psychological issues.

Winch first used comparative experiments in investigations of transfer of training in memory, a notion that provided a rationale for some then prevalent teaching methods. In these studies Winch (1908, 1910a, 1910b) wanted to show that "special practice" in memory would lead to improvement in "productive imagination," as indicated by story writing. Winch divided a group of school children from the same class into two "equivalent groups" on the basis of preliminary tests in "productive imagination." Then, he administered "special practice" in memory to one of the groups, while the other group, the control group, had undergone "natural growth" only. This involved the performance of some "neutral" task, such as, simple arithmetic problems. With this exception all students were to be treated the same way by their teacher. Winch then compared the performances of the two equivalent groups on another test of "productive imagination." In order to carry out the comparison, Winch calculated the percent difference for the average scores of the two groups as a whole.

When Winch introduced what is now called the control group design he was aware that it was a methodological innovation in educational psychological research (Winch, 1911). Although his design was accepted, his way of making inferences from the data was criticized. An early criticism came from Spearman's student, W. G. Sleight (1911), who pointed out that differences in improvement between the practice and control groups were small enough to have been produced by chance alone, and that Winch did not account for this possibility. In his own investigations Sleight followed the Winch design, but tried to show that the obtained difference was not likely due to chance by calculating the ratio of the obtained difference to its probable error. If this ratio was greater than 3, it was considered to indicate a statistical difference described as "significant," "reliable," "certain," or "true." It was held to

indicate the presence of a systematic factor over and above any accidental factors producing a group difference. The operation of this systematic factor was presumably linked to the difference in experimental conditions. The index used by Sleight became commonly known as the "Critical Ratio" during the 1920s (Walker, 1929).

STATISTICS TAKES OVER

The statistical procedures developed by Karl Pearson and adopted by Spearman and other mental testers required that large sample sizes be used in their investigations. A large sample size made it possible to use the well-known characteristics of the normal probability curve when estimating the significance of an obtained statistic. Thus, the probability values needed for testing the reliability of an obtained correlation, or the significance of the critical ratio of an obtained difference between two means were defined in terms of areas under the normal probability curve, assuming a normal sampling distribution. In comparative experimentation, however, the sample sizes were often quite small. Under this condition, the sampling distribution of the statistical constants deviated greatly from the expectations based on a normal distribution. This made the normal distribution useless for a whole range of practical problems. A new set of hypothetical sampling distributions had to be discovered for these types of problems.

The first such distribution was proposed by W. D. Gosset, working on problems of quality control at the Guinness brewery. In 1908, he published his results under the pseudonym "Student." Despite its practical importance, the value of his contribution was not generally recognized beyond the Guinness brewery. As late as September 1922, Gosset wrote R. A. Fisher: "I am sending you a copy of Student's Tables as you are the only man that's ever likely to use them" (Cochran, 1976, p. 13). Indeed, it was not until Fisher generalized Gosset's results that his work received wide application.

Ronald A. Fisher was a mathematician who first started working in statistics in order to resolve the conflict between Mendelian experimental geneticists and Biometricians (Mackenzie, 1981). In 1919, he was asked to join the Rothamstead Experimental Station to examine about 70 years' worth of results from agricultural field experiments in order to "see whether he could elicit from it further information" (Box, 1978, p. 96). In the context of this work he developed the analysis of variance (ANOVA) procedure and his interest in the design of experiments.

However, Fisher did not see ANOVA as merely a solution to local problems in agricultural field experimentation. Rather, he saw it as part of a general solution to the philosophical problem of describing the nature of inductive inference from any experiment, whether of the field or the laboratory

variety. Fisher's *Statistical Methods for Research Workers* (1928), first published in 1925, differs strikingly from other contemporary statistical textbooks in the emphasis placed on methods of statistical inference and on small-sample problems typical of experimental research. Commenting on the notion of statistical inference, Fisher explicated an assumption implicit in the first applications of the "Normal Law of Error" to the reliability of experiments. In Fisher's (1928) words:

> Just as a single observation may be regarded as an individual, and its repetition as generating a population, so the entire result of an extensive experiment may be regarded as but one of a population of such experiments. The salutary habit of repeating important experiments, or of carrying out original observations in replicate, shows a tacit appreciation of the fact that the object of our study is not the individual result, but the population of possibilities of which we do our best to make our experiments representative. The calculation of means and probable errors shows a deliberate attempt to learn something about that population. (p. 3)

Once experimentation is conceptualized as a sampling process, then, in Fisher's words, "statistical procedure and experimental design are only two different aspects of the same whole" (Fisher, 1935/1966, p. 3).

In the 1930s, Fisher's ideas on statistics and experimental design gained a wide following among North American agricultural researchers and statisticians. Between 1924 and 1937, Fisher made extended visits to the United States and Canada (Box, 1978). In psychology, during this time comparative experimentation became a major alternative to the laboratory experiments modeled on experimentation in physiology. At the same time there was a growing trend to use the "critical ratio" as the criterion of the value of the obtained results. "Statistical significance," indicated by the critical ratio, was becoming the arbiter of what Spearman had called the "evidential value" of an experiment. Fisher's Analysis of Variance was introduced into psychology in the mid 1930s. Its first applications were in educational psychology, parapsychology, and occasionally in animal psychology. After World War II, ANOVA was applied widely to experimentation in all areas of psychology (Lovie, 1979; Rucci & Tweney, 1980).

One fundamental consequence of this development was the implicit incorporation in the methods of psychology of Fisher's belief that the design of experiments should be governed by statistical considerations (Gigerenzer & Murray, 1987). But if, as Fisher says, an entire experiment must be considered one of a population of such experiments, the question of the nature of this population becomes rather important. Unfortunately, it is easier to give a statistical than a psychological answer to this question, a fact that constituted an ever-present temptation to substitute statistical for psychological solutions to research questions.

CONCLUSIONS

Although all psychologists who thought of their subject as a science agreed that replicability constituted an essential feature of their work, they found it much more difficult to agree on the criteria of acceptable replication and on the role played by replication in their research practice. For Wundt and Titchener successful replication necessarily depended on laboratory control of the conditions that made it possible to obtain accurate reports on simple conscious experience. That view of replication was a consequence of the way they had defined the subject matter of psychology. As soon as others ceased to accept that definition they found themselves turning to alternative conceptions of replication. A broader conception of the kind of experience that could form the legitimate subject matter of experimental psychology led to a corresponding loosening in conceptions of acceptable replication on the part of the Würzburg experimenters. Once the subject matter shifted from conscious experiences to the responses of different kinds of people, as it did in Baldwin's experiments, the question of the selection of subjects could no longer be ignored in discussions of replication.

Spearman stood at the beginning of a long tradition of psychological investigation which defined its task as the estimation of the "real value" of stable patterns of interindividual differences in performance on set tasks. Replication, therefore, involved the reproducibility of particular statistical patterns among the scores of a group of individuals. An analogous emphasis on repeatable statistical patterns was characteristic of the turn to comparative experimentation, though here the question was one of establishing a reliable link between such patterns and specific experimental interventions. It is clear that throughout the history of modern psychology investigators have defined and redefined the issue of replication in terms of their assumptions about their subject matter and the basic research goals that they adopted. General discussions of the topic have no absolute significance but always need to be interpreted in the light of the prevailing historical context.

This does not mean that when psychologists adopted diverging views on the subject of replication they necessarily differed on every conceivable point. In fact, their profound disagreements did not preclude tacit agreement on certain shared assumptions. We briefly mention two of these. In the first place, there appears to have been a pervasive tacit agreement to eliminate history from psychology. Whatever the changes that were imposed on the definition of the subject matter of psychological investigation, one thing did not change, namely, the belief that what psychologists studied were processes and functions not subject to historical change. In other words, the things that were posited as the objects of psychological research were conceived to be analogous to the objects of research in the natural sciences and not at all like the objects of historical research. Given this basic assumption

the notion of replicating psychological experiments appeared to be unproblematic in principle, though there was still plenty of room for disagreement on what this meant in practice. If the things one experimented on were stable and unchanging—individual abilities and specific environmental effects, for example—one could expect investigations to be repeatable and findings to be replicable.

The one major figure for whom this assumption appeared to be problematical was Wundt. He accepted its reasonableness for certain areas of psychological investigation, notably sensation and perception, but rejected it as implausible for other areas. So Wundt projected two psychologies, only one of which could be experimental. Subsequent generations of psychologists, however, seem to have taken it for granted that ahistoricity was a basic property of any phenomena that psychologists might choose to investigate. Among other things, this conviction made it appear perfectly reasonable to adopt the Fisherian model of the infinitely repeatable experiment as a basic paradigm for all psychological research. One unfortunate by-product of the heavy institutionalization of a single methodological paradigm (Gigerenzer, 1987) was that it precluded systematic examination of the relative appropriateness of its assumptions for different areas of psychological research.

A second widespread preconception concerned the independence of methods of investigation from whatever they were supposed to investigate. In the historical discussions that are relevant to the topic of replication it is possible to discern a somewhat ghostlike figure that seems to preside over the proceedings. The ghost is that of a completely method-independent absolute knowledge that everyone seemed to believe in although no one had ever actually encountered it. Belief in this ghost was clearly implied in all those contributions that advocated some set of methodological principles that was to maximize objective knowledge irrespective of the nature of the object one wanted to know. However, as we have seen, methodological priorities always depended on the kind of knowledge that was desired. Contrary to the illusions of many investigators, the attempt to separate the conditions and the products of psychological research was never successful. That is hardly surprising, for knowledge is after all a relationship with an object and not a thing in itself.

REFERENCES

Angell, J. R., & Moore, A. W. (1896). Reaction time: A study in attention and habit. *Psychological Review, 3*, 245–258.

Baldwin, J. M. (1895). Types of reaction. *Psychological Review, 2*, 259–273.

Baldwin, J. M. (1896). The type theory of reactions. *Mind, 5*, 81–90.

Böhme, G. (1977). Cognitive norms, knowledge-interests and the constitution of the scientific object: A case study in the functioning of rules for experimentation. In E. Mendelsohn, P.

Weingart, & R. Whitley (Eds.), *The social production of scientific knowledge* (pp. 129–141). Dordrecht: Reidel.

Boring, E. G. (1954). The nature and history of experimental control. *American Journal of Psychology, 67*, 573–589.

Box, J. F. (1978). *R. A. Fisher, The life of a scientist*. New York: Wiley & Sons.

Bühler, K. (1908). Antwort auf die von W. Wundt erhobenen Einwände gegen die Methode der Selbstbeobachtung an experimentell erzeugten Erlebnissen. *Archiv für die gesamte Psychologie, 12*, 93–122.

Cattell, J. McK. (1890). Mental tests and measurements. *Mind, 15*, 373–381.

Cochran, W. G. (1976). Early development of techniques in comparative experimentation. In D. B. Owen (Ed.), *On the history of statistics and probability*. New York: Dekker.

Collins, H. M. (1985). *Changing order: Replication and induction in scientific practice*. Beverly Hills, CA: Sage.

Cowan, R. S. (1972). Francis Galton's statistical ideas: The influence of eugenics. *Isis, 63*, 509–528.

Danziger, K. (1980). The history of introspection reconsidered. *Journal of the History of the Behavioral Sciences, 16*, 241–262.

Fancher, R. E. (1984). The examined life: Competitive examinations in the thought of Francis Galton. *History of Psychology Newsletter, 16*, 13–20.

Fancher, R. (1985). *The intelligence men: Makers of the IQ controversy*. New York: Norton.

Fechner, G. T. (1897). *Kollektivmasslehre*. Leipzig: Engelmann.

Fisher, R. A. (1928). *Statistical methods for research workers* (2nd ed.). Edinburgh: Oliver and Boyd.

Fisher, R. A. (1966). *The design of experiments*. Edinburgh: Oliver and Boyd. (Original published in 1935)

Gigerenzer, G. (1987). Probabilistic thinking and the fight against subjectivity. In L. Krüger, G. Gigerenzer, & M. S. Morgan (Eds.), *The probabilistic revolution, Vol. 2: Ideas in the sciences* (pp. 11–33). Cambridge, MA: MIT Press.

Gigerenzer, G., & Murray, D. J. (1987). *Cognition and intuitive statistics*. Hillsdale, NJ: Erlbaum.

Hacking, I. (1983). *Representing and intervening*. New York: Cambridge University Press.

Hearnshaw, L. S. (1964). *A short history of British Psychology: 1840–1940*. New York: Barnes and Noble.

Humphrey, G. (1950). *Thinking*. London: Methuen.

Krantz, D. L. (1969). The Baldwin-Titchener controversy: A case study in the functioning and malfunctioning of schools. In D. L. Krantz (Ed.), *Schools of psychology: A symposium* (pp. 1–19). New York: Appleton-Century-Crofts.

Lilienfeld, A. M. (1982). Ceteris paribus: The evolution of the clinical trial. *Bulletin of the History of Medicine, 56*, 1–18.

Lovie, A. D. (1979). The analysis of variance in experimental psychology: 1934–1945. *British Journal of Mathematical and Statistical Psychology, 32*, 151–178.

Mackenzie, D. A. (1981). *Statistics in Britain 1865–1930*. Edinburgh: Edinburgh University Press.

Mehner, M. (1885). Zur Lehre vom Zeitsinn. *Philosophische Studien, 2*, 546–602.

Norton, B. (1979). Charles Spearman and the general factor in intelligence: Genesis and interpretation in the light of socio-personal considerations. *Journal of the History of the Behavioral Sciences, 15*, 142–154.

Popper, K. R. (1959). *The logic of scientific discovery*. New York: Harper & Row.

Porter, T. M. (1986). *The rise of statistical thinking 1820–1900*. Princeton, NJ: Princeton University Press.

Quetelet, L. A. (1969). *A treatise on man and the development of his faculties*. Gainesville, FL: Scholars' Facsimiles.

Rucci, A. J., & Tweney, R. D. (1980). Analysis of Variance and the "second discipline" of scientific psychology: A historical account. *Psychological Bulletin, 87*, 166–184.

Rusk, R. R. (1932). *Research in education, an introduction*. London: University of London Press.

Schultz, D. P. (1970). The nature of the human data source in psychology. In D. P. Schultz (Ed.), *The science of psychology: Critical reflections*. New York: Appleton-Century-Crofts.

Sidman, M. (1960). *Tactics of scientific research, evaluating experimental data in psychology*. New York: Basic Books.

Sleight, W. G. (1911). Memory and formal training. *British Journal of Psychology, 4*, 386–451.

Solomon, R. L. (1949). An extension of control group design. *Psychological Bulletin, 46*, 137–150.

Spearman, C. (1904). The proof and measurement of association between two things. *American Journal of Psychology, 15*, 72–101.

Spearman, C. (1907). Demonstration of formulae for true measurement of correlation. *American Journal of Psychology, 18*, 161–169.

Spearman, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology, 4*, 271–295.

Spearman, C. (1930). Autobiography. In C. Murchison (Ed.), *A history of psychology in autobiography* (Vol. 1). Worcester, MA: Clark University Press. "Student." (1908). The probable error of a mean. *Biometrica, 6*, 1-25.

Thurstone, L. L. (1931). *The reliability and validity of tests*. Ann Arbor, MI: Edwards Bros.

Titchener, E. B. (1895). The type theory of the simple reaction. *Mind, 4*, 506–514.

Titchener, E. B. (1896). The "type theory" of simple reaction. *Mind, 5*, 236–241.

Walker, H. (1929). *Studies in the history of Statistical Method: With special reference to certain educational problems*. Baltimore: William & Wilkins.

Winch, W. H. (1908). The transfer of improvement of memory in school children. *British Journal of Psychology, 2*, 284–293.

Winch, W. H. (1910a). Some measurement of mental fatigue in adolescent pupils in evening schools. *Journal of Educational Psychology, 1*, 13–23.

Winch, W. H. (1910b). Accuracy in school children. Does improvement in numerical accuracy "transfer"? *Journal of Educational Psychology, 1*, 557–587.

Winch, W. H. (1911). Some relations between substance memory and productive imagination. *British Journal of Psychology, 4*, 95–125.

Winch, W. H. (1914). *Children's Perceptions: An Experimental Study of Observation and Report in School Children* (Educational Psychology Monographs, No. 12). Baltimore: Warrick and York.

Wundt, W. (1883). *Methodenlehre*. Stuttgart: Enke.

Wundt, W. (1907). Über Ausfrageexperimente und über die Methoden zur Psychologie des Denkens. *Psychologische Studien, 3*, 301–360.

CHAPTER 2

Replication, Restudy, And The Nature Of Anthropological Fieldwork*

Peter Kloos
Vrije Universiteit
Amsterdam, The Netherlands

INTRODUCTION

In March 1989, two physicists, Stanley Pons and Martin Fleischmann, announced that they had produced surplus energy through nuclear fusion using a simple, table-top apparatus. Within a few weeks, scores of scientists throughout the world were trying to replicate their experiment. Everywhere, the results were inconclusive but then, not all the details of the Pons/Fleischmann experiment had been made available immediately.

Here, we have a straightforward example of an alleged important discovery in physics inducing other scholars to test the conclusions drawn from empirical research by independently replicating the experiment. Although in actual practice replication may occur less often than physics textbooks might suggest (see Broad & Wade, 1982), replication and replicability are regarded as essential elements of research in the natural sciences.

In anthropology, both the ideal and the practice of replication are very different. When Malinowski published his *Argonauts of the Western Pacific* in 1922—a work widely acknowledged as the first modern ethnographic study—no one rushed to the Trobriand Islands to see whether that impressive interinsular network of barter called *kula* really existed. When Mead (1928) published her views on the cultural background to adolescent *Sturm und Drang* six years later, no one traveled to Samoa to find out whether adolescence really was such a peaceful phase in the life cycle of the Samoans. And when in 1940, Evans-Pritchard published his seminal explanation of how

*I am indebted to Zella Carriere for improving my use of a language not my own.

200,000 quick-tempered Nuer managed to live together without daily blood shed, even though they lacked centralized legal and political institutions, his readers believed what he wrote and refrained from going to the South Sudan to look for themselves.

In cultural anthropology, replication is definitely *not* a basic ingredient of research practice. It hardly features in textbooks on methods of ethnographic research. In fact, it is often argued that replication in anthropological fieldwork is not possible at all. Still, anthropological field workers do use the term "*restudy*"—*which seems to come close to replication.*

In writing this chapter I first clarify the anthropologist's apparent aversion to replication which, at first sight, seems such a self-evident scientific operation. Secondly, I review the various ways in which anthropologists deal with differences in fieldwork results where these emerge as a consequence of a restudy. To be sure, anthropological research is not restricted to field research. However, I still focus on problems of fieldwork because this form of research dominates the profession and thus later, I briefly sketch the development and nature of fieldwork. I then turn to problems of replication as far as anthropological fieldwork is concerned. The phenomenon called "restudy" and the anthropologist's interpretation of differences in fieldwork results are also discussed in this chapter. Concrete cases of ethnographic discrepancies are briefly described as well. In the final section, I return to the basic question: Is replication possible in anthropological fieldwork and, if not, what does this mean for the status of anthropology?

ANTHROPOLOGICAL FIELDWORK: DEVELOPMENT AND NATURE

Although fieldwork is now by far the most conspicuous form of research in anthropology, this has not always been so. Anthropology as a discipline originated from an awareness of cultural differences among various societies. Within the Western tradition this awareness, as well as the need to deal with cultural differences, can be traced back at least to ancient Greek writers such as Herodotus. Although he is often claimed by historians as their primeval ancestor, anthropologists have as good a claim to this intellectual line of descent as historians have (see de Waal Malefijt, 1974). In his writings, Herodotus described more than 50 different peoples and their cultures, some of them in great detail as in the case of the Scythians. He also asked questions that are basic to anthropology as it is known today such as: Why do people in different societies behave differently?

Awareness of cultural differences as a necessary condition for anthropology to exist implies that the development of the discipline is directly related to intercultural contacts. When geographic horizons shrank during

the European Middle Ages, interest in foreign cultures dwindled. However, the age of discovery, especially the first transoceanic voyages to other continents, brought the problem of cultural differences among human societies forcefully back to the forefront of European thought. The so-called "discovery" of the Americas posed special problems. Here, European thought was confronted with people whose existence was hitherto completely unknown: Even the Bible and the ancient Greek writers make no mention of a continent in the Western hemisphere, let alone of the multifarious cultures that existed there. Travelers of many kinds, discoverers, *conquistadores*, missionaries, and later colonial administrators reporting on strange customs all found avid readers in Western Europe. From the 16th century onward countless writers, like historians, philosophers, and also novel writers, commented on the problems of different ways of life (see Slotkin, 1965), even though very few of them should be called "anthropologists."

Anthropology as a more or less circumscribed discipline came into being in the early 19th century with the establishment of Ethnological Societies in Paris (1839) and in London (1843). These societies were created to promote greater knowledge of cultures in other continents (see Stocking, 1983). Knowledge of differing cultures was of practical significance, particularly for colonial powers such as the Netherlands. Dutch colonial administrators realized very early on that the administration of the Netherlands East Indies required knowledge of the cultures and languages of its peoples. Toward the end of the 19th century, anthropology was given academic recognition through the establishment of university Chairs. Professor P.J. Veth was appointed to the first Chair of Anthropology in the Netherlands in 1877; in England, E. B. Tylor was first appointed lecturer in "Ethnology" in 1883 and was made a professor later on.

Although many of those who began to call themselves anthropologists or ethnologists had been abroad, their scientific work was largely based on information supplied by discoverers, *conquistadores*, sailors, traders, missionaries, and colonial administrators. At that time, anthropologists were not accustomed to gathering their own data to formulate their theories, hence, the term "armchair" anthropology applied to this kind of scientific work.

These armchair anthropologists were, of course, aware of the unreliability and the limitations of the data they had at their disposal. They tried to improve the quality of the data by drawing up lists of questions designed to systematically cover the subjects that attracted their interests (see, for an early example, Degérando's *Considérations sur les méthodes à suivre dans l'observation des Peuples Sauvages*, published as early as 1800).

Pressure to improve the quality of ethnographic data finally led to a methodological revolution in the first quarter of the 20th century (see Kuper, 1983). Thanks to the British Anthropological Society, trained observers were sent to other continents to collect ethnographic data. Trained observers

quickly became *anthropologically* trained observers. A key figure eventually emerged in the person of Bronislaw Malinowski who, during World War I, lived for more than a year among the Melanesian Trobrianders.

Of Polish descent, Malinowski studied anthropology in England, went to Melanesia, learned the local language, and carried out research by participating as much as he could in whatever was taking place. Although the designation "participant observation" did not yet exist, it covers exactly what he did.

Back in London, Malinowski wrote a detailed account of what he had witnessed. He received an appointment at the London School of Economics and, being a gifted teacher, he convinced his pupils that they should carry out research in the way he said he had done. Within a few years, the distinction between the armchair anthropologist based at home and the data collector who ventured abroad disappeared. Henceforth, every anthropologist was expected to do fieldwork. It has even been stated that no one who has not done fieldwork could be regarded as an anthropologist. This indicates that fieldwork acquired a meaning beyond what it is basically about: the anthropologist's way of collecting, or rather of constructing, theoretically relevant data.

Without going into details, the Malinowskian revolution had a number of consequences that have to be mentioned. In the first place, modern fieldwork gave rise to a new kind of ethnography. Although there are many examples of valuable ethnographic accounts written, for example, by missionaries who often spent many years among those whose cultures they describe, the modern anthropological account combines theoretical relevance with rich and thorough descriptive coverage. A prolonged stay among the people, a learning of the language, and a relying on direct observation rather than on answers to questions all resulted in a wealth of information regarding a particular society and culture in its totality and in its individuality. The rise of fieldwork strengthened a holistic approach that was, to some extent, already present in armchair anthropology. The new approach brought many completely new phenomena to the attention of anthropologists. It also offered fresh insights into phenomena that were already known but poorly understood. For example, although the 19th-century American anthropologist Lewis H. Morgan may be credited for having established kinship studies, it was the modern approach that brought to light the manifold functions and complexities of kinship systems. Although 19th-century evolutionary anthropologists had written about political behavior, it was modern fieldwork that brought an understanding of the mechanisms of social control in societies lacking central political institutions. And although religious ideas and behavior had always attracted much attention, it was fieldwork in the style of Malinowski that showed what religion in its multifaceted forms meant in small-scale societies.

Given this exciting growth of detailed knowledge about other societies and cultures as well as the very small number of qualified anthropologists, it is no surprise that fieldworkers did not really consider going to societies that had already been studied. They preferred hunting grounds of their own: a society that was hardly known at all, a blank spot on the ethnographic map.

In the second place—and this only strengthens the aversion to replicating older research—fieldwork is not just a method of acquiring scientific data. It quickly became what is called in anthropology a *rite de passage*. It was the fieldwork experience as such, with all its hardships, which transformed the academically trained student of anthropology into an anthropologist in the true sense of the word. In a *rite de passage* the novice as a rule does not really know what is going to happen—it is something to be experienced, not something to be expounded beforehand. A *rite de passage* is a way toward a new and higher status and to a higher form of knowledge. In anthropology, this meant that the first generation of students sent into the field by such people as Malinowski in England and Boas in the United States received little advice on how to do field work. It was a swim-or-sink approach. Margaret Mead, a student of Boas, once wrote:

> There was, in fact, no *how* in our education. What we learned was *what* to look for. Years later, Camilla Wedgewood, on her first field trip to Manam Island, reflected on this point when she wrote her first letter back: "How anyone knows who is anybody's mother's brother, only God and Malinowski know." (Mead, 1972, p. 151)

When Beals, about to go off to the field, asked a distinguished British anthropologist what the secret of successful fieldwork was, he received this answer: "Never accept free housing and always carry a supply of marmalade" (Beals, 1970, p. 38).

The specific function of fieldwork in an anthropologist's career again puts a premium on finding a place where no one has carried out research before. Anthropologists going into the field are interested in finding something new and exotic rather than in finding support for something already known. Marvin Harris, referring to this tendency, once wrote:

> I remember well my own chagrin at having chosen to do fieldwork among the Bathonga, a patrilineal group in Southern Mozambique, when with a little more foresight I could have convinced the Ford Foundation to let me go to a more exotic *and hence professionally more rewarding* matrilineal culture slightly to the north. (1977, p. 82; emphasis added)

The status of the anthropologist and his or her authority in the profession is closely connected with the uniqueness of the fieldwork experience (see Mead's autobiography, 1972). Again, this attitude does not favor replication.

In the third place, there are elements of commitment and power involved. Fieldwork can be an almost traumatic experience. Because it implies living in a foreign culture—and to some extent, also living a foreign culture—fieldworkers often develop a very intense relationship with the people they study. This was especially true during the early years of modern fieldwork when two or three years of work in the field were often followed by a lifetime of writing on the society and culture studied. The relationship between fieldworker and society can, to some extent, be described in terms of ownership and of exclusive rights to carry out research, thus excluding other scholars. Paul Radin drew attention to this peculiar, though understandable, behavior as early as 1933. Such "claims" to exclusive rights also seem to be recognized by fellow anthropologists, especially when the first person to have carried out fieldwork in a particular society holds an important position. The expression "my village," which is often heard, provoked Brown (1981) to introduce the term *My Tribe*-syndrome for this behavior. Again, once a community has been investigated, recognition of certain rights both by the investigator him- or herself and by his or her colleagues tends to prevent replication.

This explanation of the absence of replications in anthropology is partly historical and partly sociological. It has nothing to do with methodological or epistemological considerations. It is to these considerations that I shall now turn.

IS REPLICATION IN ANTHROPOLOGICAL FIELDWORK POSSIBLE?

The question now under consideration is not so much why replication of anthropological field research is so rarely undertaken, but why many anthropologists seem to believe that replication in the true sense of the term is not really possible.

Replication in scientific research is directly related to one of the basic features of knowledge: reliability. Reliability defines the consistency between two or more scientific statements supposed to describe reality. Because reality can never be known other than through statements about reality, reliability is usually defined in an instrumental manner. The reliability of a statement about empirical reality increases in proportion to the number of independent observations that have led to the same statement. When we do not believe the outcome of an observation—for example when we have reason to believe that it may be unreliable—we repeat or *replicate* the observation. In its simplest form: If we do not trust the outcome of an arithmetical operation, we do it a second time. If the result is the same, the outcome of the two operations taken together is more reliable than either of

them taken separately. There can never be 100% reliability: Observations and measurements are replicable, but then again so are errors.

Replication as a way to more reliable knowledge is based on a number of assumptions, and I mention two of them. In the first place, it is assumed that empirical reality has not changed during the time that has elapsed between the first and the second observation or, to put it differently, that the same reality is being observed. Secondly, it is assumed that the way in which reality is measured is the same for the various observations. In physics experiments in which reality itself can be controlled and the same instruments can be used, these conditions can indeed be met.

In nonexperimental sciences where reality cannot be controlled, a problem arises: Differences between observations, that have been made with the same instrument may, at least *logically*, be attributed to changes in reality that have taken place during the time that has elapsed between the two observations. As long as one can make the measurements with the same instrument, or with instruments that can be calibrated, this is a logically acceptable conclusion.

In anthropological research however, the object, namely sociocultural reality, is not only beyond our control; we may also safely assume that it is continually changing. Moreover, we have to work with an instrument that cannot be calibrated and whose working is only partially known: The human individual. Powdermaker succinctly summarized this basic feature of anthropological fieldwork as follows: "The anthropologist is a human instrument studying other human beings and their societies" (Powdermaker, 1967, p. 19).

Given these fundamental uncertainties, one has to conclude that replication as carried out in experimental situations with instruments that can be controlled, or as carried out in natural situations but executed with instruments that can be controlled, is impossible in anthropology. In fact, as far as I know, there is only one study that makes that claim to some extent, namely Holmes's study of Samoa (Holmes 1957; cf. Holmes 1987).

Samoa was the subject of a famous study by Margaret Mead who went to this Polynesian island in 1925 to study adolescence. One of the intellectual issues of the 1920s was the so-called Nature-Nurture controversy. Was human behavior basically determined by ultimately genetic, biological factors or was behavior culturally determined, that is, learned? Mead's teacher, Franz Boas, was a staunch supporter of the idea that it was culture, not biology, that molded human behavior. A strategic phenomenon in this controversy was the problem of stress in adolescents. Thus, Boas wanted his gifted pupil to study adolescence

> to test out, on the one hand, the extent to which the troubles of adolescence (. . .) depended upon the attitudes of a particular culture, and on the other hand, the extent to which they were inherent in the adolescent stage of psychobio-

> logical development with all its discrepancies, uneven growth, and new impulses. (Mead, 1972, p. 137)

Mead's *Coming of Age in Samoa* is sufficiently known for me to refrain from repeating the details of her research. It is enough to remind ourselves of her conclusion which was that among Samoan girls, there is no question of the state of mind in Western society called "Sturm und Drang," the implication being that this psychological state is related to culture and is not the outcome of adolescent psychobiological changes.

In 1954, the American anthropologist Holmes went to Samoa. He went to the village were Mead had carried out her research, worked with many of her informants and "systematically investigated and evaluated every word that Mead wrote in *Coming of Age in Samoa* (1928), *Social Organization of Manua* (1930) and several articles" (Holmes, 1983, p. 929).

Although every fieldworker comments on the reliability of data that is already available, Holmes' seems to be the first attempt to systematically verify an earlier study in anthropology. Can it be regarded as *replication* in the strict methodological sense of the term? I think not, even though Holmes comes quite close to it. There are two flaws in Holmes' replication. In the first place, his restudy took place 29 years after Mead's original research. This period, which includes the Pacific war, did not leave Samoa untouched. Logically, the differences which Holmes noted between Mead's work and his own may be due to changes within the society. Secondly, one should realize that as a young woman, Mead could enter into a relationship with Samoan girls in a way that was impossible for Holmes who was a man. This simply means that the research instrument used by Holmes was basically different from the one used by Mead. Because human beings always differ from one another in terms of age, sex, personality, and cultural background, replication in the strict sense is simply not possible in anthropological fieldwork.

Holmes himself uses the term "restudy" for his research in Samoa, not the term "replication." What, then, is a restudy?

RESTUDIES IN ANTHROPOLOGY

The first scholar to use the term restudy was probably Oscar Lewis when he called a book based on his Mexican fieldwork *Life in a Mexican Village: Tepoztlán Restudied* (1951). Tepoztlán is a village in the Mexican state of Morelos. Robert Redfield carried out research in Tepoztlán in 1926 and 1927. His book, *Tepoztlán. A Mexican Village* (1930), is one of the first studies of what later came to be known as a "peasant society." Tepoztlán became famous only after Oscar Lewis's research there in 1943. A historian by training, Lewis later turned to anthropology. He first carried out fieldwork

among the Blackfoot Indians in the United States. In 1943 he went to Mexico to assist in aid programs aimed at Mexican Indian communities. It is in this context he began to do research in Tepoztlán because the earlier work by Redfield offered Lewis a baseline from which to measure and analyze change (Lewis, 1951, p. xiii). Yet, while working in Tepoztlán, Lewis became convinced that the picture that Redfield had painted was strongly biased. Redfield has described a harmonious community of people who worked collectively rather than individually. Lewis stressed individuality and distrust in interpersonal relations as well as conflict.

This was not the first case in anthropology in which two fieldworkers arrived at radically different pictures of the same community. In 1937, for example, the Chinese student Li An-Che had published an account of the Zuñi Indians in which he stressed elements of Zuñi thought and behavior that were very different from Ruth Benedict's rendering of Zuñi culture and personality in her well-known *Patterns of Culture* (1934). The difference between the Zuñi controversy and the Tepoztlán controversy is that Redfield and Lewis entered into a discussion about the backgrounds to these differences in perception of reality.

Today "restudy" is an acknowledged term in anthropological jargon. However, it is not at all clear what it stands for. Of course it means what is literally says, that is, the study of a society and culture already studied, but it also has a special connotation. It often refers to those studies in which the second researcher arrives at a quite different picture of the "same" society. As we have already seen, conscious efforts to systematically evaluate earlier studies are rare in anthropology since fieldworkers prefer to have their "own" society. In the case of a restudy, however, a fieldworker turns to a society or even a community that has already been studied. Does this mean that field work practice has changed? What motivates a fieldworker to carry out a restudy?

Looking at studies of the same society in which the second researcher reached more or less different results, there seem to be two categories. In the first place, there is the restudy that deliberately wishes to use an earlier study as a point of departure to analyze processes of change. What occasionally happens during the restudy is that the second fieldworker takes issue with the way in which the earlier situation has been portrayed. Lewis's restudy of Tepoztlán falls into this category. Secondly, there is the restudy which looks into hitherto neglected aspects of a society and, while taking a second look, takes issue with the previous study's characterization of that society. Li An-Che's study of Zuñi belongs to this category.

All this points to a change in fieldwork practices. Two circumstances should be mentioned. The first is that while the number of professional anthropologists has risen enormously, the number of societies not yet studied has gone down. Apart from a few Amerindian groups in South America and perhaps a

handful New Guinean societies, there are no truly blank spots on the ethnographic map. Furthermore, the absolute number of different societies and cultures has also gone down. Due to rapid social change, especially since World War II, as well as to ethnocide and even genocide, many societies and cultures have disappeared. It is no longer possible for fieldworkers to find a society of their own; they simply have to work on a society that has already been the object of much research.

The second change is a change in perspective. The emphasis of the first generation of fieldworkers was on the uniqueness of sociocultural systems found on the spot. They were impressed by the complexity of sociocultural systems and the interdependence of sociocultural elements. Contrary to a previous, strongly evolutionary approaches, they tried to explain a sociocultural system not in terms of its history but in terms of how it functions here and now.

The historical perspective was reintroduced mainly during the 1960s, although there are earlier examples as is evidenced by the work of Oscar Lewis. Older accounts, often written by the first generation of anthropological fieldworkers, acquired a new meaning: They offered the possibility of analyzing change. After all, one of the problems of studying long-term processes confronting fieldworkers is the paucity or even the absence of reliable historical data.

In a few cases, fieldworkers went back to the place where they themselves had carried out research before World War II (see Firth, 1936, 1959). In other cases, pupils of first-generation fieldworkers went back to the community where their teachers had pioneering work (see, for instance, Adams, 1972; Gillin, 1936). In still other cases, fieldworkers revisit the peoples about whom interesting studies have been written (see Stearman, 1987; Holmberg, 1950).

In many cases, consistency between earlier and later research dominates. In a few notorious cases, however, the points of view are so radically different that the restudy gave rise to a controversy over the question of who was right and who was wrong. The first clear case of such a controversy to reach the anthropological journals was probably that of the Zuñi which pitched Ruth Benedict against Li-An Che. The latest is the Mead–Freeman controversy regarding Samoa. How can we explain the differences between fieldworkers describing the same society and culture? By and large, there are four possibilities (see Heider, 1988). The differences may be due to:

1. Cultural variation within one society
2. Variation in time
3. Errors of observation or analysis
4. Differences in research perspective

A combination of these factors is, of course, also possible. In the next section, I elaborate on these four possibilities.

WHY FIELDWORKERS DISAGREE

Cultural Variation Within One Society

One explanation of disagreement between two fieldworkers lies in intrasocietal variation. Although they appear to have carried out research in the "same" society," what they actually describe are different *versions* of the same society. However, this explanation covers only a minority of the ethnographic controversies. In many cases of restudy, research is carried out in exactly the same community (in Samoa, Holmes even used the same informants as Mead!). This choice is usually a deliberate one because the objective of the research is to study changes that have taken place since the first fieldworker made his or her observation. Still, there is sometimes room for doubt. In 1976, there was a debate concerning the degree of nomadism among the Brazilian Nambikwara, a society that had been studied by Lévi-Strauss in 1938. Certain characteristics of their society and culture became a key elements of his ideas regarding dualism and binary opposition (see Lévi-Strauss, 1955). His portrayal of the Nambikwara was challenged by later fieldworkers such as Aspelin (1976), who argued that the Nambikwara were sedentary. However, there is a possibility that Lévi-Strauss worked with an atypical, marginal subgroup of the Nambikwara who really were nomadic (see Kloos, 1988a, pp. 73–82).

Variation in Time

Disagreement between fieldworkers can also be attributed to variations in time following from either cyclical changes or from structural change. The second fieldworker is usually quite aware of structural change; indeed, that is what he or she is trying to study. However, there are also changes of a cyclical nature. The disagreement between Mead (1949) and Fortune (1939) regarding the warlike habits of the New Guinea Arapesh seems to be related to such a cyclical change. One should appreciate the fact that the structure of many societies is related to cyclical processes requiring up to 30 or 40 years. This holds true, for example, for those New Guinean societies in which herds of pigs are slowly built up to be collectively slaughtered. Patterns of life before and after these events differ markedly.

Again, reviewing the ethnographic controversies as an explanation of disagreement, time variations do not play an important role.

Errors of Observation or Analysis

Fieldworkers do, of course, make mistakes and this would be the easiest way to dispense with ethnographic disagreements. The question of who is wrong and who is right arises especially where widely differing images of the "same" society have been put forward. When in 1983, Freeman challenged almost every statement made by Mead regarding Samoa, the one important question seemed to be whose Samoa was the right one. In a number of cases of ethnographic controversy there is a fairly simple answer. When Redfield painted his rosy picture of Tepoztlán and failed to heed more than 100 cases of crime reported to the police during his sojourn there, it is reasonable to call this an error. In the majority of cases, errors are not that simple to define. Instead, one has to conclude that the differences are due to differences in research perspective.

Differences in Research Perspective

Sociocultural systems of even small-scale, preindustrial and preliterate societies are complex. Although the ambition of anthropology, and of early anthropological fieldworkers in particular, is to apprehend a society and its culture in its totality, no fieldworker is really able to cover this complex totality. Research invariably means looking at reality from a specific point of view. This also implies selection. By far the majority of ethnographic discrepancies are due to differences in research perspective. Several categories can be distinguished. In the first place, one should remember that in anthropological field work, the dominant instrument is a human individual. Human individuals differ in their outlook in terms of age, gender, and personality, and these differences affect their observations. In the second place, one should take into account that ethnographic facts are already constructs in which the theoretical orientations of the fieldworker are as important as the actual reality. Third, one should realize that behind more or less explicit theoretical orientations, more implicit cultural assumptions held by the fieldworker color his or her observations. In the fourth place, we should not forget the more or less explicit political objectives of the scholar.

To illustrate how these four possible explanations cover specific ethnographic discrepancies, I now turn to three concrete cases.

THREE ETHOGRAPHIC DISCREPANCIES

The Pueblo Zuñi

In 1934, Ruth Benedict published her *Patterns of Culture*. It is probably one of the most widely read anthropological studies. The main theme of her book is

the ruling motivation of a culture. In this work, she contrasts the sober, highly disciplined Zuñi Indians with the exuberant Kwakiutl of the American northwest coast as well as with the Indians of the Plains.

A few years later a young Chinese scholar, Li An-Che, spent three months with the Zuñi. He saw many things of which Benedict had not taken note: Emotionally charged prayers, marital trouble, and conflict. Where Benedict had seen a highly integrated culture based on such values as equanimity, modesty, and cooperation, Li An-Che saw hidden tensions, suspicion, and ambition.

Because Benedict and Li An-Che had carried out their research in the same *pueblo*, cultural variation as an explanation of these differences can be ruled out. Because there were only a few years between the two periods of research, change as an explanation can also be ruled out.

Was one of them simply wrong? It is interesting to see that the discrepancy between the two representations of Zuñi is not restricted to the work of Benedict and Li An-Che. Observers of Zuñi (and to some extent also of the closely related Hopi) seem to stress either overt harmony or hidden tension.

The discrepancy disappears when one realizes that scholars look at Zuñi society and culture from different perspectives. The Zuñi live in a harsh environment (present-day Arizona). They are cultivators who have to rely on irrigation and on an unreliable rainfall. Benedict described the values of the cultural program which the Zuñi developed to cope with their environment; Li An-Che described the costs that have to be paid by individuals in order to fit into a rigid program that, nevertheless, makes it possible for them to survive. Put that way, it becomes clear that the two perspectives cannot be regarded as mutually exclusive. This means that on their own terms Benedict and Li An-Che were both right—they only failed to be explicit about the limitations of their statements regarding Zuñi culture (see Bennett, 1946).

Tepoztlán

At first sight, the Tepoztlán discrepancy seems to be of the same kind as the Zuñi one. As already said, Redfield presented this Mexican village as relatively homogeneous, isolated, and functioning smoothly. According to him people were content and well-adjusted. Lewis on the other hand, observed the underlying individualism, the lack of cooperation and a "pervading quality of fear, envy, and distrust in inter-personal relations" (Lewis, 1951, p. 429).

Discussing the discrepancies, Redfield stressed the complementarity of the two accounts. He saw in contrasting accounts of a community a means to better understanding and pleaded in favor of the deliberate construction of complementary descriptions (Redfield, 1960, pp. 132–133). He argued that the difference in the case of Tepoztlán was due to a "hidden question": "The

hidden question behind my book is "What do these people enjoy?" The hidden question behind Dr. Lewis' book is, "What do these people suffer from?" (Redfield, 1960, p. 136).

Lewis also seemed to believe that they had studied the same village and that the differences were due to a difference in perspective. Lewis, in particular, did not want to present a "neutral" picture of the village. Politically more committed than Redfield, he wrote:

> It seems to me that concern with what people suffer from is much more important than the study of enjoyment because it is more productive of insights about the human condition, about the dynamics of conflict and the forces for change. . . . To stress the enjoyment in peasant life [as Redfield did] is to argue for its preservation. (Lewis, 1961, p. 179)

But did they really describe the same village, albeit from a different perspective and for a different purpose? There is serious doubt here. Redfield arrived in Tepoztlán just after the atrocities of the revolution. Life in Tepoztlán slowly returned to "normal." Between Redfield's stay there in 1926–27 and Lewis's fieldwork in 1943–44, Tepoztlán's population rose rapidly. Lewis pays attention to the growth in population but gives little heed to its possible consequences. Several of the differences between Redfield and Lewis can be explained in terms of population growth (disappearance of communal land and communal labour, growing antagonism between landowners and landless people, distrust in interpersonal relations). I think Coy is largely right when he said,

> that both accounts were legitimate in the context of the situation ruling at the two periods of time . . . at some time between 1929 and 1935 economic and demographic changes in Tepoztlán favored the expression of different behaviour patterns from that of the exhausted survivors of the revolution. (Coy, 1971, p. 56)

There is little doubt that Redfield misrepresented Tepoztlán to some extent but Lewis's account is biased too. Yet, the differences between the two accounts are also due to changes taking place during the intervening period of 17 years.

The Trobriand Islanders

One of the first modern ethnographic accounts is Malinowski's *Argonauts of the Western Pacific* (1922). This book was followed by several others and by many articles, totaling some 2,500 pages covering all aspects of the life of the inhabitants of a small Melanesian island. Malinowski described production

and trade, kinship and marriage, magic and religion, and so on and so forth, in great detail. How true was his account of Trobriand society?

In 1971, more than 50 years after Malinowski, the American anthropologist Annette Weiner carried out fieldwork in a village only a few miles away from where her famous predecessor had lived. Weiner went to the Trobriand Island to study the influence of tourism on Trobriand art. However, on the very first day of her stay, a number of women took her to a large ceremony in which only women participated, and from that moment she knew "that women were engaged in something of importance that apparently had escaped Malinowski's observations. The women's mortuary ceremony was a Pandora's box; it opened up the whole question of relationship between men and women" (Weiner, 1976, p. 8).

What emerged from this Pandora's box goes far beyond the limits of this article. Women occupy a strategic position in this society. Trobriand descent is matrilineal, which means that inheritance and succession to political office go from a man to his sister's son. However, Malinowski did not pay attention to women's mortuary ceremonies although there are one or two passing references to them in his publications.

Malinowski saw Trobriand society predominantly in terms of males. This bias was his own and it was probably connected to his 19th-century, Central European background (see Ellen, Gellner, Kubica, & Mucha, 1988). His bias can hardly be explained in Trobriand terms: Although men did not participate in women's ceremonies, they were keenly aware of their significance. Whereas Malinowski, as a man, could hardly participate in these ceremonies, Weiner, as a woman, could not avoid them—and the result was that she presented an entirely different picture of the relationship between men and women, between husband and wife, brother and sister, and so on.

The difference between Malinowski's and Weiner's account of Trobriand society cannot be explained in terms of subcultural variation or of social change. It is the difference between the fieldworkers and the awareness of the existence of a male bias in anthropology that is responsible for the contrast.

REPLICABILITY AND THE NATURE OF ANTHROPOLOGICAL KNOWLEDGE

To assess the different views arrived at by two or more anthropological fieldworkers two preliminary remarks should be made: one about the nature of the object of study, or perhaps rather about the anthropological approach to sociocultural reality, the other regarding epistemological assumptions.

Since the rise of science in its modern form, the dominant goal of scientific research has been the formulation of general statements or scientific laws. Such a way of approaching reality owes much to the awareness that knowl-

edge in this form can be readily used to influence and possibly even control hitherto uncontrolled processes. Philosophers like Bacon (*Novum Organon*, 1620, Aphorisms II and IV) and Descartes (*Discours de la méthode*, 1637, Part 6) were instrumental in forcefully expressing this new approach to knowledge. This led to a scientific practice in which isolating critical variables and quantifying them became key strategies. Among the various forms of empirical research, the laboratory experiment stands out as the apogee of this approach.

There is a long distance between the laboratory experiment and anthropological field research for at least two reasons. First of all, in field work the degree of control, crucial for an experiment, is almost zero. Second, anthropological fieldworkers rarely seek to determine whether or not there is a relationship between two variables only. Fieldworkers usually deal with a sociocultural system in its comprehensive complexity and therefore have to handle large numbers of variables simultaneously. Even if these variables could be isolated and measured (which is impossible in practice), their sum total is regarded as less than what sociocultural life is supposed to be; in other words the idea of *holism* is very strong in anthropology (cf. Peacock, 1986).

This brings me to my second point. One of the current debates in anthropology concerns the nature of knowledge. A rather firmly entrenched positivist tradition is now being vigorously challenged by a constructivist point of view (see Kloos, 1988b).

Few, if any, anthropologists would subscribe to the extreme view that knowledge is part of an independent reality and thus has an existence apart from the observers. Knowledge consists of statements about reality. These statements are obviously made by observers: Without observers there would be no knowledge in a form that we would acknowledge. In this sense, all knowledge is constructed. Yet according to many anthropologists, their statements do refer to an objective reality. Although the statements do not exist in reality, what they stand for does. A corollary of this assumption is that in the last analysis, only one of various rival statements can be true.

The alternative view is that reality does not dictate what is true or not. It only responds to our way of asking questions, in the natural sciences no less than in the social sciences (cf. Heisenberg, 1969, p. 57). Some answers correspond to what was expected, others have to be rejected. According to this view knowledge is inevitably a product of a reality having an existence of its own, *and* a particular method. Methods of observation cannot be seen as epistemologically neutral tools but as tools partly responsible for the ways in which our perception of reality is cast.

This places the traditional concept of reliability in a different perspective. According to received wisdom, observation and measurement may be contaminated by accidental elements, tainting the results (any textbook on research methods in the social sciences abounds with examples that illustrate this idea). Reliable methods are supposed to be free from these accidental elements. Yet to what extent can we speak of "contamination"? Could it be

that at least some of these so-called "accidental elements" refer not to accidents at all but to highly relevant parts of both method and reality? If so, removing them would result in a biased portrayal of reality. If it is true that the anthropologist is a human instrument studying other human beings (Powdermaker, 1967, p. 19), these so-called "accidental elements" should be utilized rather than removed. This step, however, makes replication problematic because fieldworkers differ very much from one another in terms of historical and sociocultural background assumptions, scientific convictions, gender, age, let alone more individual idiosyncracies.

If the insights of a fieldworker cannot really be recreated, why should we listen to him or her anyway? How far can we trust that what he or she reports corresponds to the sociocultural reality he or she purports to portray?

There seem to be at least three solutions to this problem. The first is to trace back our steps in an effort to recast the fieldworker into a neutral tool by eradicating, or at least repressing, individual and collective idiosyncracies. In my judgment this would lead us up a blind alley because it removes from our research the very essence of our being human.

The second solution is to accept an almost solipsist position that discards the traditional notion of science altogether. Description becomes autobiography. Leach came close to that position when he wrote: "the data which derive from fieldwork are subjective not objective. I am saying that every anthropological observer . . . will see something that no other such observer can recognize, namely a kind of harmonious projection of the observer's own personality" (Leach, 1984, p. 22).

Leach went on arguing that anthropological texts can be read in two different ways. They can be read as texts that are interesting in themselves, full of hidden meanings, but without assuming that "what is discussed in the text corresponds to any kind of reality" (p. 22). Or they can be read "with the set purpose of discovering projections of the author's personality, of finding a record of how he or she reacted to what was going on" (p. 22).

This implies, Leach added, that we should pay much closer attention to the personal background of the authors of anthropological works. Although Leach referred to texts about the history of anthropology, the same applies to ethnographic texts.

The third solution to the problem of credibility follows from Leach's remark. It consists of a fieldworker supplying his or her reader with the information necessary to follow and assess the process by which the results were obtained. Basically the idea is plain enough: A fieldworker should make clear what his or her basic assumptions are, what the empirical basis of his or her statements is, what procedures have been used, what choices have been made, what the limitations are, and so on. In this way, the conclusions reached can be made *credible* even though the procedures are not really replicable (cf. Glaser & Strauss, 1967, p. 228ff).

Sensible though such a solution may sound, it is not feasible. Even in a

relatively simple case like the Pons/Fleischmann experiment alluded to at the beginning of this chapter, there was so much confusion around the exact nature of the experiment that replication proved to be extremely difficult. In anthropological research where the main research instrument is a human individual, any description of its relevant characteristics would require an independent study of the researcher but, for obvious reasons, this approach is self-defeating.

Paradoxically, the way out should not be sought in replication in the strict sense of the term—that is, in a kind of research in which every effort is made to eradicate all differences except the relevant variables and to eradicate the idiosyncracies of the fieldworker. As already argued, neither is possible without eliminating what is the very nature of anthropological research. The only alternative is to accept the complexity of sociocultural life and to acknowledge the essential contribution of the idiosyncratic approach to our understanding of it. The nature of these idiosyncracies and their consequences can only be assessed where reports by at least two different fieldworkers are available. When our knowledge of sociocultural reality is based on the work of a single scholar, apart from such formal criteria as internal consistency, there is no way to judge the credibility of his or her report. It is precisely cases of contrasting and even conflicting representations of the same society and culture that offer the possibility of improving our understanding of the essentials of social life in a particular place and at a particular time. However, to be content with the idea that the mere existence of different accounts automatically adds to our understanding, as Redfield seems to have been in the case of Tepoztlán (see earlier), underestimates the potential contribution of conflicting descriptions. They should be studied, not merely accepted. They should even be deliberately *created* by sending scholars holding very different views to study the same society at the same time.

I agree with Bennett's contention, written almost 50 years ago, that the problem of conflicting interpretations of the same reality cannot be solved by simply collecting more facts and making more interpretations. Obviously, "a reflexive analysis of the meanings of the respective interpretations" (Bennett, 1946, p. 374) is needed to avoid the trap of Western dualistic thought according to which only one several conflicting points of view can be true. The real challenge lies in coming to terms with these contrasting and conflicting interpretations in the context of empirical research.

REFERENCES

Adams, K. (1972). *The Barama River Caribs of Guiana restudied. Forty years of cultural adaptation and population change.* Unpublished doctoral thesis, Case Western Reserve University, Cleveland, OH.

Aspelin, P. L. (1976). Nambicuara economic dualism: Lévi-Strauss in the garden once again.

Bijdragen Taal-, Land-en Volkenkunde, 132, 1–31.
Beals, A. R. (1970). Gopalpur 1958–60. In G. W. Spindler (Ed.), *Being an anthropologist. Fieldwork in eleven cultures* (pp. 32–57). New York: Holt, Rinehart and Winston.
Benedict, R. (1934). *Patterns of culture*. New York: Houghton Mifflin.
Bennett, J. (1946). The interpretation of pueblo culture. A question of values. *Southwestern Journal of Anthropology, 2*, 361–374.
Broad, W., & Wade, N. (1982). *Betrayers of the truth. Fraud and deceit in the halls of science*. New York: Simon & Schuster.
Brown, P. J. (1981). Field-site duplication: Case studies and comments on the "my-tribe" syndrome. *Current Anthropology, 22*, 413–414.
Coy, P. (1971). A watershed in Mexican rural history: Some thoughts on the reconciliation of conflicting interpretations. *Journal of Latin American Studies, 3*, 39–57.
Degérando, J-M. (1969). *The observation of savage peoples*. London: Routledge & Kegan Paul. (Original work published 1800)
Ellen, R. F. (Ed.). (1984). *Ethnographic research. A guide to general conduct*. London: Academic Press.
Ellen, R., Gellner, E., Kubica, G., & Mucha, J. (Eds.). (1988). *Malinowski between two worlds. The Polish roots of an anthropological tradition*. Cambridge, UK: Cambridge University Press
Evans-Pritchard, E. E. (1940). *The Nuer*. Oxford: Oxford University Press.
Feinberg, R. (1988). Margaret Mead and Samoa: Coming of age in fact and fiction. *American Anthropologist, 90*, 656–663.
Firth, R. (1936). *We, the Tikopia. A sociological analysis of kinship in primitive Polynesia*. London: G. Allen.
Firth, R. (1959). Social change in Polynesia. London: George Allen & Unwin.
Fortune, R. (1939). Arapesh warfare. *American Anthropologist, 41*, 22–41.
Freeman, D. (1983). *Margaret Mead and Samoa. The making and unmaking of an anthropological myth*. Cambridge, MA: Harvard University Press.
Gillin, J. (1936). *The Barama River Caribs of British Guiana* (Papers of the Peabody Museum No. 14) Cambridge, MA: Harvard University Press.
Glaser, B. N., & Strauss, A. L. (1967). *The discovery of grounded theory. Strategies for qualitative research*. Chicago: Aldine.
Harris, M. (1977). *Cannibals and kings. The origins of culture*. New York: Random House.
Heider, K. (1988). The Rashomon effect: When ethnographers disagree. *American Anthropologist, 90* (1), 73–81.
Heisenberg, W. (1969). *Physics and philosophy. The revolution in modern science*. London: George Allen & Unwin.
Holmberg, A. R. (1950). *Nomads of the Long Bow. The Siriono of Eastern Bolivia*. Washington, DC: Smithsonian Institution.
Holmes, L. D. (1957). *The restudy of Manu'an culture. A problem in methodology*. Unpublished doctoral thesis, Northwestern University, Ann Arbor, MI. (Ann Arbor Microfilms. No. 23.514).
Holmes, L. D. (1983). A tale of two studies. *American Anthropologist, 85*, 929–935.
Holmes, L. D. (1987). *Quest for the Real Samoa. The Mead/Freeman controversy and beyond*. South Hadley, MA: Bergin and Garvey.
Kloos, P. (1988a). *Door het oog van de antropoloog. Botsende visies bij heronderzoek*. Muiderberg: Coutinho.
Kloos, P. (1988b). No Knowledge Without A Knowing Subject. In R. Burgess (Ed.), *Studies in qualitative methodology* (Vol. I, pp. 221–241). Greenwich, CT: JAI Press.
Kuper, A. J. (1983). *Anthropology and anthropologists. The Modern British School*. London: Routledge & Kegan Paul.
Leach, E. R. (1984). Glimpses of the unmentionable in the history of British Social anthropology. *Annual Review of Anthropology, 13*, 1–23.
Lévi-Strauss, C. (1955). *Tristes Tropiques*. Paris: Plon.

Lewis, O. (1951). *Life in a Mexican Village: Tepoztlán restudied.* Urbana: University of Illinois Press.

Lewis, O. (1961). Some of my best friends are peasants. *Human Organization, 19* (4), 179–180.

Li An-Che. (1937). Zuñi: Some observations and queries. *American Anthropologist, 39*, 62–76.

Malinowski, B. (1922). *Argonauts of the Western Pacific.* London: G. Routledge and Sons.

Mead, M. (1928). *Coming of age in Samoa.* New York: W. Morrow.

Mead, M. (1949). *The Mountain Arapesh* (Anthropological Papers, Vol. 41, Part 3). New York: American Museum of Natural History.

Mead, M. (1972). *Blackberry winter. My earlier years.* New York: William Morrow.

Peacock, J. L. (1986). *The anthropological lens. Harsh light, soft focus.* Cambridge, UK: Cambridge University Press.

Powdermaker, H. (1967). *Stranger and friend. The way of an anthropologist.* London: Secker & Warburg.

Radin, P. (1933). *The method and theory of anthropology.* New York: McGraw-Hill.

Redfield, R. (1930). *Tepoztlán: A Mexican village.* Chicago: University of Chicago Press.

Redfield, R. (1960). *The little community and peasant society and culture.* Chicago: The University of Chicago Press.

Slotkin, J. S. (Ed.). (1965). *Readings in early anthropology.* Chicago: Aldine.

Stearman, A. M. (1987). *No longer nomads: The Siriono revisited.* Lanham: Hamilton Press.

Stocking, G. W. (Ed.). (1983). *Observers observed. Essays on ethnographic fieldwork. History of anthropology* (Vol. 1). Milwaukee: The University of Wisconsin Press.

Urry, J. (1984). A history of fieldwork. In R. F. Ellen (Ed.), *Ethnographic research. A guide to general conduct* (pp. 35–61). London: Academic Press.

Waal Malefijt, A. de. (1974). *Images of man. A history of anthropological thought.* New York: Knopf.

Weiner, A. B. (1976). *Women of value, men of renown. New perspectives in Trobriand exchange.* Austin: University of Texas Press.

CHAPTER 3

A Process Model Of Replication Studies: On The Relation Between Different Types Of Replication*

Marinus H. van IJzendoorn
Center for Child and Family Studies
Leiden University
The Netherlands

INTRODUCTION

In this chapter we argue that the "received view" of replication as the exact or algorithmic repetition of an original study has become obsolete. Exact repetition only appears to be a boundary case of a series of more or less "varied" replications, that is, replications that systematically vary one or more parameters of the original study to see whether its outcome remains stable or changes in a predictable way. Only in case of failure to produce any predicted outcome through a series of varied replications does the suspicion of an irreproducible effect arise, and therefore the need for a more or less exact replication. From this "constructivistic" perspective it is to be expected that exact replications are rather rare, although methodologically replicability is considered to be one of the most important cornerstones of science. The production of so-called "objective knowledge" (Popper, 1980) would be dependent on the intersubjectivity of observations in accordance with the hypothesis. Intersubjectivity of observations means that observational reports would only be taken seriously if they could in principle be reproduced by any competent researcher. A theory would only be considered falsified in case of a replicable observation contradicting one of the implications of that

*This chapter was partly written during a stay at the National Institute of Child Health and Human Development, Bethesda, MD, as a senior Fulbright fellow. Preparation of this chapter was supported in part by the Netherlands Organization of Scientific Research (NWO) through a Pioneer grant. The author would like to thank Michael Lamb for his constructive criticism on an earlier version of this chapter.

theory (Popper, 1980). We argue here that the central idea of replicability should not be interpreted literally, that is, as the necessity of exact replications. Only if we take a more constructivistic perspective does it become possible to notice replication studies being carried out as part and parcel of everyday research. To emphasize its central role in methodology as well as in research practice, we present a process model of replication in which different kinds of replication efforts are systematically interrelated. Against the background of this process model of replication it becomes clear that research programs allow for continuous replications of their data and results, although isolated studies may remain unreplicated.

LACK OF EXACT REPLICATIONS AND THEIR IMPACT

It seems as if a tremendous lack of replication studies exists in the natural sciences as well as in the behavioral and social sciences. A few examples may illustrate this point for the behavioral and social sciences. The famous Harlow (1958) experiments with primates raised by an "iron" mother or by a "cloth" mother never have been exactly replicated, although their impact on the theory of primary attachments in animals and in humans have been substantial (Paul & Blumenthal, 1989). According to L.L. Ainsworth (1984) we even should doubt the possibility of replicating the experiments because their design has been described with too little detail. Maybe more influential still is J.B. Watson's conditioning experiment with Little Albert. This experiment showing the force of conditioning in human learning has never been replicated, neither by Watson himself nor by his numerous students and followers. Nevertheless there is some reason to believe that the experiment's exact replicability should be doubted. Its design and procedures have been described in a superficial way, and with sometimes contradicting details. Furthermore, the experiment with Albert was preceded by many failures to show that operant conditioning played a major role in human learning (Samelson, 1980).

Sometimes very influential studies are replicated after a long period of time. Examples are: Burt's twin studies on the heredity of intelligence, Mead's Samoa study focusing on the cultural context of adolescence (see Chapter 2), and Efron's media-sociological study on the political biases in television news about the candidates in an American presidential election campaign. On the basis of a detailed and scrutinous (secondary) analysis of Burt's data set, Kamin (1974) concluded that Burt had at least partially faked his data and had invented imaginary subjects and experimenters. Freeman (1983) showed that Mead probably had not taken into account all relevant and available information, and had sketched a somewhat unidimensional and optimistic picture of an idyllic tropical island. Stevenson and his colleagues (Stevenson, Eisin-

ger, Feinberg, & Kotok, 1973) reanalyzed part of Efron's raw data after coding these data with a more detailed and reliable coding system. They concluded that Efron's theory could be falsified by her own data.

These replications seemed to disconfirm at least partially the original studies, but they could not restrict their influence. On the contrary, the replication studies provoked sometimes heated discussions about the merits of replication (Holmes, 1983; Kloos, 1988), about the meaning and possibility of "real" replications, and about their impact on scientific discourse (Collins, 1985).

It would seem, then, that replication studies do not seriously influence discourse about the value of certain theories and their underlying data, and that the request for more replications remained unanswered. Caplow (1982) found that about 1% of the sociological studies is replicated. Shaver and Norton (1980) observed that replications are nearly absent in educational research: They found that only 14% of 151 research projects reported on in the *American Educational Research Journal* could be considered replications. Bahr and his colleagues (Bahr, Caplow, & Chadwick, 1983) traced 300 replications (half of which are from psychology) in the Social Science Citation Index over a period of almost 10 years (1973–1981), looking for the concept of replication or its equivalents in the title of the papers. Studies that refer explicitly to the concept of replication may be often considered exact replications. It seems, therefore, that only a small amount of behavioral and social research is being exactly replicated, and that even very influential studies remain unreplicated in the strict sense. Furthermore, if a study has been replicated, and results diverge from the original outcome, more discussion about the merits of the replication and its author(s) seems to be provoked than about the original study and the theory by which it was originated. On the one hand, replications are being considered the cornerstones of modern science, but on the other hand they seem to be carried out seldomly and they seem to have little impact. How can this paradox be explained? We think that a naive definition of replication as exact repetition of a single original study is the reason replications remain unnoticed.

REPLICATIONS IN THEORY: ALGORITHMIC REPLICATIONS

Many researchers think that studies can be replicated in an exact way (Collins, 1985). These so-called exact or algorithmic replications would be carried out strictly according to the technical rules of the original author(s); an identical outcome should consequently be attained (Collins, 1975). From a methodological perspective, however, such replications cannot exist, simply because technical guidelines will always leave room for minor variations, and because spatiotemporal parameters will have changed anyway (Musgrave,

1975). Whether the same outcome will be produced will, therefore, depend on the robustness of the original results in the presence of minor variations in the research design, and of spatiotemporal variations. A replication will be called an exact replication if it is essentially similar to the original study. Every comparison, however, is being made from a certain perspective (Popper, 1980, p. 421). As we have said, replications necessarily change the design of the original study, but on the basis of available background knowledge it will be assumed that the changing parameters will not have any influence on the outcome. For example, in a replication of a study on the conditioning of emotions, spatial parameters are being considered irrelevant; that is, the conditioning study may be carried out in different parts of the world and is nevertheless supposed to yield the same kind of results.

Our background knowledge, however, is not infallible, and can be considered "true" only for the time being (Popper, 1980). If the original study and its replication show divergent results, differing parameters constitute possible alternative hypotheses to interpret the divergencies. With respect to cultural-anthropological research, for example, Kloos (Chapter 2) composed a list of important parameters: Replications usually are being carried out in another social group, historically in another society, from a different cultural and personal background of the researcher, and often also from a different theoretical perspective. Differences between Mead's (1961) and Freeman's (1983) study, for example, may be explained by the differences in social group in which both researchers carried out their observations and interviews, differences in historical period (before and after the second World War), and differences in theoretical perspective (cultural determinism versus sociobiology). Last but not least, it certainly made a difference whether a woman or a man carried out interviews on sexual issues (Chapter 2).

If little knowledge about a certain domain exists, and if therefore our background knowledge is still minimal, alternative hypotheses about the causes or reasons for divergent replication results will be mushrooming. In uncharted territory small differences in age of subjects, in apparatus, in physical environment, in psychological atmosphere, and in the experimenter's personality may imply cumulative differences in outcome (Smith, 1970). For example, research on attachment relationships between a caregiver and an infant is usually based upon the Strange Situation procedure (Ainsworth, Blehar, Waters, & Wall, 1978). This procedure consists of a series of infant's separations from and reunions with the caregiver in a strange environment, as well as a confrontation with an unknown person who tries to interact with the infant. This procedure is often used in research on the effects of infant day care and other practically relevant issues, but it is not clear to what extent variations in playroom, play material, personality, and sex of experimenter and "stranger," and the order in which the procedure is carried out during a measurement session influence coding results. In the influential Bielefeld

study (Grossmann & Grossmann, 1990), the large playroom with the life-sized play material may be hypothesized to be one of the causes of the infants' seemingly avoidant behavior to the parent when they enter the room after being away for 3 minutes (Sagi, Van IJzendoorn, & Koren, 1989). It also has been suggested that the overrepresentation of resistantly attached infants in Israeli kibbutzim may partly be explained by the stressful sociability test that preceded the Strange Situation (Sagi et al., 1989). Because a strong background theory about essential ingredients of the Strange Situation is lacking, many alternative hypotheses are possible in case of diverging attachment classification distributions.

In sum, strictly seen it is impossible to copy an original study in the strict sense of the word; necessarily some potentially relevant parameters will inevitably change, and our background knowledge determines whether these changes are essential or not. The stronger our background knowledge is, that is, the more frequently confirmed or at least not falsified this knowledge is, the more exact our replications may be considered to be. If background knowledge implies that spatiotemporal parameters are irrelevant in producing a certain effect or outcome, the experiment can more easily be replicated: In this case, we do not have to control for evasive spatiotemporal variables, as these variables cannot influence our evaluation of the identity of the original and the replication study.

THE PRACTICE OF REPLICATION: A CONSTRUCTIVISTIC VIEW

The constructivistic view on replication studies is based on sociological research in scientific laboratories. Through participant observation in eminent laboratories, and through in-depth interviews with leading scientists, a description of the practice of replication has been developed that contradicts "received views" of the replication process. Research on "gravitational radiation" (Collins, 1975), on the construction of the "ETA laser" (Collins, 1985), and on conditioning of "planarian worms" (Travis, 1981), for example, has been the object of sociological studies. The results of these studies can be summarized as follows: First, interviews show that exact replications do occur only very infrequently in the natural sciences, and they constitute a low-status activity (Mulkay, 1986). Researchers are much more inclined to change some parameters of the original experiment systematically, so as to be able to discover new applications, and they are much less inclined to make a "carbon copy" of the study. The outcome of the original study is accepted—until the contrary has been proven. A falsification appears to be taken into consideration only if an outcome repeatedly cannot be established, although the variations on the original theme seem to be minor. Exact or algorithmic

replications should be considered a boundary case of "varied" replication. Starting with rather large variations of the design of an important study, researchers first try to add something new to the original outcome, for example by showing that the same effect can be produced in some other species or with another chemical substance. If such larger variations do not yield the expected results, more refined variations are being practiced, until some doubt about the validity of the original study arises. In the end, "exact" replication is applied to (dis-)confirm the doubts, and to check the assumptions of the varied replications. In sum, these social studies imply that many scientists feel that exact replications may be carried out, but usually are irrelevant for scientific progress.

Participant observations in laboratories also show that exact replications are seldom carried out, and that every replication is subject to a negotiation process on the relevance of changed parameters for diverging results (Collins, 1985). In other words, replications are not accepted without discussion about the relevance of changed parameters for the production of (un-)expected results. The so-called "enculturation model" is assumed to be a more adequate description of replication than the algorithmic model. The enculturation model emphasizes the process of socialization of the researchers who implicitly learn to differentiate between relevant and irrelevant parameters. Complicated measurement procedures are not considered to be replicable by just reading a pertinent paper. Because of lack of space, in published papers essential details about the procedures are being left out; sometimes the researcher intentionally gives a global description to defend him- or herself against premature and incompetent replication, or to prevent colleagues from producing new facts before he or she finds time to "discover" them. Researchers who want to apply the procedure are urged to contact their colleague for further information, and maybe even for training through which unreported presuppositions and details are being transferred. Until such enculturation has taken place, a researcher who cannot replicate a certain result may expect to be accused of being not competent enough to carry out the experiment (Travis, 1981).

In the case of attachment theory, the same enculturation process appears to exist for procedures like the Strange Situation (Ainsworth et al., 1978), and the Adult Attachment Interview (Main & Goldwyn, in press). A training of several weeks in one of the American research centers is considered necessary for the reliable and valid coding of the observations or interviews, and therefore for a plausible and persuasive contribution to the international discourse on attachment. When central theses of attachment theory are in danger of being falsified by "untrained" researchers, this "incompetence" and lack of enculturation will be explicitly used against the "dissident" (see, for an example, Waters, 1983). On the one hand, such arguments appear to belong to the category of *ad hominem* arguments (Fogelin, 1987), that should be

forbidden in rational scientific discourse; on the other hand, however, some measurement procedures can be so complicated that an intensive training indeed is necessary to get insight into essential elements, to be able not only to code the material reliably, but also in agreement with the investigators who constructed the procedure. Sometimes it is hard to know which procedural variations indeed are relevant and which are irrelevant, unless the "novice" is being instructed in personal training sessions with the constructors. Background knowledge cannot always be made explicit as Popper (1980) assumed. We should leave room for "tacit knowledge" (Polyani, 1973) that can be crucial for the adequate application of complicated measures, not only in the domain of the natural sciences but also in that of the behavioral sciences.

Accepting that background knowledge always remains fallible, and partly implicit, and that, therefore, the application of measurement procedures is subject to an enculturation process the effectiveness of which can be doubted, the possibility of a so-called "experimenter regression" has to be considered. This paradox means that replication logically can not be the ultimate test of a statement's truth. Replication is a matter of trained application of measurement procedures that cannot be described algorithmically. Therefore, the application of the procedures always can be criticized, and, in fact, a further test would be necessary to see whether the quality of the replication meets the accepted standards or not. But this test would also be based on a complicated measurement procedure, and would therefore again be subject to criticism. The "experimenter" regression could be stopped if the final decision about the adequateness of a measure's application would be taken by the investigator who constructed the measure. Such a short circuit, however, would imply methodological solipsism: Every researcher would in the end have the right to decide about the truth value of his or her own propositions. Constructivists, however, locate the decision on a social level: The scientific forum—or at least its most powerful part—is supposed to finally decide about the validity of the replication. Collins (1985) and Travis (1981) describe this process of consensus construction as a rhetorical discourse resulting in a paradigm shift or continuation (Kuhn, 1962). Replications are social constructions; they are the final product of a discourse on what is considered adequate and relevant in a certain domain of research.

THE PARADOX OF REPLICATION AS NEGOTIATION

The constructivistic argumentation for replications as the result of negotiations is elegant. According to constructivists, participant observation studies in different domains of natural science have shown that replications can seldom be carried out algorithmically, and that replication results are consid-

ered "true" only after intensive debate about relevant parameters. This view on replication appears to demystify the idea of exact replications in the natural sciences, and in this respect seems to close the gap between the social and the natural sciences. If we apply the constructivistic sociology of science on its own methods and results, however, a paradox seems to be implied (Mulkay, 1984). Through several replications it was established that replication studies constitute contingent social constructions, and can never be considered as unproblematic validations of knowledge claims. The same holds true for constructivistic replication studies: In these cases, too, rhetorics are being used to make clear that several different replications did only differ in irrelevant aspects, and, therefore, should be seen as real replications. Metastudies should not be exempted from the constructivistic interpretation of replication. And if every researcher—or better still, every forum of researchers—would construe his or its own interpretation of reality and, therefore, his or its own idea of replication, why should we take the constructivistic view more seriously than the view of the researchers in the field?

Although constructivism criticized the concept of exact replication, and showed the relativity of replication outcome, this theory of science did not succeed in rejecting replication as a regulatory criterion for science. Replications can always be heavily discussed because of changing parameters explaining divergent results, but replication studies remain a very strong motive for the spiral of ever more detailed discussions about knowledge claims, and the studies that constitute their foundation. For example, the constructivist Collins (1985) still concludes that "Replicability . . . is the Supreme Court of the scientific system." The simplistic view on replication as the ultimate criterion for the validity of the original study, however, should be replaced by the idea that all results of scientific research, and, therefore, also the replication results, can only get their impact on scientific developments through the medium of discourse. Replications may lead to refinements of original results, or to their rejection, but only as a result of debates in the scientific forum. A study that fails to replicate the original results, or that succeeds in replicating them, cannot determine the outcome of the scientific discourse. The possibility of replication, however, appears to be a condition sine qua non if a study is to be taken seriously in the discourse. That is, every research project should in principle be replicable; if this is not the case, the study should not be taken into account by the forum (Popper, 1980). Whether a study will in fact be replicated is a different question. Only if a series of varied replications does not succeed in producing new "facts," researchers feel urged to replicate the original as exactly as possible. And in that case, it would still be difficult to know whether an exact replication is being carried out: Replications never completely succeed in controlling all potentially relevant parameters (for example, spatiotemporal parameters). Therefore, the exactness of the repli-

cation is the result of competent researchers discussing the replication against the background of available knowledge in the field.

A PROCESS MODEL FOR REPLICATION

Replications can be classified in several different ways. Lykken (1968) for example introduced the concepts of "literal replication," "operational replication," and "constructive replication," and they constitute a continuum along which ever more parameters are being varied. La Sorte (1972) developed a somewhat more extensive classification that also includes research on the validity of measures and longitudinal research (repeated measurements). Bahr et al. (1983) present the most systematic and complete classification. They differentiate four important parameters that may vary in replication studies—time, place, method, and sample—and they show that combinations of these parameters result in 16 different kinds of replication. These types of replication can be placed on a continuum from constant parameters (the exact replication) to completely different parameters. They found that about 27% of 300 replications could be classified in the category: constant sample and method, but different time and place. About 21% could be classified as keeping the sample characteristics constant, and changing all the other parameters. These two types of replication, therefore, include about half of the replication studies.

Such classifications do not take into account two important types of replication—that is, secondary analysis and meta-analysis. Secondary analysis is a kind of replication in which all parameters except the researcher and the method of data analysis are kept constant. Secondary analysis starts with the data as collected in a certain way in a certain sample. The replication of Burt's studies by Kamin (1974) in fact can be considered a secondary analysis, in which only Burt's statistics were tested against probability. If such a secondary analysis is carried out with minimal methodical and theoretical variations, and nevertheless leads to divergent results, it can be considered one of the most powerful falsifications through replication. Secondary analysis also is one of the most inexpensive and efficient types of replication, because it is based on existing data sets. One of the main barriers to secondary replication is, however, the accessibility of the original data sets. Wolins (1962) reported that 24% of 37 authors were willing to make their data sets available. Twenty-one researchers claimed that their data sets were lost. Craig and Reesen (1973) found that 38% of 53 researchers were unwilling to make their raw data available for secondary analysis (see also Miedema, 1986).

The results of the secondary analyses are sometimes rather disappointing.

Results do not always appear to be replicable, because the first researcher used other methods that now sometimes seem outdated (Bryant & Wortman, 1978). We would like to state, therefore, that secondary analysis should have the status of an intermediate type of research situated between the original study and its actual replication. If secondary analysis of the existing data set already leads to falsification of the results, further replication including data collection should be considered an intolerable waste of time and money (see Figure 3.1).

Secondary analysis can be subclassified in two categories. First, secondary analysis may include recoding of the original raw data. Stevenson et al. (1973), for example, did recode part of the Efron's material on media reports of a presidential campaign. We propose to call this type of replication the "complete secondary analysis," because two phases of processing the raw data are involved: the coding and analyzing of the data. In many studies in the behavioral sciences data are recorded on video- or audiotape; such raw data always can be used for a complete secondary analysis, at least during the time that the recordings still have a reasonable quality. Second, secondary analysis may be restricted to coded data; in this case, we propose to use the concept of "restricted secondary analysis." In this type of secondary analysis the coding system is not changed but only the methods of analyzing the data, to see whether the original results survive statistical criticism or the application of refined methods of statistical analysis (see, for example, Cronbach & Webb's [1975] unsuccessful reanalysis of a classical Aptitude-Treatment Interaction study).

If the restricted or complete type of secondary analysis shows convergent results, replications should be carried out in which new data under different conditions are being collected (see Figure 3.1). From the start, the original study will be "trusted" so much that rather significant variations in the design will be applied. Larger variations may lead to more interesting discoveries in addition to the original study, but they will be followed by smaller variations if more global replications fail to produce new "facts." This process of varied replications seems to be inherent to modern science, and to be especially characteristic for the phase of "normal" science (Kuhn, 1962). This type of replication is not algorithmically carried out, but it is tried, on the contrary, to determine the robustness of the original results for spatiotemporal and other variations, and possibly to refine the original results with additional hypotheses. If even modest variations fail to reproduce the results, a more or less exact replication is needed (see Figure 3.1).

We consider meta-analyses to be replications because these analyses test the replicability of the original study in a series of varied replications. Until recently, the integration of the results of a great number of empirical studies on a certain domain seemed exclusively to be a matter of qualitative analysis in which the reviewer structures his material intuitively, and reaches a global

conclusion. In the last decade a trend toward formalizing this qualitative process by means of meta-analyses can be observed (Glass, McGaw, & Smith, 1981; Chapter 6). Through meta-analysis a combined effect size for different studies may be computed that gives a more precise and replicable description of the outcome of a series of related studies (Rosenthal, 1991). Furthermore, a meta-analysis can trace the effects of systematic variations in design on the outcome of a series of studies, thereby locating the spatiotemporal boundaries of replicability.

In attachment theory, for example, several instructive examples of such meta-analyses can be found. Goldsmith and Alansky (1987) showed through meta-analysis that the average effect size for the relation between maternal responsiveness and infants' attachment classification appeared to be much smaller in a series of replications (with more or less strong variations in methods and samples) than Ainsworth and her colleagues (1978) had found in their original study. Although this meta-analysis was based on studies that were very divergent in quality of design and instrument, its outcome should be taken into account as a warning not to foreclose the discussion about the central theme of attachment and responsiveness. Van IJzendoorn and Kroonenberg (1988) were able to show through meta-analysis that cross-cultural differences of attachment classification distributions were smaller than the intracultural differences. Before this meta-analysis it was assumed that distributions were culture-specific, just on the basis of a few samples from Germany, Japan, and Israel. McCartney and Phillips (1988) showed that the average effect size for the relation between type of care (day care or "home-care") and quality of attachment, specifically avoidance to the caregiver, was very small, although a few isolated studies seemed to have proven a negative effect of day care on attachment. The authors also fruitfully used the variations in replication studies to analyze their influence on the relation between type of care and attachment. They showed that neither sex of experimenter nor sex and age of infant were related to the effect size, but that it did matter whether the coders were blind to the infants' membership of type of care. If the coders were not blind to group membership, effect sizes were relatively higher and pointed toward a negative influence of day care on attachment quality (McCartney & Phillips, 1988, p. 167). In this way, varied replications can be combined to get insight into the overall effect size as well as into the influences of the variations on the relations studied.

As is shown by Lytton (Chapter 6), however, meta-analyses require many decisions to be taken in several different parts of the study. These decisions may or may not be made explicit, and other meta-analysts may or may not reach the same conclusions based upon the same "raw data" but departing from different presuppositions and decision rules. For example, in a meta-analysis on parental punishment strategies the exact definition of this construct should be given, and the researcher should decide what measures are

considered to be valid operationalizations of this construct. Different meta-analysts may propose different definitions and measures, and therefore select different (sub-)samples of pertinent studies, yielding different effect sizes (Chapter 6). Therefore, meta-analyses should also be replicated, and through variations in decision rules, it should be tested how robust meta-analytic outcomes in fact are. Only after intensive discourse on implicit and explicit decision rules—against the background of available knowledge (temporarily) considered true—the scientific forum may decide on one or the other outcome, and consider this result as (provisional) part of true background knowledge (see Figure 3.1).

CONCLUSIONS

Although many researchers and philosophers of science think that replications represent an important cornerstone of science, much confusion exists as to what we should consider replication studies to be, and what kinds of replication efforts can be distinguished. In this chapter, we tried to show that replications should not be narrowed down to the rather small category of exact or algorithmic replications. Usually replications consist of variations upon the theme of the original study, and the scientific discourse focuses upon the question in what sense it may theoretically be supposed that varied

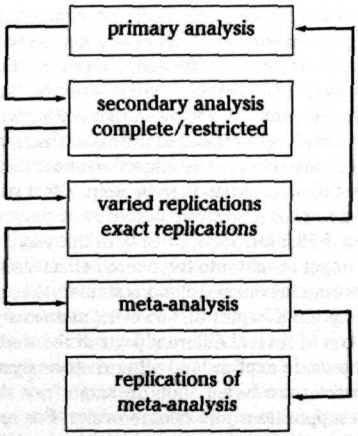


Figure 3.1. A process model of different types of replications.

replications produce results converging with those of the original study. In a broader interpretation of replication, we should not only consider the partial or complete repetition of the original study through secondary analysis or through collecting new data as "real" replication, but also meta-analysis of a series of related studies. Through meta-analysis it can be tested what variations in relevant parameters influence the results, and whether the outcome of the original study indeed can be considered to be replicated—not in an isolated replication study, but in a series of varied replications. In our process model of replication, the relations between the different kinds of replication studies have been shown, and the discursive context of replications has been emphasized. Replicability is a necessary condition for every study that is intended to have an impact on the scientific discourse, but an exact replication is only a boundary case indicating that varied replications did not yield fruitful discussions about the theory to which the original study contributed.

REFERENCES

Ainsworth, L. L. (1984). Contact comfort: A reconsideration of the original work. *Psychological Reports, 55,* 943–949.

Ainsworth, M. D. S., Blehar, M. C., Waters, E., & Wall, S. (1978). *Patterns of attachment.* Hillsdale, NJ: Erlbaum.

Bahr, H. M., Caplow, T., & Chadwick, B. A. (1983). Middletown III: Problems of replication, longitudinal measurement, and triangulation. *Annual Review of Sociology, 9,* 243–264.

Bryant, F. B., & Wortman, P. M. (1978). Secondary analysis. The case for data archives. *American Psychologist, 33,* 381–387.

Caplow, T. (1982). La répétition des enquetes: Une methode de recherche sociologique. *L'Année Sociologique, 32,* 9–22.

Collins, H. M. (1975). The seven sexes: A study in the sociology of a phenomenon, or the replication of experiments in physics. *Sociology, 9,* 205–224.

Collins, M. H. (1981). Son of seven sexes: The social destruction of a physical phenomenon. *Social Studies of Science, 11,* 33–62.

Collins, H. M. (1985). *Changing order. Replication and induction in scientific practice.* London: Sage.

Craig, J. R., & Reese, S. C. (1973). Retention of raw data: A problem revisited. *American Psychologist, 28,* 723–730.

Cronbach, L. J., & Webb, N. (1975). Between-class and within-class effects in a reported Aptitude x Treatment interaction: Reanalysis of a study by G.L. Anderson. *Journal of Educational Psychology, 67,* 717–772.

Fogelin, R. J. (1987). *Understanding arguments. An introduction to informal logic.* New York: Harcourt.

Freeman, D. (1983). *Margaret Mead and Samoa. The making and unmaking of an anthropological myth.* Cambridge, MA: Harvard University Press.

Glass, G. V., McGaw, B., & Smith, M. L. (1981). *Meta-analysis in social research.* Beverly Hills: Sage.

Goldsmith, H. H., & Alansky, J. A. (1987). Maternal and infant temperamental predictors of attachment: A meta-analytic review. *Journal of Consulting and Clinical Psychology, 55,* 805–816.

Grossmann, K. E., & Grossmann, K. (1990). The wider concept of attachment in cross-cultural research. *Human Development, 33*, 31–48.

Harlow, H. F. (1958). The nature of love. *American Psychologist, 13*, 673–685.

Holmes, L. (1983). A tale of two studies. *American Anthropologist, 85*, 929–936.

Hunt, K. (1975). Do we really need more replications? *Psychological Reports, 36*, 587–593.

Kamin, L. J. (1974). *The Science and Politics of I. Q.* New York: Wiley.

Kloos, P. (1988). *Door het oog van de antropoloog. Botsende visies bij heronderzoek*. Muiderberg: Coutinho.

Kuhn, T. S. (1962). *The structure of scientific revolutions*. Chicago: University of Chicago Press.

La Sorte, M. A. (1972). Replication as a verification technique in survey research: A paradigm. *The Sociological Quarterly, 13*, 218–227.

Lykken, D. T. (1968). Statistical significance in psychological research. *Psychological Bulletin, 70*, 151–159.

Main, M., & Goldwyn, R. (in press). Interview-based adult attachment classifications: Related to infant-mother and infant-father attachment. *Developmental Psychology*.

McCartney, K., & Phillips, D. (1988). Motherhood and child care. In B. Birns & D. F. Hay (Eds.), *The different faces of motherhood* (pp. 157–183). New York: Plenum.

Mead, M. (1961). *Coming of age in Samoa*. New York: The New American Library.

Miedema, s. (1986). *Kennen en handelen. Bijdragen aan het theorie-praktijk-debat in de opvoedingswetenschap* [Knowledge of action]. Leuven: Acco.

Mulkay, M. (1984). The scientist talks back; A one-act play, with a moral, about replication in science and reflexivity in sociology. *Social Studies of Science, 14*, 265– 282.

Mulkay, M. (1986). Relication and more replication. *Philosophy of Social Science, 16*, 21–37.

Musgrave, A. (1975). Popper and "diminishing returns from repeated tests." *Australian Journal of Philosophy, 53*, 248–253.

Paul, D. B., & Blumenthal, A.L. (1989). On the trail of little Albert. *The Psychological Record, 39*, 547–553.

Polyani, M. (1973). *Personal knowledge. Toward a post-critical Philosophy*. London: Routledge & Kegan Paul.

Popper, K. R. (1980). *The logic of scientific discovery*. London: Hutchinson.

Rosenthal, R. (1991). *Meta-analytic procedures for social research* (rev. ed.). Newbury Park, CA: Sage.

Sagi, A., Van IJzendoorn, M. H., & Koren, K. (1991). *Primary appraisal of the Strange Situation: A cross-cultural analysis of preseparation episodes. Developmental Psychology, 27*, 587–596.

Samelson, F. (1980). J.B. Watson's little Albert, Cyril Burt's twins, and the need for a critical science. *American Psychologist, 35*, 619–625.

Shaver, J. P., & Norton, R. S. (1980). Randomness and replication in ten years of the American Educational Research Journal. *Educational Researcher, 9*, 9–15.

Smith, N. C. (1970). Replication studies: A neglected aspect of psychological research. *American Psychologist, 25*, 970–975.

Stevenson, R. L., Eisinger, R. A., Feinberg, B. M., & Kotok, A. B. (1973). Untwisting the NewsTwisters: A replication of Efron's study. *Journalism Quarterly, 50*, 211–219.

Travis, G. D. L. (1981). Replicating replication? Aspects of the social construction of learning in planarian worms. *Social Studies of Science, 11*, 11–32.

van IJzendoorn, M. H., & Kroonenberg, P. M. (1988). Cross-cultural patterns of attachment. A meta-analysis of the Strange Situation. *Child Development, 59*, 147– 156.

Waters, E. (1983). The stability of individual differences in infant attachment. Comments to the Thompson, Lamb, and Estes contribution. *Child Development, 54*, 516–520.

Wolins, L. (1962). Responsibility for raw data. *American Psychologist, 17*, 657–658.

CHAPTER 4

Constructivism And Pragmatism: How To Solve The Problematic Relation Between Methodology And Epistemology In The Debate About Replication

Siebren Miedema and Gert J. J. Biesta
University of Leiden
Leiden, The Netherlands

INTRODUCTION

There has to be replication! If we take the statements of scientists and methodologists there can be no doubt of the need for us to be able to replicate original research. In the discussions, however, one distinguishes several connotations accorded replicability (cf. Lykken, 1968; Mulkay & Gilbert, 1986; Chapter 3). Complete replicability of research ("perfect replication"; "exact replication") was, and is still seen as, the ideal situation, because it provides the strongest guarantee for the *reliability* of research results. However, there is a lot of doubt as to whether one can speak of an exact replication, in the narrow sense of the word, outside of the laboratory setting. In trying to find a solution for this problem the concept of replication has been *broadened* in terms of varied replications (cf. Chapter 6). This means that almost always one or more parameters, such as time, place, method, and sample will be varied. It is Hunt's opinion that if replicating experiments are to be "powerful tests" they should differ from the original experiments in one or more aspects (Hunt, 1975). In the contemporary discussions about inquiry, replicability continues to be a regulative criterion for scientific research, notwithstanding the change from perfect replication to varied replication.

Replicability is used in the discussions as a methodological criterion. However, it is not clear whether replicability is only *a methodological criterion*,

that is, that the demand for replicability only means that the genesis of the research results be made traceable for other researchers or that it is *a methodological criterion with epistemological pretensions* or presuppositions, that is, replicability is a guarantee that the research effects tell us something about reality and that they are "true" in the sense of the correspondence theory of truth. In this chapter we first want to examine which of the two criteria distinguished above forms the basis for the notion of replicability actually used in the ongoing scientific debate. On the basis of recent literature we argue that notwithstanding the Kuhnian turn in the philosophy of science (cf. Kuhn, 1970), epistemological arguments still play a decisive role in legitimizing the results of research in terms of replicability. Given the fundamental critique of logical empiricism it is a generally accepted idea that the process of acquiring knowledge is not simply a naive realistic mirroring of reality. This critique, however, has not yet led to a rejection of the "subject–object" scheme which underlies logical empiricism. Philosophers of science—with Popper to the fore—have placed method as a neutral link between reality and knower, between subject and object. In this view method becomes the argumentative cornerstone for scientific discourse. The naive mirror view based on the idea that "all individuals in principle have the same observation apparatus at their disposal" (Kunneman, 1986, p. 21) is merely substituted for a more "sophisticated" *variant* of what is still basically the same mirror view. The connotations accorded such methodological criteria, reliability, and validity, for instance, are still closely connected with this classical dual "subject–object" scheme.

To develop our critique, we first refer to the changes in content that have taken place in the conceptualization of the notion of replicability and the use of replication in scientific practice. Then we will clarify how the epistemological scheme outlined earlier is still present as a hidden factor in the methodological descriptions of replicability. In our subsequent contribution we argue that the methodological interpretation of the demand for replicability as formulated in constructivism is a tenable one. Nevertheless, in respect to epistemology, the presupposition of the dual "subject–object" scheme still plays a decisive role in constructivism. On the basis of the work of John Dewey, we put forward some suggestions for solving this problem and for constructing a more consistent view on the relation between epistemology and methodology.

VARIETIES OF REPLICATION

Conceptualizations of Replication

Traditionally the canon of replicability is central to (natural) science. The results of research are based on observations and—so it is held—it is crucial

that the observations, which are the basis of the scientific statement being made, are in principle repeatable by anyone (cf. Heisenberg, 1975, cited in Mulkay & Gilbert, 1986).

A lot of lip service is paid to this widespread idea, but it is remarkable that there is hardly any theoretical reflection on this concept of replication in standard textbooks. Draaisma (1989) points out that if one tries to locate the concept of replication in Nagels's *The Structure of Science* (1961) (the codex and canon of the traditional philosophy of science) one does so in vain. There is also no entry in the *Philosopher's Index*. De Groot's (1961) methodological handbook for the behavioral sciences is an exception, however. In his *Methodology*, the matter of replication (in research) is mentioned in five places, according to the index. For De Groot replication has to do with repeating an earlier study in exactly the same way, but with the introduction of fresh samples and new experimenters. Replication means repeated tests of the same hypothesis on fresh samples. The guideline given is that "the reader must be put in possession of all the relevant facts about the experimental design, the data processing, and the outcomes so that, should he be so inclined, he can replicate the investigation (De Groot, pp. 167, 168). Because of "practical reasons (space in journals, readability, and the like)" (p. 168), this rule will sometimes have to be relaxed, but the more detailed data and the original materials must, according to De Groot, remain available for some time for study by others who wish to replicate the study.

In Popper (1972a) we find the view that in order to generate objective, valid, justifiable, and scientific knowledge, such knowledge in principle must be testable by anyone. Testability is an aspect of the more general idea of intersubjective criticism: Rational control by means of critical discussion. Popper is only prepared to talk about testing our observations where it is possible that "certain events recur in accordance with rules or regularities, as is the case with repeatable experiments" (Popper, 1972, p. 45). If an empirical scientific statement is presented in such a way that the experimental arrangements are described and can be tested by any competent researcher (anyone who has learned the relevant technique), then scientific observations are testable and repeatable. And a theory is only falsified if a reproducible effect is discovered that can be seen as a refutation of the theory.

On the question as to how often an effect in reality must be reproduced before it can be called a reproducible effect or discovery, Poppers answer is:

> In some cases *not even once*. . . . If I assert that there is a family of white ravens in the New York zoo, then I assert something which can be tested *in principle*. If somebody wishes to test it and is informed, upon arrival, that the family has died, or that it has never been heard of, it is left to him to accept or reject my falsifying basic statement. As a rule, he will have means for forming an opinion by examining witness, documents, etc.; that is to say, by appealing to other intersubjectively testable and reproducible facts. (Popper, 1972, p. 87)

Thus, for Popper the idea of an arbitrary reproducible physical effect (Popper, 1972, p. 203) is closely connected with the idea of intersubjectivity.

In the above sketched algorithmic view on replication Popper is pointing to a complicating factor:

> All repetitions which we experience are approximate repetitions; . . . for any given finite group or set of things, however variously they may be chosen, we can, with a little ingenuity, find always points of view such that all the things belonging to that set are similar (or partially equal) if considered from one of these points of view; which means that anything can be said to be a repetition of anything, if only we adopt the appropriate point of view. This shows how naive it is to look upon repetition as something ultimate, or given. (Popper, 1972, pp. 420, 422)[1]

Replication in Practice

When we read the latest literature on replication in anthropology, sociology, and the sociology of science, it becomes clear that the normative and ideal view of replication is not actually found in the practice of scientific research. Scientists do not carefully criticize each others' research outcomes by replicating the original experiment.

Mulkay and Gilbert (1986), philosophers of science, have shown on the basis of interviews with biochemists that in the business of science the "official view of replication," that is, "exact replication," "close copying," or "*routine copy*" *is not everyday practice. A thorough analysis of the complex*, diverse, and flexible interpretative work usually done by scientific researchers yields a different picture from the "official" view on replication. Usually replication is seen in practice as "doing something *different* which (leads) to the same scientific conclusion" (Mulkay & Gilbert, 1986, p. 26). *Experiments are seen as either supporting or weakening a general problem* solution or proposition under discussion in the scientific community.[2] Such a proposition is "taken to be a more general statement than that inherent in the observation furnished by any particular experimental set-up. . . . (I)t can be verified, validated or replicated in various different ways" (Mulkay & Gilbert, 1986, p. 26).

When replication is used in this sense, the most adequate metaphor is Peirce's cable metaphor (1955, p. 229). Along the lines of several experimental paths outcomes are generated which can throw light on a problematic proposition. If in the scientific discourse these several outcomes can be as

[1]Collins rightly calls this "the Popper problem" (Collins, 1985, p. 30).

[2]In the words of Hunt: The hypothesis must be tested, not the experiments (Hunt, 1975, pp. 588, 590). "In fact, the best replication may be one which only replicates the hypothesis, since it is the strongest test of the hypothesis' generalizability" (Hunt, 1975, p. 589).

intimately connected as the fibers of a cable, then the argumentative force of a proposition becomes stronger. Mulkay and Gilbert characterize this way of acting in research as "replication through experimental variation" (1986, p. 27). A number of research techniques and research designs are deployed to test a *conclusion*, or to confirm it. The conclusions of earlier research function as starting points for new original experiments and analysis. Only in those cases where there is doubt about the tenability of conclusions from earlier research is there a likelihood that the research on which those conclusions are based will be replicated as exactly as possible. But as long as this is not the case, it is assumed that the results of the original research can, in principle, be replicated. The possibility of exact replication is a presupposition which is seldom effected. "Even in accounts of cases where there is marked scientific disagreement, speakers do not usually question the reproducibility of each other's observations. Rather, they query the scientific meaning of the other's experimental regularities; and questions of scientific *meaning*, they maintain, are best settled, not by means of simple repetition, but through the strategy of experimental variation" (p. 30). Only in those situations in which the results of original research "come to be defined as increasingly problematic, (then) exact replication comes to seem more appropriate" (p. 34).

From Reliability to Validity

When the need to replicate conclusions is so strongly emphasized, it is not surprising that one starts using the term *validation* instead of *reliability*. In the practice of replication it is not the repeatability of the results which is at stake but the validity of the conclusions.[3] Experiments take place along an argumentative continuum in which primarily the validity of the conclusions is the subject under discussion. The most important as well as the most prevalent contribution to such a discussion is the strategy of the varied replication. In such a discussion original research outcomes have the status of "facts": It is assumed that they are repeatable and that there is no need to put this to the test. Only when strong doubts arise does "exact replication" appear on the stage of the validity discussion. "Close experimental copying, then, is treated (. . .) as a strategy to be used very occasionally in order to try to find out how doubtful findings are being produced" (p. 34).

EPISTEMOLOGICAL PRESUPPOSITIONS

One conclusion then is very clear: Reflection and prescription in the philosophy of science do not correspond to actual research practices. In research

[3]See also Chapter 6, in which Van IJzendoorn discusses the validation of knowledge claims. Popper, as we have seen above, also talks about the validity of generated scientific knowledge (Cf. Hunt, 1975, p. 587).

practice it is not the results of research that have primacy, but the *conclusions* that can be drawn from these "research outcomes." The pivotal question now is: What is the value of the repeatability of the results of research in relation to the conclusions that are built on these research outcomes? To put it differently: Does the repetition (or repeatability in principle) function as an argument for the "reality value" (the truth) of the conclusions in the (theoretical) discourse between scientists?

To be able to answer this question we must examine what the epistemological presuppositions are that form the basis of how philosophers of science perceive the relation between research outcomes, conclusions, and the reality content of these conclusions. First we analyze the standard view with regard to the status of research results, including Popper's version, and then we take a look at the ideas that have been developed more recently.

The Standard View and Popper's Variant

The clearest and most outspoken view on the relation between epistemology and methodology is the orthodox or standard view. Epistemology and methodology are seen as being seamlessly connected. In this view epistemology is fundamental in the sense that, on the basis of a separation between subject (the knower) and object (the existing reality independent of the knowing subject), and taking the fallible knowledge apparatus of the knowing subject into account, the purpose is to capture reality as accurately as possible. If the research outcomes are reproducible then the claim to truth of the conclusions is stronger, that is, the objectivity of the conclusions is better substantiated. Methodological repeatability of the same process—and on the basis of that, also the same effects—renders increased certainty, which strengthens the claim to truth. As certainty increases, the implication is that independent reality has actually been captured as a result of the research process. Replication in this view "automatically" constitutes the argument for the validity of knowledge.

Popper makes a sophisticated variant out of this orthodox view. He still accepts the classical connection of methodology and epistemology but at the same time adds a communicative component to it and stresses the importance of the critical discussion between scientists. He points to the decisive role of common background knowledge in empirical testing procedures. In contrast with monological verificationism, Popper draws attention to intersubjective testability and intersubjective criticism. The process of testing empirical statements has no "natural" final end: One can always derive new consequences (new conjectural and testable statements) from the statements to be tested and so ad infinitum (see Kunneman, 1986, p. 21). In Popper's own words: "(I)ntersubjective testability always implies that, from the statements which are to be tested, other testable statements can be deduced. Thus, if the

basic statements in their turn are to be intersubjectively testable, *there can be no ultimate statements in science*" (Popper, 1972a, p. 47).

Specifically one of the characteristic features of empirical testing procedures is their finiteness. A stop is put to a procedure such as Popper describes when the scientific community decides to provisionally accept certain, easily testable basic statements as truth. Because of the introduction of this idea of "unproblematic background knowledge" (Popper, 1978, p. 390), the theories that are tentatively accepted as truth and that form the empirical basis of scientific action and decisions about the empirical falsification of theories and hypotheses, have become communicative (see Kunneman, 1986).

Replication in the sense of "repeatable effect," however, remains for Popper the ultimate methodological criterion for judging the value of research. He never reversed his view. This is not surprising when we take a look at his later work and especially at *Objective Knowledge*, in which Popper (1972b) develops an evolutionary epistemology. Scientific knowledge has to do with *a reality that is autonomous to the knowing subject*. When we look at his ideas about theories of truth, it turns out that he is an adherent of Tarski's correspondence theory of truth. In trying to give a further underpinning to that theory of truth he introduces his concept of the three worlds in *Objective Knowledge*. World-1 is the reality of material processes and things, the physical world; world-2 is the domain of the psychic reality of experience and thought, of language and understanding; and world-3 finally is knowledge without a knowing subject, the domain of objective knowledge, objective thought. In his theory of the three worlds Popper separates the *communicative processes* in the scientific community and the existing beliefs (world-2), from the problems, problem solutions, conjectures and refutations, and from the growth of objective knowledge, in short, from the *cognitive validity* of the beliefs (world-3).

> The purpose of this separation is that it can prevent the *having* of ideas and views by a specific group of scientists from being identified with the *validity* of these ideas and views. This immunizes the truth and the falseness of theories against a relativistic equalization of the valued and the valid. An ontological and epistemological reserve is created, named world-3, in which objectivity and validity can continue to exist unthreatened and in which the growth of knowledge is not contaminated by social influences and communicative processes. (Kunneman, 1986, p. 23)

Methodological decisions, so it turns out, are located in world-2 which is why the gap between world-2 and world-3 cannot be bridged. From the perspective of the validity of scientific knowledge, the methodological recommendations are kept up in the air. With regard to the criterion of falsifiability, in world-2 we are no longer able, from a methodological perspective, to say

anything about the *correspondences* between statements or between knowledge and reality, in terms of the truth content of these statements (see also the conceptual problems with Popper's concept "verisimilitude," Popper, 1978). This means that Popper, in spite of having introduced the communicative component, nevertheless takes as his point of departure a rigorous separation between subject and object. In the end he ties his ideas on replication to the object-world outside of us. This being the case, we can say that his view is indeed a variant of the standard view. For him too, replication guarantees the truth of our knowledge.

In summary both positions start with an epistemology that presumes a rigorous separation between subject and object. In the process of acquiring knowledge this gap has to be bridged. Given this epistemology, replication (and actual replicability) is seen as a methodological criterion that ensures the gap is bridged as adequately as possible.

Kuhn

If we consider Popper's methodological remarks—given the separation of world-2 and world-3—outside of his epistemological view, what we are left with is a concept that stresses both *intersubjective criticism* of the results of research and the demand for optimal *communicability* (cf. Glastra van Loon, 1980).

It is precisely this direction that the post-Popperians, led by Kuhn, have taken. The evaluation and the validity of a statement is, according to them, one process. Now the object of research, as well as the subject, has been made "communicative fluid" as Kunneman so beautifully puts it (Kunneman, 1986). There is no independent criterion by which we can judge whether theories adequately mirror reality. Stress is laid either on the social and communicative processes (Popper's world-2), on the paradigmatical (Kuhn, 1970), or on disciplinary (Toulmin, 1972) empirical testing procedures. The implication is that epistemology is no longer a directive for methodology. Instead of a methodology in which one starts with a certain, realistic, and dual epistemology, the sociology of knowledge functions as the "source of information" for methodological discussions. Constructivism (see hereafter) is the most radical elaboration of this development to date.

A central idea in Kuhn's theory of paradigms is that theories are not immediately abandoned as soon as outcomes of empirical experiments are at odds with the effects one would have expected on the basis of the theory. For Kuhn, empirical tests are of minor importance when success or failure of theories is at stake. Rather, one tries to reinterpret the outcomes of earlier research. According to Popper, replication should get a more or less algo-

rithmic connotation (replication as exact replication)[4] in the (hypothetical but never actual) case of a successful linking between world-2 and world-3. Kuhn's view on the other hand, is more in line with the above-mentioned experimental variation which makes possible a communicative process of interpretation and reinterpretation. In the scientific discussion, according to Kuhn, one should focus on the conclusions, that is, the *effects* of research. Such discussions can never be settled by pointing to an objectively given reality.

The important shift that takes place with Kuhn is that epistemology is no longer fundamental for methodology. We might even say that this is a case of "methodology without epistemology." This means that the value of replication is no longer epistemologically warranted; it has to be determined in the scientific discussion. As we shall see, this becomes even more preeminent in constructivism.

Constructivism

Completely in line with Kuhn's theory of paradigms (in which methodological criteria are not absolutely fixed but are the result of the discourse of paradigmatically bound communities), discursive and argumentative processes are also central in the work of constructivists like Collins, Knorr, Latour, & Woolgar. According to the constructivists, replications of original research must always be seen as the *outcome* of a process of *negotiation*. The negotiations entered into by scientists deal with the conclusions, and the results, as claims not as evidence. That is why science as a process of negotiation is in principle a never-ending event, with no fixed Archimedean point.

The main characteristic of constructivism is its complete reversal of the usual perspective. This is made clear for example by Tibbets (1988) when he discusses the status of ontology in constructivism.

> (F)or realists ontological assumptions constitute a *sine qua non* for understanding scientists' empirical and theoretical activities and apart from such a realist ontology such activities would make no sense at all. For constructivists, on the other hand, *the* issue is not the respective merits of one ontological

[4]We pass over Lakatos's methodology of scientific research programs (cf. Lakatos, 1978), because the problem of the relationship between ontology and methodology that is found in Popper, comes up even stronger with Lakatos. His evaluation of the history of the natural sciences and his reference to the basic value judgments of scientists form the basis for his arguments for the increasing verisimilitude of theories and increasing "grip" on the structure of reality. Lakatos only gives us the philosophical postulate of an independently given reality, but he cannot articulate a substantial relationship between methodology and ontology (cf. Kunneman, 1986).

framework over another but inquirers' discourse, accounts, cognitive constructions and *praxis*. (p. 128)

The constructivist "reversal" is also nicely illustrated in Latour and Woolgar's (1979) well-known book *Laboratory Life*. In this book, which is partly based on Latour's two years of empirical, cultural-anthropological research in the Salk Institute, the authors state that: "The difference between object and subject or the difference between facts and artefacts should not be the starting point of the study of scientific activity; rather, it is through practical operations that a statement can be transformed into an object or a fact into an artefact" (Latour & Woolgar, 1979, p. 236). In science there is no passive description and representation of an autonomous, already existent reality (Facts, Nature), but "reality (is) the *consequence* of the settlement of a dispute rather than its *cause*" (p. 236). Facts are not found, but scientists strive to produce them. So the "reality effect" of facts is constructed. "Reality cannot be used to explain why a statement becomes a fact, since it is only after it has become a fact that the effect of reality is obtained" (p. 180). This is the reason then why Latour and Woolgar do not talk about reality, but prefer to speak of stabilizing a scientific fact as a statement, a proposition. When in the research process a fact stabilizes, a bifurcation takes place: "On the one hand, it is a set of words which represents a statement about an object. On the other hand, it corresponds to an object in itself which takes on a life of its own" (p. 176). Latour and Woolgar do not deny that in this process researchers are manipulating matter, but the acceptance of a statement about this matter as a scientific fact is based on the constructed consensus.

Latour and Woolgar also point to the "agonistic field" (p. 237) in which these constructions are located. As much as possible scientists try to get rid of "modalities which qualify a given statement" (. . .). "An advantage of the notion of agonistic is that it both incorporates many characteristics of social conflict (such as disputes, forces, and alliance) and explains phenomena hitherto described in epistemological terms (such as proof, fact, and validity)" (p. 237). Negotiations about what constitutes a demonstration or a replication in a certain setting, or what makes a good "biological assay," are not in any way different from the discussions between politicians, judges, and lawyers. The psychological or personal evaluations of the various parties are *never the sole issue* at stake. The central focus is always on the force of the argument ("the solidity of the argument").

According to Latour and Woolgar construction and the notion of agonistic are not undermining the power of scientific facts; one should avoid a relativistic interpretation and use of these concepts. For this reason Latour and Woolgar point to the process of *materialization* or *reification*. As soon as a statement stabilizes in the agonistic field, it becomes part of the implicit, the "tacit" skills or the material equipment of other laboratories. It is reintro-

duced in the form of a machine, inscription, plan, skill, routine, presupposition, deduction, program, and so on. In sum: "The result of the construction of a fact is that it appears unconstructed by anyone" (p. 240).[5]

Collins's work concurs nicely with the views held by Latour and Woolgar. However, Collins goes much further in elaborating the concept of replication (Collins 1985; see also Collins, 1975, 1981). He tries to reverse the discussion about replication in showing that replicability is not an axiomatical term with which to start a discussion, but that replication too must be seen as the *outcome* of process of negotiation.

> The actual replicability of a phenomenon is only a cause of its being seen as replicable in the same way as the colour of emeralds is the cause of their greenness. Rather the belief in the replicability of a new concept or discovery comes hand in hand with the entrenching of the corresponding new elements in the conceptual/institutional network. This network is the fabric of scientific life. Replicability . . . turns out to be as much a philosophical and sociological puzzle as the problem of induction rather than a simple and straightforward test of certain knowledge. It is crucial to separate the simple idea of replicability from the complexities of its practical accomplishment. (Collins, 1985, p. 19)[6]

According to Collins, the notion "isomorphic copy," that is, repeating an experiment "similar in every knowable, and readily achievable, significant respect" (Collins, 1985, p. 170), is problematic because of the conceptual problems that are connected with the term "the same" (see Collins, 1985, pp. 38, 170; see also Popper's remarks on this topic earlier). For this reason Collins argues for what he refers to as the enculturation model of replication. What is considered as a "working" (and therefore in the discussion a decisive) experiment is, according to Collins, the outcome of a process of negotiation between researchers who form a "community of skills." Implicit statements about the nature of the discovered effect in the experiment can also play a role in such a process of negotiation.

A fine example of Collins' enculturation thesis of replication is given by Draaisma in his reconstruction of the so-called "Benveniste-affair" (Draaisma,

---

[5]Latour and Woolgar add: "The result of rhetorical *persuasion* in the agonistic field is that participants are convinced that they have not been convinced; the result of *materialization* is that people can swear that material considerations are only minor components of the thought process; the result of the investment of credibility, is that participants can claim that economics and beliefs are in no way related to the solidity of science; as to *circumstances*, they simply vanish from accounts, being left to political analysis than to an appreciation of the hard and solid world of facts!" (p. 230).

[6]"(M)ost published results cannot be easily re-generated, or validated—a fact which may be surprising to the outsider. Hence what is selected will not depend on the evaluation of a product in itself; instead, it depends on whether a result fits into the framework of a current undertaking and works out successfully in instrumental manipulation" (Knorr, 1977, p. 689).

1989). In 1988, according to Draaisma, Benveniste et al. published an article in *Nature* which contained the empirically corroborated statement (so he said) that even in an extreme dilution an allergen keeps its biological efficacy despite the fact that, theoretically speaking, the dilution no longer contains any molecule of the original substance. Although not explicitly stated, one of the main premises of homeopathy was here at stake. In describing the way in which a research group established by *Nature* carried out a "replication on the spot," Draaisma shows that although such a type of replication succeeds in circumventing some of the difficulties of replicating research, the research process is best typified as a negotiation process, even with respect to the research outcomes (as distinguished from the more general research conclusions).

Having the same people who carried out the "first" study carry out the replication in Benveniste's own laboratory did away with the problem of giving an exhaustive enumeration of instructions. Because of the fact that the same laboratory was used, and the same materials, techniques, and instruments as in the "first" study, differences between the "first" and the "second" study could not be ascribed to these factors. Before carrying out the research, consensus was reached on the precise way of carrying out the experiments—and this was of special importance—on the meaning of possible outcomes, so that the experiments themselves, when carried out, would be decisive (cf. Draaisma, 1989, p, 102).

Because the classical ontology is not sound (white or black crows do not exist, only crows that are *taken* to be black or white, which means that the central question is what is *taken* as black and what is *taken* as white) this variety of replication (on the same spot and with the same experimenters, etc.) shows that in the end, even in the case of an exact replication, decisions have to made by negotiation. The process of negotiation has to establish what counts and what does not count as a satisfactory—and in the end—a decisive experiment.

Even a case of "replication on the spot," carried out by the original experimenters and with a preestablished consensus on procedures and the meaning of the conceivable results, does not necessarily guarantee agreement, is shown by Draaisma in a reconstruction of an "on the spot replication" instigated by *Scientific American* in 1923, of research on the power of psychic mediums. Although everything was done to rule out possible disagreement in advance, the editor of *Scientific American*, Bird, and the famous magician Houdini could not come to a unanimous conclusion. From this we may conclude that disciplined "on the spot replication" increases the possibility of reaching consensus but does not guarantee it (cf. Draaisma, 1989, p. 104).

A "Realistic" Problem

With constructivism we have again a case of "methodology without epistemology." We even might call the constructivist position a radicalization

of the Kuhnian ideas, because arguments against a traditional epistemological interpretation of the process of gathering scientific knowledge are given much more explicitly. One of the main arguments is that we cannot use "reality" as an object that can be relied on and referred to by the knowing subject because "reality" is considered to be the *outcome* of the "settlement of a dispute" and cannot, therefore, be used to settle the dispute while it lasts. A similar view is held on replication: Belief in the replicability of research outcomes must be established step by step, it cannot be used as an irrefutable axiom.

Seen from a methodological perspective, constructivism fits actual scientific research practice very well in which—as we have seen—disputes about research outcomes and their significance for research conclusions take a central place. We have also seen that in scientific practice experiments never speak for themselves, nor do they automatically lead to specific conclusions. This "fit" is no wonder, seeing as the origin of constructivism lies in empirical research of scientific practice. However, the main characteristic of constructivism is—in our opinion—also its main problem. Because constructivism is "methodology without epistemology," it leaves the scientist empty-handed with respect to the epistemological significance of his research outcomes and the conclusions reached on the basis of them. It is no wonder, then, that constructivism is often criticized for its idealistic tendencies and implications.

Not only critics are aware of this danger; constructivists themselves do their utmost to avoid such an interpretation of their position. For example, Latour and Woolgar are careful to stress that they do not deny that experimenters manipulate matter. And although they state on the one hand that reality is the outcome of a process of negotiation, they add at once: "We do not wish to say that facts do not exist nor that there is no such thing as reality. In this simple sense our position is not relativist. Our point is that out-there-ness is the *consequence* of scientific work rather than its *cause* (Latour & Woolgar, 1979, pp. 180–182). According to Latour and Woolgar science does not aim at constructing "reality" but at stabilizing statements (cf. Latour & Woolgar, 1979, p. 180). This means that the set of stabilized statements "fills up," as it were, the constructivist concept of "Reality."

It must be noted that in their defense of their position Latour and Woolgar use the concept "reality" in two different ways: a constructivist one ("reality" as the outcome of a process of negotiation) and a "realistic" one ("reality as such," about which we are informed that its existence is not denied). When they speak about "reality" or "matter" in a constructivist sense, they do not refer to some existence outside of us, but to the set of statements, the collection of facts about which, at this moment, consensus has been reached. Of course this set of statements can be used to test other statements. But it must not be forgotten that this is not a confrontation of a statement with "reality-outside-of-us," but with "reality-as-viewed-at-this-moment-with-respect-to-the-present-state-of-our-knowledge."

In fact, the defense strategy that Latour and Woolgar use is a disguised form of subject–object dualism; Latour and Woolgar ward off their critics by referring to the existence of "facts" and the existence of "such a thing as reality." In spite of the constructivist declaration of intent to provide an adequate way of talking about "materiality," the remark cited above shows that, when the problem is discussed, Latour and Woolgar fall back on a rather unreflective use of dualistic epistemology in which the relationship between "facts" and "such a thing as reality" on the one side and the constructivist concept of "reality" on the other remains unclear.

This means that for the moment constructivism brings the discussion only one step further. Although it gives an account of replication that is more in agreement with scientific practice, it does not succeed in doing away totally and radically with the epistemological presuppositions of either the received view or Popper. To be able to take this second step we are in need of a more adequate view of the concept of "reality"; that is, a view without dualistic epistemological presuppositions. Only then might we be able to give an answer along constructivist lines on our "pivotal question" about the epistemic status of replication.

THREE FORMS OF REALISM

In recent times the philosophy of science has seen renewed attention paid to questions about the "reality" of (scientific) knowledge. Such different philosophers as Van Fraassen, Salmon, and Harré have all tried to develop a more adequate realistic position than "standard" realism (cf. Radder, 1988). In this section we briefly discuss two forms of this "new" realism—Van Fraassen's *constructive empiricism* and Radder's *referential realism*—in order to see whether their ideas can further our discussion. Then we introduce John Dewey's *transactional realism*.

Constructive Empiricism

Van Fraassen tries to solve the problem of realism by making a distinction between knowledge claims that refer to observables and knowledge claims that refer to nonobservables. His suggestion is that only the first are relevant for the relation between knowledge (science) and reality. Apart from the fact that Van Fraassen tries to found the validity of knowledge on something outside knowledge (the epistemic status of which remains unclear), the main problem is why we should restrict the epistemic (and the realistic) significance of scientific knowledge only to what is observable. Why, we can ask, is observation with the naked eye better and more real than observation with, for example, a contact lens? (cf. Radder, 1989). Furthermore, we have seen

that even what is observable does not always (or maybe: never) speak for itself. Even on the elementary level of research results the significance of the results has to be established by negotiation.

Referential Realism

With Radder we find a more sophisticated position, called *referential realism*. Contrary to the usual practice, Radder does not lay stress on theory but on experiment. Based on the distinction made by Habermas between *objectivity* (the objectivity of experiences only guarantees the identity of the objects of experience in the manifold of possible interpretations) and *truth* (the truth or falsity of statements is dependent on a consensus on the validity of interpretations and theories; a consensus that has to be reached anew in every case; (cf. Habermas, 1973, 1982), Radder wants to maintain the idea of knowledge as referring to a human-independent reality, but at the same time he rejects both the idea that theories depict such a reality and the idea that subsequent theories do this job either more accurately or adequately. For this Radder proposes the following criterion of reference: "The descriptive terms occurring in the theoretical description of an experimental situation refer to elements of a domain in human-independent reality, if the material realization of this experimental situation is possible in a reproducible manner" (Radder, 1988, p. 102).

Radder's main stress is on the "material realization" of experiments. In order to guarantee that a material realization can be effected independently of specific theories (and solve the problem of incommensurability), Radder introduces the layperson. If two competing theoreticians succeed in instructing laypersons, in everyday terms, to carry out an experiment (i.e., can accomplish a material realization of the experiment), both theoreticians have a commensurable starting point for their theoretical descriptions and explanations. When it has been shown that such experiments are repeatable, we can be certain, according to Radder, that the terms in both theoretical descriptions refer to independent existing entities in a certain domain. The material realization, however, does not say anything about the adequacy of the description of those entities. This means, according to Radder, that "we can say of terms in scientific propositions *that* they refer to elements in a human-independent reality, but not that these propositions describe reality as it is 'in itself' (approximately or not). Thus this kind of realism implies reference, but not correspondence, picturing, reflection, representation or something like that" (Radder, 1988, p. 102). If, from the perspective of two competing theories, reference has been established, we can also conclude that co-reference has been established. The theoretical terms in both theories refer to a reality which is independent both of theories and of men.

The advantages of Radder's *referential realism* are found in his suggestions

for solving both the problems of incommensurability and co-reference. But because of the fact that Radder bases his criterion of reference on replicability and actual replication, his ideas cannot make a contribution to our present discussion. In Radder's *referential realism* replication is used as a primitive term, which serves as an argument for reference, while earlier we stressed the fact that a replication is not the beginning of a scientific discussion but the outcome of it. This means that questions about the status of replication and replicability cannot be answered by Radder's realism.

Apart from the formal point that Radder's realism does not forward our present discussion, his rather uncritical use of the concepts of "replication" and "replicability" also indicates that other more substantial problems might be encountered with his referential realism.

Transactional Realism

The same cannot be said of John Dewey's *transactional realism* (Sleeper, 1986), because he does not make use of a dualistic epistemology. It was precisely this—as we have seen—that was the main problem for constructivism, as a consequence of which the reproach of idealism could only be defended by a form of "standard realism." In our opinion Dewey's *transactional realism* points a way out of this epistemological dilemma. It provides us with a different line of approach, one in which a more adequate interpretation of the concept "materiality," and the concept "reality," comes within reach. We will not present Dewey as having all the answers to the methodological and epistemological problems of the replication debate. But we will argue that Dewey's ideas can help supplement the methodological ideas of constructivism with a consistent epistemological position so that a better relation between methodology and epistemology can be established.

The most important characteristic of Dewey's epistemology is that from the outset he does not postulate any dualisms, and especially not the dualism of a knowing mind over and against "the world." Instead, Dewey starts by stressing the continuity of man and nature and expresses the view that knowing is a process that takes place *inside* nature. He characterizes his monistic, "transactional" *approach* as a procedure "in which is asserted the right to see together, extensionally and durationally, much that is talked about conventionally as if it were composed of irreconcilable separates" (Dewey & Bentley, 1949, p. 69). In separating subject and object, mind and matter, and by viewing knowledge as a process which takes place outside of nature, the dualist arrives inevitably at such problems as "how it is possible to know at all; how an outer world can effect an inner mind" and "how the acts of mind can reach out and lay hold of objects defined in antithesis to them" (Dewey, 1958, p. 10). From a transactional perspective the question is not how subject and object can reach each other in the process of knowing, but

"how and why the whole is distinguished into subject and object, nature and mental operations" (Dewey, 1958, p. 9).

Dewey elucidates this process by introducing a combination of three concepts: "transaction," "experience," and "knowledge." The concept of "experience" is used for those transactions, those transactional processes, in which there is a living organism which is bound to its environment in a double way (cf. Dewey, 1980a, p. 7). "The organism acts in accordance with its own structure, simple or complex, upon its surroundings. As a consequence the changes produced in the environment react upon the organism and its activities. The living creature undergoes, suffers, the consequences of its behavior. This close connection between doing and suffering or undergoing forms what we call experience" (Dewey, 1948, p. 86). According to Dewey we can speak of knowledge if the relationships between "doing and undergoing," between actions and the consequences of the actions, are formulated by means of symbols (cf. Dewey, 1939, p. 532). This means that "knowing is not the act of an outside spectator but of a participator inside the natural and social scene . . . the true object of knowledge resides in the consequences of directed action" (Dewey, 1980b, p. 196). What, we must now ask, is the exact status of this object?

According to Dewey, objects must *not* be seen as "being in a way marked out by nature always and for all beings" (Dewey, 1975a, p. 78). Also our perception of objects is not the beginning of our knowledge but the provisional outcome of it. Objects-as-known are not independent of the process of knowing. For Dewey this means that it is absurd to have the intention of saying something about "the nature of things" before, and apart from, "experimental operations," "inquiry," "directed action" (cf. Dewey, 1939, p. 546). This also means that the "objects as they are for us" change with the development of our judgments and theories. Contrary to what is said in some realistic positions, progress, change, and development of our knowledge are not processes in which new judgments, new predicates are affixed to an already existing object. The object of knowledge is no "antecedent being," "antecedent reality," or "antecedent existence" with "antecedent properties." It is precisely at this point, Dewey holds, that in spite of their opposite points of view "objective idealism and sensational empiricism" meet each other by presuming that

> *reflective* thought . . . has its test in antecedent reality as that is disclosed in some nonreflective immediate knowledge. Its validity depends upon the possibility of checking its conclusions with the terms of such prior immediate knowledge . . .. The basic premiss is also shared by realists. The essence of their position is that reflective inquiry is valid as it terminates in apprehension of that which already exists. (Dewey, 1980b, pp. 109, 110)

According to Dewey it is also not adequate to presume that with every change in our knowledge a new object, totally independent of the old, comes into

being (cf. Dewey, 1975b). In contrast to these ideas, Dewey holds that "the object of knowledge is eventual, that is, it is an outcome of directed experimental operations, instead of something in sufficient existence before the act of knowing" (Dewey, 1980b, 171). The object of knowledge is the consequence of acting experimentally. That this does not imply an idealistic position is made clear in the following remark: "When we are trying to make out the nature of a confused and unfamiliar object, we perform various acts with a view to establishing a new relationship to it, such as will bring to light new qualities that will aid in understanding it" (Dewey, 1980b, p. 87). The establishment of such a new relationship is, of course, not intended as a relationship between an independently existing object and an independently existing subject. The new relationship is the outcome of a transactional process of action. The object, or more precisely, the concept "object" that is the result of this process, is not viewed by Dewey in terms of institution (i.e., idealism) or predication (i.e., realism) but in terms of reconstruction.

The main advantage of Dewey's position in regard to constructivism is that it provides us with an *epistemological* account of the way in which we might see knowledge as the outcome of our "transaction," our "manipulations," an account which avoids the danger of falling back on idealism. By placing knowing "back" into nature, by viewing it as a process that takes place *within* nature and not as something outside of it, Dewey succeeds in maintaining (or from the perspective of the history of philosophy: reestablishing) the bond between knowledge and reality.

Although Dewey provides us with an epistemological account, this does not mean that a methodology can be based on this epistemology. An important consequence of Dewey's position is that there is no longer any difference between epistemology and methodology. From a transactional perspective we do not need a separate method to bridge the epistemological gap between subject and object. Methodology *is* knowing-in-action; the one *is* inseparable from the other, and neither has primacy. From a Deweyan perspective we must see method, "logical form," as something originating in processes of knowing, of inquiry (cf. Dewey, 1938). We can never start with an epistemology and build our methodology on it; we can only improve our methodology step by step in the very processes of knowing and of inquiry. This, as we have seen, is in perfect agreement with constructivism.

CONCLUSIONS

The question that has to be answered now is whether the concepts of "replication" and "replicability" can still be used in scientific and philosophical discourse in the light of the developments in the philosophy of science sketched earlier—and if so, how?

In the introduction we made a distinction between a "methodological" and an "epistemological" interpretation of the demand for replicability. Our first conclusion is that, in the light of the foregoing discussion, the methodological interpretation of the demand for replicability is still worthwhile. Both the interviews of Mulkay and Gilbert and the research carried out by Latour and Woolgar show that in scientific practice *discussion* about the research results is the most important and most fundamental activity. However, in order to conduct such a discussion meaningfully, it is of utmost importance that the genesis of both the research results and the research conclusions is—in principle—retraceable for fellow researchers. The reconstruction of the "Benveniste-affair" makes clear that even when this demand has been fulfilled, nothing has been said about the argumentative force of the research outcome, which also has to be established argumentatively—a process in which varied replication can play a significant role.

As we have seen, replicability can also be taken as a sign that the research results tell us something about "Reality." Here it is often argued that if it is possible to repeat an experiment with the same outcome, truth—viewed as a correspondence with "Reality"—has been attained. This "ontological" interpretation of replicability can be found both in the "received view" and with Popper, albeit in the last instance in a methodological "disguise."

It is exactly at this point that constructivism reverses the case. Here "World" and "Reality" do not function as something given, but as the outcome of a process of negotiation. What counts as "real" or as "replication" or as "replicable" is not the starting point of the negotiation but the eventual outcome. "(R)eality (is) the consequence of the settlement of a dispute, rather than its cause" (Latour & Woolgar, 1979, p. 236). Following Collins we can make the same statement about replication. This constructivistic approach is an improvement on the traditional point of view. First it is in better agreement with scientific practice; secondly it seems to do away with a naive "mirroring" epistemology. The main problem, however, was that constructivism presented us with a "methodology without epistemology." By focusing exclusively on the methodological aspects of scientific practice, constructivists were no longer able to say anything about the epistemic status of the conclusions of science and the role of replication in the discussion. By means of Dewey's transactional realism it becomes possible to take out the epistemological "sting" still remaining in the methodological "flesh" of constructivism.

The main significance of Dewey's perspective for the discussion about replication is, in our opinion, what one might call his "situalization" of knowledge. From a Deweyan perspective we can speak of knowledge when the relations between actions and the consequences of these actions are formulated by means of symbols. By determining the consequences of specific actions, those actions acquire significance. This leads to a situation in

which specific actions start to refer to events in the future. The consequences of actions function "as necessary tests of the validity of propositions, *provided* these consequences are operationally instituted and are such as to resolve the specific problem evoking the operations" (Dewey, 1938, p. iv). However, knowledge of the relationship between actions and their consequences is not simply given, but must be realized anew in every case. We have a succession of "inquiries" with no beginning and no ending. The outcome of one process of inquiry, called "warranted assertibility" (Dewey, 1938), is introduced in the next "inquiry" and has to prove its relevance for the problem at hand in *that* process of inquiry. This means that knowledge, or warranted assertibility—viewed as the outcome of one process of inquiry—is strictly situational, strictly bound to that specific situation. Viewed against the background of Dewey's monistic transactional realism, this implies that although the "warranted assertibility" acquired in one situation is strictly bound to that situation, it does have an epistemic significance. It does tell us something about the relationship between our actions and the consequences of those actions in that situation. Whether the same action will have the same result in another situation cannot be said with total certainty; it has to be established anew each time.

From a methodological perspective, one of the main tasks of replication is to test the "robustness" of the original result for time-space and other variations (chapter 3). This can always be done with two different objectives: to establish a *certain* relationship with a menindependent reality, or to see if the results reached so far can also be used in other situations. In the last option we meet the aspect of generality of transactional knowledge. Here one goes further than the strictly individual, strictly situational, and strictly unique character of the research results reached; the question then posed is what is the meaning of this knowledge, this "warranted assertibility" for other situations? Is there still a relationship between specific actions and specific consequences in other situations? Methodologically seen, the strategy of "varied replication" can inform us about the usefulness of these actions in terms of antecedent and consequence. The strategy of "on the spot replication" seems to be the most adequate way of establishing the relationship between antecedent and consequence as strictly as possible.

This leads to our second conclusion. Besides the "methodological" interpretation of replication, we can, on the basis of a combined perspective of constructivism and transactional realism, uphold the view that the "epistemological" interpretation of replication also makes sense. This is not because actual replication gives us absolute certainty about "reality outside of us," but because the "warranted assertibility," the "knowledge" acquired in one specific situation might be found useful in another. This, however, is not an automatism, but has to be established anew for every case.

REFERENCES

Brownstein, L. (1987). Relevance of the rationalist constructivist/relativist controversy for the "validation" of scientific knowledge-claims. *Knowledge: Creation, Diffusion, Utilization, 9,* 117–144.

Collins, H. M. (1975). The seven sexes: A study in the sociology of a phenomenon, or the replication of experiments in physics. *Sociology, 9,* 205–224.

Collins, H. M. (1981). Son of seven sexes: The social destruction of a physical phenomenon. *Social Studies of Science, 11,* 33–62.

Collins, H. M. (1985). *Changing order. Replication and induction in scientific practice.* London: Sage.

De Groot, A. D. (1961). *Methodology. Foundations of Inference and Research in the Behavioral Sciences.* The Hague and Paris: Mouton & Co.

Dewey, J. (1938). *Logic: The theory of inquiry.* New York: Holt.

Dewey, J. (1939). Experience, knowledge and value: A rejoinder. In P. A. Schilpp (Ed.), *The philosophy of John Dewey* (pp. 517–608). Evanston/London: Northwestern University.

Dewey, J. (1948). *Reconstruction in philosophy* (enlarged ed.). Boston: Beacon Press.

Dewey, J. (1958). *Experience and nature.* New York: Dover.

Dewey, J. (1975a). *The early works, 1882–1898. Vol. 3, 1889–1892.* Carbondale and Edwardsville: Southern Illinois University Press.

Dewey, J. (1975b). *The early works, 1882–1898. Vol. 4, 1893–1894.* Carbondale and Edwardsville: Southern Illinois University Press.

Dewey, J. (1980a). *The middle works, 1899–1924. Vol. 10, 1916–1917.* Carbondale and Edwardsville: Southern Illinois University Press.

Dewey, J. (1980b). *The quest for certainty. A study of the relation of knowledge and action.* New York: Perigee.

Dewey, J., & Bentley, A. F. (1949). *Knowing and the known.* Boston: Beacon Press.

Dicker, G. (1972). John Dewey on the object of knowledge. *Transactions of the Charles S. Peirce Society, 8,* 152–166.

Draaisma, D. (1989). Voorbij het getal van Avogadro: de Benveniste-affaire. *Kennis en Methode, 13,* 84–107.

Glastra van Loon, J. F. (1980). *De eenheid van het handelen. Opstellen over recht en filosofie.* Meppel/Amsterdam: Boom.

Habermas, J. (1971). *Knowledge and human interests.* Boston: Beacon Press.

Habermas, J. (1973). A postscript to knowledge and human interests. *Philosophy of the Social Sciences, 3,* 157–189.

Habermas, J. (1982). A reply to my critics. In J. B. Thompson & D. Held (Eds.), *Habermas critical debates* (pp. 219–283). London: Macmillan.

Hunt, K. (1975). Do we really need more replications? *Psychological Reports, 36,* 587–593.

Knorr, K. D. (1977). Producing and reproducing knowledge: Descriptive or constructive? Toward a model of research production. *Social Science Information, 16,* 669–696.

Kuhn, Th.S. (1970). *The structure of scientific revolutions.* Chicago: The University of Chicago Press.

Kunneman, H. (1986). *De waarheidstrechter. Een communicatie-theoretisch perspectief op wetenschap en samenleving.* Meppel/Amsterdam: Boom. (German translation Frankfurt: Campus Verlag, 1990).

Lakatos, I. (1978). *The methodology of scientific research programmes. Philosophical papers* (Vol. 1). Cambridge, UK: Cambridge University Press.

Latour, B., & Woolgar, S. (1979). *Laboratory life. The social construction of scientific facts.* Beverly Hills: Sage.

Lykken, D. T. (1968). Statistical significance in psychological research. *Psychological Bulletin, 70,* 151–159.

Mulkay, M., & Gilbert, G. N. (1986). Replication and mere replication. *Philosophy of the Social Sciences, 16,* 21–37.

Nagel, E. (1961). *The Structure of Science. Problems in the logic of scientific explanation.* London: Routledge & Kegan Paul.

Peirce, Ch. S. (1955). Some consequences of four incapacities. In J. Buchler (Ed.), *Philosophical Writings of Peirce* (pp. 228–250). New York: Dover.

Popper, K. R. (1978). *Conjectures and refutations. The growth of scientific knowledge.* London: Routledge and Kegan Paul. (Original work published 1963)

Popper, K. R. (1972a). *The logic of scientific discovery* London: Hutchinson. (Original work published 1959)

Popper, K. R. (1972b). *Objective knowledge. An evolutionary approach.* Oxford: Oxford University Press.

Radder, H. (1988). *The material realization of science. A philosophical view on the experimental natural sciences. Developed in discussion with Habermas.* Assen/Maastricht: Van Gorcum.

Radder, H. (1989). Rondom realisme. *Kennis en Methode, 13,* 295–314.

Sleeper, R. W. (1986). *The necessity of pragmatism. John Dewey's conception of philosophy.* New Haven, CT/London: Yale University Press.

Tibbets, P. (1986). The sociology of scientific knowledge: The constructivist thesis and relativism. *Philosophy of the Social Sciences, 16,* 39–57.

Tibbets, P. (1988). Representation and the realist-constructivist controversy. *Human Studies, 11,* 117–132.

Toulmin, S. (1972). *Human understanding. The collective use and evolution of concepts.* Princeton, NJ: Princeton University Press.

PART II

The Methods Of Replication

Explicit replications indeed appear to be very rare in the social and behavioral sciences. Siegfried (Chapter 7) found 277 replication studies, published in the major psychological journals during 1983 to 1988. In experimental and clinical psychology, about 10 replications per year may be expected, whereas in child psychology this rate drops to about 5 per year (Siegfried, Chapter 7). In Part II of this volume, three chapters deal with the issue of replication in a different but interrelated way.

Siegfried stresses the epistemological differences between the natural and social/behavioral sciences. He doubts whether we should strive for more replications. In the natural sciences, experiments conducted 300 years ago still can be replicated easily. But social reality is constantly in flux, people change their behavior and attitudes because of historical events, or because of the influence of research results. The "same" experiment never can use exactly the same subjects and instruments, and a changing reality may be the reason many of the replication studies yield disappointing and diverging results. Siegfried emphasizes the complexity of the replication issue, and uncovers the ambivalent status of replication in the social and behavioral sciences.

Fiske (Chapter 5) also stresses the differences in methodological and epistemological status of the natural and behavioral sciences, but his approach aims at a concrete and feasible solution for the replication problem. Departing from the famous theory of the multitrait–multimethod matrix, Fiske shows us how several different kinds of method effects exist that preclude exact replication of research results. Method effects come into existence when (a) the conceptualization of the construct implies different observa-

tional categories compared to the measure used to operationalize the construct, (b) when subjects react to the intrusive aspects of measures which are irrelevant for the operationalization of the construct, and (c) when measuring operations have impact on the person generating the measurements. Fiske formulates several useful guidelines to enhance the probability of reproducible results, related to the influence of method effects. A clear-cut definition of basic concepts can prevent implicit shifts between different measures aiming at operationalizing the same concept. In accordance with the idea of a multitrait–multimethod matrix, Fiske also emphasizes the necessity of a built-in replication in every single study: Researchers should use at least two different measures for the same concept to check the stability of the outcome. And last but not least: Method effects can be diminished in using unobtrusive procedures which prevent subjects from reacting to being studied.

Fiske considers meta-analysis to be a major methodological innovation to trace problems in replication of research results. Because different characteristics, or methods, of replication studies can be considered as variables influencing their outcome, meta-analysis may uncover systematic variations in results depending on the way in which the original studies where conducted. Lytton (Chapter 6), however, shows that meta-analysis has to be considered as a study that is not by definition replicable itself. In describing his meta-analysis of sex effects in parents' socialization practices, Lytton shows how many decisions have to be taken in conducting a meta-analysis. For example, an important problem to be solved is how to decide what studies will be included in a meta-analysis: Some studies using the same concepts may operationalize those concepts extremely differently, and inclusion in the same meta-analysis may not seem feasible. A series of studies using different concepts, however, may in fact be focusing on the same phenomenon, and therefore be considered valuable replications. Lytton acts like a participant observer doing meta-analysis, and uncovers hidden assumptions and implicit decisions. His report of the pitfalls of meta-analysis makes clear that is indeed an important tool in discovering replicable results. But meta-analysis should not be an exception to the rule that all research must be replicated. The meta-analysis cannot act as the final judge in scientific discourse. Meta-analysis is just a starting point for another round of replications aiming at testing the robustness of the original meta-analytic results (see Chapter 3). Nevertheless, meta-analysis may represent the best evidence we have on a topic at a certain point in time, provided that it has been carefully done and allows—at least in principle—for replication of all steps conducted.

The three chapters show different solutions to the problem of replication: Fiske points out that the multitrait–multimethod matrix can be conceived as a replicability index; Lytton concentrates on quantitative meta-analysis as a tool to test the replicability of separate studies on the same topic; and Siegfried emphasizes the need for conceptual clarity and theoretical analysis

in paving the way for replication studies. The three authors, however, have the same stance toward their own approach: They do only suggest that their approach is one way of clarifying the issue of replicability, and they are all convinced that other perspectives are valuable as well. In this respect, the three chapters supplement each other in making clear that there is no single solution to the problem of replicability: Different approaches should be used, and if convergence is being reached, we may be convinced that this outcome reflects the most recent state of the art in a specific domain.

CHAPTER 5

Do Our Methods Work for or Against Us?*

Donald W. Fiske
Department of Psychology
University of Chicago

Forty to fifty years ago, psychological research in the United States was dominated by the study of rats learning mazes. The topic seemed very promising: Psychologists could work experimentally and objectively on the important problem of how responses are learned, how behavior is changed. But then something happened and the amount of research on rats decreased rapidly. One colleague thinks it was because the research failed to generate any new perspectives or stimulating questions. I believe that psychologists began to realize that findings holding for rats in mazes did not necessarily hold for other learning conditions and for other mammals, such as human beings. In particular, knowledge about rats learning in mazes did not help us to understand children learning in school or at home. As Beach (1950) pointed out, psychology in this country was moving toward a science of rat learning, not a general science of behavior.

One interpretation of this historical episode is that the wrong method had been chosen—using method in a broad sense. If psychologists wanted to understand how children learn, they should not have done their research on rats. Psychology, however, had not yet realized that psychological findings are usually specific to the particular method used to generate them (Fiske, 1986). Changes in the subjects (or at least in the subject population), in the context, in the stimuli, in the instructions, in the observers, or in the way in which the data are generated—almost any change—is likely to alter the

*I am indebted to Alan P. Fiske, Barbara P. Fiske, Susan T. Fiske, Louis Fogg, Kathryn F. Mason, and three anonymous reviewers for their incisive comments on an earlier draft of this chapter.

findings to some degree. As a consequence, most findings cannot be replicated exactly, and certainly not with the precision typical of research in the natural sciences. And if we cannot closely replicate our findings, we cannot generalize them with much confidence.

In a very wise article, Jerome Kagan (1988, p. 619) wrote: "The meaning of psychological terms, like the meaning of all scientific concepts, is affected by the source of the evidence. If the mechanisms mediating the observations generated by one procedure are different from the mechanisms operating in another procedure, one should not use the same conceptual term to summarize both corpora" or sets of evidence. If two procedures generate different data and different findings, it seems safe to conclude that the mechanisms are different in the two cases. But what are these mechanisms? We know very little about the processes going on when we measure behavior in psychological research. To illustrate his point, Kagan discusses procedures for assessing recognition memory and also measures for aggression: Is aggression as inferred from a child's striking a toy clown the same aggression as that inferred from peer nomination techniques? (See also Kagan, 1989, chap. 1.)

Providing another example, Gottman (1977) points out that two entirely different literatures on social isolation have been generated, one based on low frequencies of interaction with peers and the other using low level of peer acceptance, as derived from sociometric measures. He reports that there is no relationship between the interaction and the sociometric measures. Similarly, Parker and Asher (1987) note that, in studying children with difficulties in peer relationships, findings from studies using sociometric measures of peer acceptance may not generalize to children identified by assessments of social behavior (see also Hymel & Rubin, 1985).

Developmental psychologists recognize the significant effects associated with perspective: People in different roles see children differently. In their review of informant agreement on behavioral and emotional problems, Achenbach, McConaughy, and Howell (1987) report a mean correlation of .60 between pairs of informants with similar relationships to the child (e.g., two parents or two teachers), but a mean of only .28 between pairs with different relationships (e.g., a parent and a teacher). Pairs involving an informant and the self show slightly less agreement (.22). They note that, even within one perspective, different informants are likely to have different effects on a child, and hence more than one informant should, if possible, be used for each perspective. The need for avoiding just a single source and a single procedure in the assessment of childhood pyschopathology is recognized by West and Prinz (1987) who report in their review on parental alcoholism and child psychopathology that half the studies used only one source or perspective (e.g., teacher) and more than half did the assessment by only one procedure (e.g., an interview). The significance of differences in perspective is illustrated by the research of Cowan, Pederson, Babigian, Izzo, and Trost (1973), who

found that peer judgments predicted later mental health problems while an array of other measures had no statistically significant results.

The problem of method effects is pandemic, critical, and general in psychology. Psychologists are only beginning to recognize how large and widespread the problem is. Although such awareness constitutes a big step forward, it does not solve the problem. Method effects are certainly still with us. For example, the papers at a recent conference on decision making (Hogarth, 1990) are full of reported method effects, under such labels as question framing effects and response mode bias. Method effects are clearly present in research in developmental psychology (see Bank, Dishion, Skinner, & Patterson, 1990; Kagan, 1988). Have you seen any instances of multitrait–multimethod matrices that approximated the ideal, that is, showed evidence of only minimal effects from methods? Thirty years after the Campbell–Fiske paper (1959), the problem of method effects is still unresolved. Will soft psychology be held back indefinitely by this problem? (Cf. Fiske, 1986, pp. 72–77.)

But just what are method effects? The next major sections will discuss that term and the conceptual and empirical identification of such effects. The last part of this chapter examines what can be done by developmental and other researchers who take the problem of method effects seriously. The important role of judgment is indicated throughout the chapter. All researchers have to identify and deal with whatever method effects they judge to be present in the data from their own investigations.

But before looking at method effects more closely, we need to consider the implicit philosophy underlying their conceptualization and the kinds of phenomena studied by psychologists. In physical science, there are phenomena that can be measured by several discrete methods with essentially perfect agreement among the obtained measurements. Replication is exact. In contrast, many of the phenomena of interest to soft psychology (e.g., much of personality, social psychology, and developmental psychology) are largely the interpretations of physical events or the products of the cognitive processing of experiences. Such phenomena are judgments, judgments that depend on and vary with the judge, and also depend heavily on the particular circumstances in which each judgment is made.

So from a philosophical viewpoint, I believe that there is a physical reality out there, a reality for which we can construct conceptual pictures, bodies of theory that enable us to achieve an understanding of that reality. Insofar as you and I and most educated people can comprehend contemporary physical theory, we accept it, we agree on it.

In the domain of human behavior and experience, there are many phenomena with a physical aspect (including biochemical and physiological processes within the category of physical): for example, nonverbal behavior in face-to-face interactions, asymmetry in cerebral functioning, consistent re

sponses after variable-interval reinforcement. These areas seem fairly free of method effects; measurements here do not seem to be dependent on inference and judgment. I think our understanding of phenomena in these areas will continue to improve over time.

There are also phenomena that are personal, that occur within the person: private experiences and feelings, cognitive interpretations of these experiences, judgments about the behavior of others. Each of us has such experiences and cognitions, and we assume that others also have them. There have been a multitude of ways of construing them, ranging from variables derived from everyday language to the esoteric terminology of each psychoanalytic group. We have a plethora of constructs and will probably continue to have them. Rarely is a construct completely discarded; new constructs are proposed every day. These new terms—and the old ones—may enrich our experience, but do they further our scientific understanding of these interpretive phenomena? It is in this part of psychology, in trying to measure constructs pertaining to these phenomena, that method effects seem so strong and so crippling.

Cronbach (1975, p. 126) concluded that "though enduring systematic theories about man in society are not likely to be achieved, systematic inquiry can realistically hope to make two contributions. One reasonable aspiration is to assess local events accurately, to improve short-run control. . . . The other reasonable aspiration is to develop explanatory concepts, concepts that will help people use their heads" (cf. Cronbach, 1982, 1986). Although Cronbach may well be right about the prospects for understanding "man in society," I am still optimistic about other sets of psychological phenomena. We can refine some of our present approaches to studying processes within the human organism—not only neurological and physiological processes but also cognitive ones, and we can find new, even more fruitful approaches. Taken together, they should help us to achieve a better scientific understanding of human behavior.

THE DEFINITION OF A METHOD EFFECT

A method effect is a systematic contribution to a set of measurements that is confounding, intrusive, and not wanted. The source of the contribution is identified as the process of measuring or some part of it. Not included here are random error or error of measurement—that is, the residual variation after all possible sources of systematic variance have been specified. Thus, a method effect appears whenever we identify some systematic variance that is irrelevant to our purposes in making the measurements.

But what is relevant variance? That is up to the researcher to decide. For example, someone studying self-presentation in children might use children's

self-reports—perhaps their ratings of themselves. That investigator might well decide that such self-ratings are entirely appropriate to the purposes of the research. The fact that, in a study of self-presentation, children might exaggerate positive attributes could be construed as part of the phenomena being investigated, rather than as a potential defect in the technique. For that investigator, the use of peer ratings, and perhaps even the use of clinician assessments, would seem inappropriate because the peers' perceptions—and even the clinicians'—might be biased by their personal perceptions of the child based on their idiosyncratic interactions with that target person. That is, ratings by peers or clinicians might be seen as showing method effects in this case. In contrast, someone studying the social acceptance or rejection of children by their peers might well believe that peer evaluations were the method of choice (cf. Gottman, 1977) and that self-reports were not appropriate because they would introduce a method effect. Most of us would find acceptable the judgments of these investigators.

Note that these two investigators might well put the same labels on the variables they were measuring: perhaps "friendliness," "sensitivity to others," or "leadership." But their use of the same labels would not mean that they were studying the same constructs. "Friendliness as attributed to oneself" is not equivalent to "friendliness as perceived by one's peers."

It is up to each investigator to define each construct that he or she is trying to measure and to specify the conceptualization in adequate detail. Enough elaboration must be provided to permit others to judge how well the investigator's measuring procedures for that construct fit the conceptualization and to what extent these procedures generate method effects. For example, a single procedure may be condemned for getting at only part of the behavior specified as relevant to the construct. Or a procedure may be of a kind that is known to be susceptible to response sets. Although the conceptualization of a variable is up to the researcher, it is also the responsibility of that researcher to communicate that comprehensive definition to others. Similarly, although a researcher is free to choose among available methods, readers of that researcher's report are entitled to judge whether or not the chosen method is likely to generate method effects. The type of method effect produced by poor linkages between a construct and a measuring operation is widespread in psychological research. But even the careful and comprehensive specification of constructs, as in the definitions of needs by Murray (1938), may not prevent people from creating individualistic measures. As a result, even though derived from the same specifications and having the same labels for the constructs, several techniques developed to assess Murray's needs are clearly not measuring the same variables (Fiske, 1973).

Issues about specifying method effects thus raise questions about our concepts: Should they be abstract and general or concrete and constrained? If they are too open, they may be too loose, uncritically allowing a wide range

of procedures to be employed in investigations of them. If they are defined narrowly and if certain measuring operations are stated as required, then we are in danger of approaching the absurd extreme form of operational definition in which each particular operation must have its own concept. Note that there are problems of replication at both extremes. With broad concepts and diverse measuring operations, measurements will not be replicated and findings will often not be reproduced. At the other extreme, any replication of measurements and reproducibility of findings would be within such tight limits that we would not be able to reach the generalizations we seek.

Measurements for broad, loosely defined concepts are quite likely to be confounded with method effects, but it is often difficult to pin them down. At the other extreme, a concept that is defined precisely as that which the measuring operation measures can be assessed without any concern for method effects—there can be none, by definition. For concepts falling between these extremes, we regularly have problems with method effects.

In psychology, we obviously need concepts at several levels of abstraction, including a fairly concrete level, and we need careful specification of the relationships between the concepts. What we ultimately want is good reproducibility of conclusions. To attain that, we need close replication of measurements, something we can hope to achieve only if our conceptual framework includes some simple and explicit low-level concepts (cf. Fiske, 1986).

In principle, then, a researcher can assert that no method effects are present in his or her research since each concept is defined strictly in operational terms, as that which its measuring procedure measures. In practice, however, some degree of construct-method incongruence seems almost inescapable in psychological research. (Recall the earlier discussion of aggression and social isolation.) In addition to such incongruences, there are many other sources and kinds of method effects.

Kinds of Method Effects. Although method effects are widespread, they are often elusive because they appear in many forms, guises, or disguises. Some exist because people react not only to whatever stimuli they encounter in a given context but also to the fact that they are being measured. Most specifications of our concepts do not say whether the subjects should know or not know that they are being measured. (Test anxiety and evaluation apprehension are two exceptions.)

Then there is a whole set of effects associated with the dataproducer, the person providing the responses from which numerical data are obtained. That person may be either the subject to whom the data will be attached or an observer. In either role, that person reacts to the instructions given for responding, to the investigator running the research, to the content of the measuring instrument or record form (both the stems and the response format), and finally, in the case of the observer, to the target person whose

behavior is being assessed. Thus, there are three main types of method effects: (a) There are those associated with lack of congruence between the behavior from which data are extracted and the behavior described or implied in the conceptualization of the construct at which the measuring procedure is aimed. (b) Other method effects are associated with the intrusive consequences of imposing measuring operations on people. (c) Finally, some effects are associated with the impact of the measuring operations themselves on the person generating the measurements. (For a more extended discussion, see Fiske [1987] in which the problem is approached in somewhat different terms.)

At least for correlational and observational psychologists, part of the problem may be that, when we think about and conceptualize behavior, we tend to focus on behavior as it appears in the natural world, as if it were not affected by being observed and measured. We have enough trouble pinning down exactly what behavioral tendencies we are studying without adding the complication of what such behavior would look like under the constraining conditions of a measuring situation. Experimental psychologists may not face this difficulty: They may think about the behavior that occurs in their restricted context and what a planned manipulation will do to it.

There is another interesting difference. Throughout this chapter, we are assuming that the correlational psychologist trying to assess a given concept hopes that any change in method, made for convenience or economy, will not seriously affect the measurement of that concept. In contrast, the experimentalist manipulates some aspect of the situation intentionally, expecting that change to alter the measurements. In this approach, the dependent variable is measured in two or more contexts, and when two experimentalists try to employ the same manipulation, their two versions of it may differ—for example, in the tone of voice in which an instruction is given. The correlational psychologist may see the experimentalist as generating and measuring some effects due to method or procedure. Meanwhile the latter may view the correlational psychologist as naively expecting congruence between measurements made under different conditions.

Those considerations bring out an apparent conflict between two norms or ideals. In experimental psychology, the investigator is expected to describe his or her procedures so that, at least in principle, another investigator could replicate them and see if the first set of results could be reproduced. But I would argue that it is not desirable for an investigator who is explicating a conceptualization or a construct to delineate every minute detail of the measuring procedure for that construct: As noted earlier, we do not want each construct to refer only to a very particular and specific measuring operation. The conflict, however, is more apparent than real. We certainly want to have more than one method for measuring each of our concepts so that we can try to determine whether each method is contributing systematic

irrelevant variance. We also need to know the procedural details for each published study so that we can decide how much light the findings shed on the nature of and the interrelationships among the several concepts involved.

IDENTIFYING METHOD EFFECTS

Up to this point, we have been talking about method effects as two psychologists might discuss them over beer. We have said they exist, we have more or less defined them, and we have said that they are common and widespread. But suppose someone who has been listening interrupts and objects: "All this talk is fine, but now show me a method effect and tell me how big it is!" What should we do? Order three more beers?

To get at these effects, we would like to make a comparison, to say that, in this set of measurements, there is no method effect but in that set of measurements, there is such an effect. However, if one believes that method effects are universal or nearly so, then one can get no baseline data, no comparison data free from these effects. One can only say that the method effect here looks larger than the method effect there.

Let's go back to the original paper on multitrait–multimethod matrices (Campbell & Fiske, 1959). What it emphasized was the relatively high correlations between measures made by the same method, in the monomethod blocks, in contrast to the lower values for corresponding correlations in the heteromethod blocks: When two traits or other properties are measured by the same method, the correlation between them is bound to be higher than the correlation when the two are measured by different methods. We stressed the desirability of using methods that are as different as possible because it was obvious even 30 years ago that two methods could have many features in common, thus producing much the same method effects. For example, two self-report measures might differ only in response format, with one having the alternatives "Yes" and "No" while the other had "Yes," "?" and "No" so the subject could express uncertainty or avoid making a commitment.

The position I take here is simply that the identification and the size estimation of a method effect is a judgment, an appraisal of the evidence. Since the Campbell and Fiske paper was published, many techniques have been proposed for analyzing multitrait–multimethod matrices (see Marsh, 1988; Schmitt & Stults, 1986; Wothke, 1984). These are usually aimed at identifying method factors or method effects and assessing them. Each of these proposals seems to have important limitations. I agree with the conclusion of the review by Marsh (1988, p. 580): "There is currently no well-developed, generally accepted approach to the analysis of MTMM data." Does this suggest that the trait-method unit (the particular procedure as applied to the measurement of a given construct) identified in the original paper is an

organic whole that cannot be dissected? Is a trait measured by method X different from the "same" trait when measured by method Y? Is method Z different when used to measure trait A and when used to measure trait B?

One can examine a multitrait–multimethod matrix as a set of results. In that case, some statistical description may be useful for presenting or interpreting the findings. One can also look at such a matrix to see what might be improved in the next piece of research using those traits and those methods. Especially for that purpose, I continue to believe that a great deal of information can be obtained just by a thorough visual examination of the matrix. One can usually see which traits or attributes are not being measured well by the given methods and can judge whether the fault is in the methods or in the construal of the attributes. Similarly, one can judge whether each method is doing its task well or whether it is not discriminating among the attributes or is yielding much the same information as another method, but in less reliable form. At the very least, such an inspection of the matrix will provide a basis for interpreting the results of any analytic procedures that have been applied.

At this point, let us go back one step and ask "What is a method?" A method is a set of conditions in which behavior is observed and measured. But describing such a situation in full is about as difficult as defining a trait, an attitude, or a value. We can identify and specify many features of a measuring situation or method, but these are the external, objective aspects. We cannot describe the internal states of the people being measured, and we cannot even be sure what meanings they ascribe to such objective features of the situation as the stimuli or items and the instructions. Although some method effects from these sources may be inescapable, steps can usually be taken to minimize them.

A method is a complex entity. Some aspects of it are obviously relevant to measuring the variable being studied. Other aspects of it may be producing intrusive method effects. Suppose we believe that a method effect is present in some array of measurements: Can we identify the aspect or aspects of the method that have generated that effect? Usually, we cannot just compare two forms of a method, one with and one without a particular feature, to determine whether that feature has a predicted effect. What we have to do is to compare the method with that feature and the method with a different feature in its place, to see if the two features have different effects. For example, there have to be instructions to the data producer. We can compare two different sets of instructions but we cannot compare a set of instructions with no instructions.

Some method effects seem clearcut. For example, a recent review of cerebral lateralization asked whether cerebral specialization increases with age (Hahn, 1987). The author concludes that five experimental paradigms (or methods for studying the problem) agree that linguistic functions are localized in the left hemisphere from birth on. In contrast, for functions lateralized

in the right hemisphere, the results vary with the task: Some tasks show no change over development, others do show changes. Here, then, is an instance where results and conclusions vary with task or general method. Hahn also notes that the observed asymmetries are readily affected by variations within a given experimental paradigm, such as in the way that stimuli are presented, and are also affected by such psychological factors as reporting biases.

Method effects are also reported in a review of sex differences in group performance (Wood, 1987). When working either individually or in groups, men have been found to outperform women—in tasks or settings that favor men's interests and abilities over women's. This conclusion is hardly surprising. It is introduced here as yet another area in which the method—the task and the setting—affects findings.

Wood's review is also pertinent because she employed meta-analysis, a major methodological development that can help us investigate method effects. Meta-analysis is a systematic quantitative type of literature review. Unlike the typical qualitative review, a meta-analysis ordinarily lists all the studies reviewed and gives the criteria used to select them. This makes it possible for the reader to decide whether the reviewer's judgments were appropriate. The results from the several studies are converted into a common index and then summarized in various groupings. It is as though a researcher published all the raw data and then indicated the statistical analyses performed and the results obtained. Everything is explicit so that the reader can follow the reviewer's argument closely.

The steps in a meta-analysis parallel those in a research study. A general question is formulated and pertinent empirical data are compiled and weighed. Then more qualified questions are posed and the data are regrouped to answer them. The obvious difference between a research study and a meta-analysis is in our confidence in findings. We cannot place great faith in the findings of any one piece of research, taken by itself. But if several investigations study the same question and come up with essentially the same answer, the replications can give us a high level of confidence in that answer. If the replication is poor, that is, if the several answers show much variation, then the reviewer must examine the published papers to find the source of that variation, or devise a new empirical study to identify that source. Meta-analysis is a tool, not a cure-all. It provides answers, but the value of the answers depends on how insightful the reviewer's questions were.

Meta-analysis is important because it is the first comprehensive and systematic procedure for putting together empirical findings. We have a large body of methodological techniques for planning and executing single research studies but until now, we have had no systematic technique for integrating the findings from several such studies. Such integration is at least as important as the research itself (cf. Chapter 6).

Meta-analysis is a valuable tool for comparing and assessing method

effects. When a number of studies have been done on the same topic, it is possible to categorize them by method and see how the type of method relates to the outcome. To illustrate, the authors of the classic monograph on *The Benefits of Psychotherapy* (Smith, Glass, & Miller, 1980) classify outcome measures according to the degree of reactivity, a variable including the extent to which the client or the therapist could react to the measuring procedure, thus influencing the data. They found that Grade Point Average, physiological measures of stress such as the Galvanic Skin Response, and blind ratings of adjustment (in which the rater did not know whether or not the subject had had therapy) had the lowest effect sizes. (Effect size is a standardized index of the effect of a treatment. It has the advantage that it can summarize results from studies using diverse statistics to report their results.) Slightly larger effect sizes came from results using the MMPI. Much larger effect sizes were obtained from client self-reports to the therapist. The largest effects, twice the size of the GPA and GSR values, were found for outcome measures where the therapist rated client behavior produced in the presence of the therapist. The obtained benefits look larger if you ask the therapist than if you ask the registrar's office or a disinterested coder.

It is worth noting that the original multitrait–multimethod paper (Campbell & Fiske, 1959) was concerned with the validation of measures of individual differences. All of its examples were of that type. Since then, it has become quite evident that the concept of method effects applies much more generally: It is pertinent to all designs in psychological research. Method effects are associated with aspects of measuring procedures, not with any one type of research design (cf. Fiske, 1982). Comparisons between different types of psychotherapy may yield results that vary with the method used to assess outcomes. Similarly, studies of sex differences in group performance compare sex groups, not individuals, but the results are biased if the task favors one sex.

Returning to the problem of identifying method effects, there is a complication. The apparent size of a net method effect depends upon a variety of facts. For example, differences between self reports and peer ratings may vary with the type of trait or disposition being assessed and with each data producer's perception of the confidentiality that will be given to his or her ratings. In general, we can expect self ratings to be more favorable than peer ratings and to differ more from peer ratings than do outside observer judgments. We cannot, however, predict what will be found in any particular study. Also, along with fairly general kinds of method effects, there are specific or local types that the careful investigator must watch for. In survey research, for example, there can be effects associated with the institution or agency conducting the survey and with the year in which it was carried out (Turner, 1984).

The problem of context effects, another kind of method effect, has for years

plagued research using questionnaires or surveys. The content preceding a given question can affect the answer (Turner & Krauss, 1978). A thorough analysis of the problem and the literature has led Tourangeau and Rasinski (1988) to conclude that context effects can enter at each of the four stages in answering an attitude question: interpreting the question, retrieving relevant beliefs, rendering a judgment, and reporting an answer. They indicate the variables affecting each stage and the effects of each variable. They conclude that context effects are often unstable, perhaps because of the complexity of the processes involved.

More generally, we need to know a lot more about the diverse kinds of variability in responding. Messick (1989) notes that different people perform the same task in different ways and that the same person may perform differently across items or on different occasions. He argues that we should analyze the process going on when a person takes a test. Although a lot has been done on that problem (e.g., Kuncel & Fiske, 1974), a lot more needs to be done. Reiss (1986, p. 52) points out that "all methods for inquiry into nature are governed by the substantive laws of nature at the same time that these methods are the means for discovering and testing these laws." Because we do not know all the substantive laws applicable to behaving in the measuring situation, we must go out and develop our own minitheories for what goes on during psychological measuring (Fiske, 1988). At present, because we have available more assumptions than facts, "the conclusions drawn about a phenomenon are heavily dependent upon the assumptions that underlie the procedure used to investigate it" (Wyer & Srull, 1988, p. 145).

Better understanding of the psychological processes during measuring should enable us to improve the replicability of our measurements. The apparent circularity here (better replicability makes it easier to understand) can be taken as a spiral: An improvement in replicability makes it possible to gain better understanding which in turn enables us to improve the replicability of our measurements. Of course, we do have to have some replicability as a starting point. Byrne and Holcomb (1962) had the courage to report on their experience in trying to develop a measure of repressing and sensitizing defences. After obtaining disappointing results in cross-validation efforts, they went back to check on the internal consistency of their measure. It was .00! A similar result is said to have occurred during World War II when graduates of bombardier training were tested on two different days. The accuracy of their bomb drops had no stability. Obviously the concept of bombardier accuracy had to be rethought in terms of the total context and the involvement of the rest of the flight crew.

Replicability of measurements contributes to convergence among studies and replication of findings. But can we minimize method effects and thus get more replication of findings over different methods?

DEALING WITH METHOD EFFECTS

It is clear that method effects are present in psychological research. Now what can we researchers do about them? First, we can try to identify when these effects are more serious or more likely. But even that matter needs to be broken down by type of effect. Let us start with stimulus questions. Turner (1984) proposes that nonsampling effects are likely to be found when survey questions are ambiguous, when their concepts are amorphous, when the referents are vague, when the response categories are arbitrary and imprecise, and when the content has little importance to the everyday lives of the respondents. Some of these can be seen as properties of other poor measuring procedures. And they may remind some of you of the conditions under which response sets are more likely to operate: "When the task is ambiguous, when the response alternatives are vague or general, and when S is unable to give an exact answer that he can verify. To oversimplify a little, [response sets] enter when S is in doubt about what response to make" (Fiske, 1971, p. 215).

We might speculate that method effects are more likely to enter when investigators set up conditions that generate the behavioral processes that are measured: In other words, when we devise an artificial task for our subjects—a test, an inventory, an experimental challenge of some kind. In contrast, such effects are less likely when we unobtrusively observe an ongoing behavioral process that is independent of the measuring. This proposal does not deny that much valuable research is done in the laboratory under contrived conditions. Further, it is also true that unobtrusive observers can produce method effects unless they are thoroughly trained and are making simple judgments rather than inferences about the behavior (cf. Fiske, 1978, chap. 5).

Also relevant is the source of the concept that we are trying to measure, We often study concepts that come from everyday perceiving and attributing, and we are likely to be fully convinced that that concept, that attribute, exists out there ready to be measured. Problems with method effects seem more probable with measures of such concepts, perhaps because we are inclined to accept any plausible measuring techniques for them. In contrast, method effects may be less serious when concepts are developed from data, and especially when they refer to data obtained by explicit measuring operations—in other words, when we are working from the bottom up rather than the top down. We need some such low-level concepts on which to base our scientific work. They are also valuable in providing new perspectives, new ways of looking at and thinking about the phenomena we are trying to understand.

Preference is a familiar concept that illustrates this point. It is obvious, isn't it, that people have preferences and make choices? But Tversky, Sattath, and

Slovic (1988) raise the question whether, in choice theory and decision analysis, we can assume that there are preferences, pure and simple. Preferences are not invariant over alternative descriptions of options nor are they invariant over procedures for eliciting them. In some cases, preferences may be "actually constructed in the elicitation process" (Tversky et al., 1988, p. 371). Much of the writing about method effects has, like this chapter, been primarily descriptive or hortatory. Both the paper by Tversky et al. and that by Tourangeau and Rasinski (1988) go further: They take steps toward explanation, toward conceptualizations of the problem. The concept studied by Tversky and his associates, preference, is not trivial: It is basic not only to choice theory and decision making but also to studies of values and to economic theory. In addition, in psychological measurement, we commonly ask subjects to express a preference or choice. If each such decision is affected by the description of the options and by the way we elicit responses, the implications of the conclusions drawn by Tversky et al. may be much wider than they imply.

Tversky and his colleagues point out that "the assumption of procedure invariance [the same preference order emerging from presumably comparable procedures for eliciting prefences] is likely to hold when people have well articulated preferences and beliefs" (Tversky et al., 1988, p. 371). This view seems consistent with that expressed by Ericcson and Simon (1984) regarding the conditions under which verbal reports may safely be used as data: when subjects know in advance the kind of information they will be asked to report (so that they can know to what to attend), when the task is one that they can carry out, and when what is to be reported can be reported immediately (not after an intervening period of time).

To summarize, method effects are less likely to occur and are less serious when the person producing the data—the subject or an observer—is given a clear task that is easily carried out. To this we can add that the dataproducer should be motivated and not threatened. Such conditions generate replicable data. Such data seem necessary, but not sufficient, for obtaining procedure invariance and moving toward the generalizability of results from the given task to the target construct.

At this point, let me offer a highly speculative and probably highly controversial proposition. Some parts of psychology have made more substantive progress than others. Those parts that have progressed farthest are those where method effects are smaller, less serious, or less likely—are less of a problem. In those parts, the task given the dataproducer is clearly defined and fairly easy. I cannot prove that proposition empirically: To do so would require obtaining judgments about the relative progress that has been made by different areas of psychology and about the relative freedom from method effects within such areas. Even if I did that study, most psychologists would decide the validity of that proposition on the basis of their personal appraisals.

A case in point: Gerald Patterson's research Program. Although you may disagree with that speculation, I trust you will agree that the problem of method effects must be faced squarely. It is in our way and it is holding us back. One research program that has undertaken to deal with method effects is Gerald Patterson's studies of children and family processes at the Oregon Social Learning Center. The Center's work on construct-indicator relationships in this substantive area is summarized in Capaldi and Patterson (1989). A recent paper from that Center (Bank et al., 1990) describes the course of their efforts and shows their healthy blend of pragmatism and idealism.

They start with the belief that all assessment procedures introduce some degree of bias or distortion into the data they generate. They then show a willingness to explore a wide range of tactics for dealing with the problem. On the basis of empirical evidence as well as conceptual arguments, they are fully convinced that they cannot measure any of their constructs adequately by a single procedure; as a consequence, they have tried different combinations of scores from varied measuring procedures. One conceptually sound strategy was to find, for each construct, two nonoverlapping indicators (where the methods and the data producers were different) and combine their scores. The difficulty is that, with more than three or four constructs, one quickly runs out of measuring procedures.

Notable about the attack by Patterson's group on the method problem is their willingness to consider giving up earlier beliefs. For example, they recognized that the method variance problem was not always a matter of poor measurement; sometimes, it is a matter of the construct and the difficulty in defining it. Thus, they concluded that their construct of Family Stress was best indexed by simply counting the number of risks to which the family was subjected. This simple method was conceptually satisfactory because the alternative—considering Family Stress as a trait—was not acceptable in the face of the low intercorrelations among measures of it.

In their work, the matter of the data producer's perspective is central: A mother, a father, and a teacher typically see a child somewhat differently. To bring out that component, they introduce the concept of agent and distinguish between effects associated with agent and those from other aspects of method, such as the instrument used.

One of their final points deserves thoughtful attention. They note that, although the mother's view of her child may disagree with the views of the father, the teacher, or the peer, it is nevertheless a significant causal factor taken by itself. It may be of overriding importance for a family. More generally, divergences between measurements from different perspectives may indicate that, for both theoretical and applied purposes, more than one concept is necessary. There may or may not be any true core common to the several views.

There is a paradox about method effects. On the one hand, we readily accept evidence that, in self-reports, subjects may portray themselves in overly favorable terms. Similarly, we are not surprised to learn that substituting "welfare" for "aid to the needy" changes the endorsement rate for some survey questions. Such dispositions in responding seem quite natural. It seems just common sense that people should behave that way; we might do so ourselves if we were the respondents. So we have some psychological understanding of the origins of method effects.

On the other hand, we continue to use our old familiar methods in our own research. We have had experience with them. Typically, other researchers are also using them. It would cost a lot of time and effort to revise them or replace them. And any alternative ways of measuring our favorite concepts may have their own method effects. Is it surprising that people prefer to talk about their findings rather than look at the potential method effects in their data? It would be disturbing to investigate the possibility that one's findings are due to a method effect and that one's results cannot be replicated by a quite different method.

SUMMARY

The following didactic exposition summarizes this chapter in terms of advice to the psychological researcher. It is not new advice. Much of it is in the indoctrination we received in our graduate training. But this advice is not closely followed. My colleague Louis Fogg and I have been studying the critical comments that reviewers make about manuscripts submitted to editors of journals published by the American Psychological Association (Fiske & Fogg, 1990). It is evident that authors of many papers have not followed that standard advice, so it bears repeating:

1. Define each of your basic concepts fully. Expand each conceptualization to specify where it can be observed and how it is best measured—that is, what perspective should yield measurements conforming to the specifications for the variable and having minimal bias (cf. Fiske, 1987).
2. For each dependent variable, and for each independent variable where appropriate, find at least two measuring procedures that seem to have minimal method effects and that are as different as possible with respect to the samples of behavior used, the format, and the person producing the data. Keep in mind, however, the caution noted by Kagan (1988) and by Bank et al. (1990): If the two procedures yield quite different results, consider whether the data should be construed as reflecting one construct or two distinct ones.
3. Wherever feasible, use unobtrusive procedures so that the person whose behavior is being studied does not react to being studied.

4. Accept the fallibility or bias present in the products of each method. Carry out the basic analyses on the array of data from each procedure separately and then try to integrate the sets of findings. Aspire to replicating the findings across these procedures. If this approach yields unsatisfactory outcomes, consider pooling the data from your separate measures of each construct. Try to have a sensible a priori rationale for each such pooling.

5. In interpreting your findings, consider whether general, replicable method effects seem to be or could be present in each of your methods. Also look for specific, local effects that might have been generated by the particular conditions under which the study was conducted. It is obviously advantageous to identify potential effects before the study is carried out so that they can be minimized. In any case, the post hoc review is needed.

6. Before stating each conclusion, think through and explain the boundary conditions within which you believe it holds. (Is the only universal law that there are no universal laws?)

Throughout all your planning, your objective is to try to maximize reliability (reproducibility) of the measurements from each procedure and the replicability of results from one measuring procedure to another. When you can reproduce your measurements and replicate your findings, you know or will soon discover what psychological processes are going on in your research.

REFERENCES

Achenbach, T. M., McConaughy, S. H., & Howell, C. T. (1987). Child/adolescent behavioral and emotional problems: Implications of cross-informant correlations for situational specificity. *Psychological Bulletin, 101*, 213–232.

Bank, L., Dishion, T., Skinner, M., & Patterson, G. R. (1990). Method variance in structural equation modeling: Living with "glop." In G. R. Patterson (Ed.), *Depression and aggression in family interactions* (pp. 247–279). Hillsdale, NJ: Erlbaum.

Beach, F. (1950). The Snark was a Boojum. *American Psychologist, 5*, 115–124.

Byrne, D., & Holcomb, J. (1962). The reliability of a response measure: Differential recognition threshold scores. *Psychological Bulletin, 59*, 70–73.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*, 81–105.

Capaldi, D. M., & Patterson, G. R. (1989). *Psychometric properties of fourteen latent constructs from the Oregon Youth Study*. New York: Springer-Verlag.

Cowen, E., Pederson, A., Babigian, H., Izzo, L., & Trost, M. (1973). Long-term follow-up of early detected vulnerable children. *Journal of Consulting and Clinical Psychology, 41*, 438–446.

Cronbach, L. J. (1975). Beyond the two disciplines of scientific psychology. *American Psychologist, 30*, 116–127.

Cronbach, L. J. (1982). Prudent aspirations for social inquiry. In W. H. Kruskal (Ed.), *The social sciences: Their nature and uses* (pp. 61–81). Chicago: University of Chicago Press.

Cronbach, L. J. (1986). Social inquiry by and for earthlings. In D. W. Fiske & R. A. Shweder (Eds.),

Metatheory in social science: Pluralisms and subjectivities (pp. 83–107). Chicago: University of Chicago Press.

Ericcson, K. A., & Simon, H. A. (1984). *Protocol analysis: Verbal reports as data.* Cambridge, MA: MIT Press.

Fiske, D. W. (1971). *Measuring the concepts of personality.* Chicago: Aldine.

Fiske, D. W. (1973). Can a personality construct be validated empirically? *Psychological Bulletin, 80,* 89–92.

Fiske, D. W. (1978). *Strategies for personality research: The observation versus interpretation of behavior.* San Francisco: Jossey-Bass.

Fiske, D. W. (1982). Convergent and discriminant validation in measurements and research strategies. In D. Brinberg & L. Kidder (Eds.), *New directions for methodology of social and behavioral science: Forms of validity in research* (No. 12, pp. 77–92). San Francisco: Jossey-Bass.

Fiske, D. W. (1986). Specificity of method and knowledge in social science. In D. W. Fiske & R. A. Shweder (Eds.), *Metatheory in social science: Pluralisms and subjectivities* (pp. 61–82). Chicago: University of Chicago Press.

Fiske, D. W. (1987). Construct invalidity comes from method effects. *Educational and Psychological Measurement, 43,* 285–307.

Fiske, D. W. (1988). Measuring to understand and understanding measuring. In S. G. Cole & R. G. Demaree (Eds.), *Applications of interactionist psychology: Essays in honor of Saul B. Sells* (pp. 121–143). Hillsdale, NJ: Erlbaum.

Fiske, D. W., & Fogg, L. (1990). But the reviewers are making different criticisms of my paper! Diversity and uniqueness in reviewer comments. *American Psychologist, 45,* 591–598.

Gottman, J. M. (1977). Toward a definition of social isolation in children. *Child Development, 48,* 513–517.

Hahn, W. K. (1987). Cerebral lateralization of function: From infancy through childhood. *Psychological Bulletin, 101,* 376–392.

Hogarth, R. (Ed.). (1990). *Insights into decision making: A tribute to Hillel J. Einhorn.* Chicago: University of Chicago Press.

Hymel, S., & Rubin, K. H. (1985). Children with peer relationship and social skills problems: Conceptual, methodological, and developmental issues. In G. J. Whitehurst (Ed.), *Annals of Child Development, 2,* 251–297.

Kagan, J. (1988). The meanings of personality predicates. *American Psychologist, 43,* 614–620.

Kagan, J. (1989). *Unstable ideas: Temperament, cognition, and self.* Cambridge, MA: Harvard University Press.

Kuncel, R. R., & Fiske, D. W. (1974). Stability of response process and response. *Educational and Psychological Measurement, 34,* 743–755.

Marsh, H. W. (1988). Multitrait-multimethod analysis. In J. P. Keeves (Ed.), *Educational research, methodology, and measurement* (pp. 570–580). New York: Pergamon.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York: American Council on Education/Macmillan.

Murray, H. A., & others. (1938). *Explorations in personality.* New York: Oxford.

Parker, J. G., & Asher, S. R. (1987). Peer relations and later personal adjustment: Are low-accepted children at risk? *Psychological Bulletin, 102,* 357–389.

Reiss, A. F., Jr. (1986). Measuring social change. In N. J. Smelser & D. R. Gerstein (Eds.), *Behavioral and social science: Fifty years of discovery* (pp. 36–72). Washington, DC: National Academy Press.

Schmitt, N., & Stults, D. M. (1986). Analysis of multitrait-multimethod matrices. *Applied Psychological Measurement, 10,* 1–22.

Smith, M. L., Glass, G. V., & Miller, T. I. (1980). *The benefits of psychotherapy.* Baltimore, MD: Johns Hopkins University Press.

Tourangeau, R., & Rasinski, K. A. (1988). Cognitive processes underlying context effects in attitude measurement. *Psychological Bulletin, 103,* 299–314.

Turner, C. F. (1984). Why do surveys disagree? Some preliminary hypotheses and some disagreeable examples. In C. F. Turner & E. Martin (Eds.), *Surveying subjective phenomena* (Vol. 2, pp. 159–214). New York: Russell Sage Foundation.

Turner, C. F., & Krauss, E. (1978). Fallible indicators of the subjective state of the nation. *American Psychologist, 33,* 456–470.

Tversky, A., Sattath, S., & Slovic, P. (1988). Contingent weighing in judgment and choice. *Psychological Review, 93,* 371–384.

West, M. O., & Prinz, R. J. (1987). Parental alcoholism and childhood psychopathology. *Psychological Bulletin, 102,* 204–218.

Wood, W. (1987). Meta-analytic review of sex differences in group performance. *Psychological Bulletin, 102,* 53–71.

Wothke, W. (1984). *The estimation of trait and method components in multitrait-multimethod measurement.* Unpublished doctoral dissertation, University of Chicago, Chicago.

Wyer, R. S., Jr., & Srull, T. K. (1988). Understanding social knowledge: If only the data could speak for themselves. In D. Bar-Tal & A. W. Kruglanski (Eds.), *The social psychology of knowledge* (pp. 142–192). New York: Cambridge University Press.

CHAPTER 6

Replication and Meta-Analysis: the Story of a Meta-Analysis of Parents' Socialization Practices*

Hugh Lytton
Department of Educational Psychology
University of Calgary
Alberta, Canada

The truism that all progress in knowledge must be founded on earlier work and knowledge, that it must be cumulative, has been obvious to the natural sciences for a long time. The cumulative nature of science is perhaps most apparent in what Kuhn (1962) has called "normal science" which assumes the task of putting flesh and bones on a new "paradigm" or model, created by a prior revolutionary reconceptualization of an area.

Cumulation is not a glamorous activity, but it is an essential one and, I would suggest, it is no less useful and necessary in the social than in the natural sciences. Links to past research reduce the temptation to "reinvent the wheel" and deepen and enrich the meaning of a novel research project. Moreover, it is only through such links that we can measure any little progress we make and define the bounds of consensus that has been reached in a given area.

We should distinguish two aspects of cumulation in science: elaboration and aggregation. Elaboration involves solving new puzzles that build on a prior revolutionary breakthrough and a new perspective. However, we are more concerned with its second aspect, namely, the aggregation, synthesis, and corroboration of existing knowledge, essential to both natural and social sciences in their normal phase. This can occur in three different ways:

1. Replication of a past study, when the essentials of a previous study are repeated; often variations of some parameters are systematically introduced and noted.

*The financial assistance of the Social Sciences and Humanities Research Council of Canada and of the University of Calgary for the research project discussed in this chapter is gratefully acknowledged.

2. Aggregation of actual data from several samples, so as to produce a larger database, something that is often called "secondary analysis."
3. Integration of existing studies in a given area by means of a review. Reviews may be in narrative form, or they may be quantitative integrations of research findings, called "meta-analyses."

The following sections discuss these three types of research cumulation briefly.

REPLICATION

Students in the natural sciences are indoctrinated in the need for replication which for them is a matter of scientific necessity. A student will perform standard laboratory experiments, the results of which he will compare with a standard result. He will thus test his own experimental procedures and materials against a benchmark and learn to remedy his shortcomings; these experiences will be part of his formation as a scientist. However, no such standard experiences are available for a social science student.

As van IJzendoorn (Chapter 3) points out, replicability is a fundamental requisite of science and any claim for a new relationship or law will normally be subjected to a test of replicability at some stage, before it can pass into the realm of accepted knowledge. However, most research in the natural sciences will not be exact replication, but will build on past research by varying certain parameters, of materials, instruments, or procedures in order to test the limits of an accepted scientific law. The accumulation of replicable findings is necessary not only for the establishment of generalized laws by induction, but also to ensure that the deductive testing of hypotheses produces dependable, replicable results. The natural sciences have been able to make relatively rapid progress precisely because they have developed programs of research that have been cumulative in the sense of replication as well as elaboration. Thereby they have extended the bounds of dependable knowledge.

In view of the probabilistic nature of findings in the social sciences, replication is at least as necessary in this domain as in the natural sciences. But for a variety of reasons it is upheld more in principle than in practice and this is certainly the case in the area of child rearing or socialization practices. The absence of replication is partly due to its being considered a nonoriginal, boring, low-level task, and to the wish of every psychologist to create a new theory, or at least a new test, or a new observation or experimental method, rooted in the specific goals of the research.

It is, of course, true that, as Yarrow, Campbell, and Burton (1968) and van IJzendoorn (Chapter 3) point out, exact replication is difficult, if not impossi-

ble, to accomplish. But partial replication—say, of one instrument among several—also has its uses and in the social sciences, as in the natural sciences, replication can take the form of an endeavor to duplicate the essentials of a study, with systematic variations of some parameters ("varied replication"). Indeed, as part of the broader research enterprise such experimental variations are desirable, for we must be interested in various parameters that affect the limits within which some established relationship holds. If parents socialize 3-year-old children to be helpful by precept and example, do they do so with 12-year-olds? And what is the effect on 3-year-olds and on 12-year-olds?

However, in the social sciences there are certain obstacles in the way of replication that are absent in the natural sciences. For instance, in chemistry materials and procedures are precisely described with generally accepted vocabulary and can be duplicated within a small area of doubt. In parent–child research we generally lack such standard vocabulary and standard procedures. Hence, replications become more problematic, because it is difficult to know whether constructs and research procedures have been consistently employed across different studies.

Yarrow et al. (1968) discuss this problem under the title "The label and the measure." One form of the problem arises when the same conceptual label is applied to different measures, because two investigators interpret the concept differently: one example is the way "control" was defined and measured by the researchers at the Fels Institute (Baldwin, Kalhorn, & Breese, 1949) and by Baumrind (1971).

The converse problem also is not infrequent: The same actions may be subsumed under differing conceptual labels. Thus, a mother giving a young child spontaneous help with a cognitive task can be viewed as showing support or warmth, but she can also be thought of as exhibiting "anxious intrusion". Research projects that use such diverging interpretations of the same concrete behavior are difficult to equate.

Conceptual issues in replication. Just because these difficulties make data imprecise and relationships unreliable and difficult to interpret, replication of existing findings that provides continuity of constructs and measures seems essential. It should perhaps start with attempting to replicate the operationalization of some constructs, for example, of the "authoritative" childrearing constellation, and of the child's socially responsible behavior. Only when these constructs have been shown to be reliably and repeatedly measurable would their interrelationship become a target for replication. The function of such replication would be to determine the generality, as well as the bounds of the generality, of "classic" relationships.

Yarrow et al. (1968), perturbed by what they saw as the lack of consistency in previous research on the relations between presumed parent antecedents and child behavioral consequences, carried out a study which repeated prior

work for example, that by Sears, Maccoby, and Levin (1957). They also built in replications of their own measures, in that they repeated questions from their interview for mother in a questionnaire for mother, and also used similar questions for teachers. The criterion characteristics on which they focused were the child's dependence, aggression, and conscience.

These researchers were not reassured by their findings. From their work and their comparison with earlier research in this area Yarrow et al. (1968) draw very pessimistic conclusions: They doubt that sound evidence really exists to support commonly held hypotheses concerning the factors crucial to conscience formation, or to the development of dependence or aggression. However, the authors somewhat understate the consistency they themselves found, for example, for the effects of parental warmth. Other writers, who reviewed a rather broader range of research in the same area by means of narrative reviews, have found a more dependable consensus on a number of factors, such as warmth and power assertion, which generally seem to have positive and negative effects, respectively, on internalized standards (Grusec & Lytton, 1988; Hoffman, 1970; Maccoby & Martin, 1983).

We owe an example of replication across four different samples, studied by the same authors, using the same measurement instrument (naturalistic observation), to Clarke-Stewart, VanderStoep, and Killian (1979). This research also was concerned with mother–child interactions and their interrelations in a variety of areas. Out of 58 significant correlations between all mother and child variables, only 5 (9%) were significant in all four samples. What marked out these replicated relations was the fact that they were more often predictable from previous research, theory, or common sense than were unreplicated relations; thus, mother's positive affect and play with the child was highly correlated with the child's positive affect and play with mother—hardly a surprising finding. It is also notable that these replicated relations tended to be based on, composite measures or standardized tests. Again, the degree of consistency was perhaps understated, since when the p values for the four samples were combined to form a composite p, 58.6 percent of possible replications turned out to be significant.

One reason why some of the findings in Yarrow et al. (1968) and in Clarke-Stewart et al. (1979) proved elusive and unreplicable may have been the fact that the variables were represented by single-occasion measurements. Epstein (1980) points out that a behavior observed on a single occasion, whether in the laboratory or elsewhere, is likely to be so situationally unique as to be incapable of generating reliable, repeatable results. One solution for obtaining replicable findings is to aggregate behavior over situations or occasions, stimuli or measures, thereby canceling out incidental, uncontrollable factors. Epstein (1980) demonstrated the utility of aggregation by showing that, when self-ratings, behavioral records, and physiological

indicators were aggregated over 12 days (in the same study), the reliability coefficients rose to high levels.

One instance of a clear and rigorous replication in the area of parent–child relations, which also incorporated improved methodological and analytical features, is the replication by Hubbard and van IJzendoorn (1987) of Bell and Ainsworth's (1972) classical study of infant crying and maternal responsiveness. The results were contrary to Bell and Ainsworth's conclusions that maternal unresponsiveness tends to increase subsequent crying by the infant.

Occasionally certain instruments or procedures become established and attain popularity—often because of their ease of administration—and this applies particularly to laboratory situations. While they remain fashionable they elicit a whole host of varied replications, though the studies are often not designed as replications. Ainsworth's Strange Situation is, no doubt, the most prominent case in point and it receives attention in Chapter 3. Note that in these studies the goal of replication is the classification of the child's attachment status in varying populations, and so on; they are not, strictly speaking, investigations of parent–child relations.

When we examine the study of the influence of parental attitudes or behavior on children's development, we find replications to be extremely rare. A recent review of parental childrearing attitude questionnaires concluded: "it is only somewhat facetious to say that when investigators wanted to sample attitudes, they developed a new instrument. Consequently, few instruments have seen repeated use; replication studies are almost nonexistent" (Holden & Edwards, 1989, p. 52).

The need for replication. Several factors underline the need for replication of socialization studies. First, significant findings from a single study, for example, correlations between parental and child variables, must be viewed with circumspection, as they may be due to *chance factors*, for example, of subject selection, and their replicability should first be demonstrated, as Clarke-Stewart (1979) has pointed out in connection with her own experience.

Second, in addition to chance factors, *personal preconceptions and attitudes* may influence the actual research process, as well as interpretation. This danger is particularly acute when emotional reactions to a given issue are strong, for example, regarding sex differences. Thus, Eagly and Carli (1981) found in studies of influenceability and of decoding of nonverbal communication that both men and women researchers portrayed their own sex more favorably than did researchers of the opposite sex. I am not suggesting that deliberate deception or suppression of data are occurring, but rather that the data gathering and the analysis process may be affected by subtle influences, of which the researcher is quite possibly unaware.

Conflicting conclusions are often a matter of interpretation of the findings, rather than of the actual findings themselves, and interpretations are powerfully influenced by the writer's theoretical preferences. Thus, Lewis (1981) interpreted the effects of "authoritative" parenting as supporting attribution theory, whereas Baumrind (1983) interpreted these effects as supporting social learning theory. It is an inescapable fact that interpretations are not value-free, and researchers' ideology and point of view will play a part in them. A simple replication will not provide an easy remedy for this problem, but even in these cases judicious replication of a study under conditions that will throw light on the point at issue, will often be a guide to the sounder interpretation of the original findings.

Moreover, lack of replication results in the overediting of data to achieve a publishable paper, or even in the partial faking of data. Thus, lack of replication in social psychology means absence of essential social control which exists in sciences where replication is routine (Campbell, 1986).

Secondary Analysis

Secondary analysis is the reanalysis of original data by means of better statistical techniques or an enlarged sample. A kind of secondary analysis, called "cluster analysis," which aggregates data across a set of studies, has been proposed by Light and Smith (1971). It requires a set of studies that are methodological replications and for which original data are available. If no important differences in study outcomes are found, the data can be combined in order to base conclusions on a larger sample size. Because original data are rarely available, it is not surprising that this method has not been used often.

However, occasionally such methodological replications become available by accident, though they can hardly be called systematic in that case. Thus Bradley et al. (1989) brought together six independent investigations, involving 11 researchers across North America, who had all investigated the relationship between maternal environment and the child's development. It so happened that all these projects used many of the same measures, in particular a common observational measure of the home environment, the "HOME." These accidental replications were then submitted to a combined data analysis, that is a secondary analysis.

Meta-Analysis

The third and broadest way of cumulating evidence and rendering conclusions more robust is a review of the existing literature in a given area. Such a review has often taken on narrative form which, however, has several defects. First, it is subjective in that there are no rules spelled out how to resolve conflicts between studies. Reviewers will do this subjectively and may

quite honestly disagree on the weight and importance to be attached to different studies in the resolution of such conflicts in the overall conclusions. The problem here—that of subjective, differential weighing up of evidence—has its parallel in clinical prediction (cf. Dawes, Faust, & Meehl, 1989). Also reviewers may be attacked on the studies they have chosen to include or exclude; while this problem also persists in quantitative integrations of research, the rules for inclusion there are made explicit and cannot be deviated from at a whim.

Second, the narrative review is scientifically unsound, if it bases conclusions on "vote counts" or "box counts" of results pointing in one direction or the other (e.g., characteristics present more with males or more with females), with the reviewer then basing his conclusions on the side that receives most "votes." This procedure ignores sample size and effect size.

Third, the traditional review is also inefficient in extracting meaningful information from the existing literature. When only a dozen or so studies existed in any given area of research it was possible to summarize and make sense of these in verbal form. The exponential growth of research in the social sciences has meant that now there typically are hundreds of studies available in a given area. It is more difficult to obtain an understanding of general trends and mentally to juggle relationships in a large number of studies by intuitive methods. Hence, a need has arisen for a method of integrating these in a quantitative way that would parallel the quantitative analyses of the primary studies. As Glass (1977) has put it: "The accumulated findings of . . . hundreds of studies should be regarded as complex data points, no more comprehensible without the full use of statistical analysis than hundreds of data points in a single study could be so casually understood" (p. 352).

In the methods of aggregation discussed under "Secondary Analysis," data points for individuals are aggregated over several occasions, or measures, or samples. If we aggregate study outcomes, that is, summary characteristics of studies, say, a difference between groups or a relationship characterizing the sample, we have a meta-analysis.

The more difficult the execution of exact replication is (as in parent–child research—see earlier) and the more it is spurned by investigators, the more important becomes meta-analysis in order to be able to bridge the differing operationalizations and provide a view on the generality of findings *across* such differences. Quantitative techniques for research integration have been available for decades, for example, as long ago as 1954 in a technical chapter by Mosteller and Bush. But it was Glass and his colleagues who coined the term "meta-analysis" and their work made the method more commonly known and accessible (cf. Glass, 1977).

The integration of data from diverse studies is beset with difficulties, whatever method one adopts. However, meta-analysis attempts to tackle

these with quantitative procedures that are common in the social sciences. Meta-analysis provides a basis for detecting regularities and trends in a set of studies with conflicting results by looking for such possible trends in the data systematically. Its procedures do not by themselves overcome all problems of disagreement and conflict (see later), and they do not eliminate subjective judgments by the reviewer. But, in contrast to verbal summaries, these judgments are made explicit, their consequences can be assessed, and hence replicability and public accountability may be assured. Since these procedures also make a priori contrasts and regression analyses possible, generalized conclusions can be drawn from a bewildering array of studies with appropriate qualifications arising from moderating factors, for example, the differing ages of the children studied. Theoretical questions of importance, it is hoped, can thereby be answered with greater clarity than by the traditional verbal review. How far such hopes can be translated into reality will be discussed below in relation to the story of our meta-analysis.

We should note that meta-analysis does not replace careful experimental replication which has its own value, noted earlier. But meta-analysis goes beyond replication in that it integrates a diversity of approaches to the same topic. One criticism of meta-analysis that is often voiced is that it tries to integrate incommensurable studies, something that has been called the "apples and oranges problem." It is true, of course, that the results of different studies to be integrated have to speak to the same issue, or address the same topic. But what may be considered "same" is not at all clear and depends on the purpose and perspective of the meta-analysis. Take the example proposed by Glass (1977): Imagine the Secretary of Health and Welfare comparing 50 studies of alcoholism treatment, 50 studies of drug addiction treatment, and 100 studies of the treatment of obesity. Though each of these would form a separate meta-analysis, if the Secretary finds the results for the first two to be negative, whereas the results for the last one are positive, he may decide to fund only obesity treatment centers. From the Secretary's point of view, the problem is public health, not alcoholism or drug addiction or obesity separately.

No doubt, it has little meaning to combine studies of the effects of chemotherapy for cancer and of psychotherapy for anxiety. But suppose a researcher wishes to compare the relative effectiveness of "talking psychotherapy" and of behavior therapy: Studies on these topics might be submitted to separate meta-analyses. But then, when the issue is the effectiveness of any type of psychological therapy versus no treatment at all, the effects of the two kinds of therapy would be considered jointly; it would have been at least equally efficient to have submitted them to one meta-analysis, with "Type of therapy" as a stratifying factor. So it is the problem under consideration that determines which studies are sufficiently alike to form a common class to be

combined for a meta-analysis. I discuss the problem of "lumping" in relation to our meta-analysis later.

The fact that meta-analysis integrates *diverse* approaches to a problem—its essential purpose—makes it more useful than replication for resolving conflicts between differing interpretations of existing findings and the trend they may show. Hunter, Schmidt, and Jackson (1982) note: "In many areas of research, the need today is not additional empirical data, but some means of making sense of the vast amounts of data that have accumulated" (p. 26). But meta-analysis is, of course, not possible unless data on a given problem *have* been accumulated: Meta-analysis feeds on replication, both exact and varied.

Take the question of the impact of environment on IQ scores. A glaring example of how verbal reviews of a group of studies can lead reputable scientists to diametrically opposite conclusions is the debate between Munsinger (1974, 1978) and Kamin (1978). Munsinger (1974) examined a group of adoption studies and concluded that environmental effects are small. Kamin later reviewed the same set of studies and arrived at the opposite conclusion. Personal beliefs are bound to enter into how reviewers resolve conflicting findings and this will explain part of the disagreement. But further, both reviewers accused each other of selectivity in reporting. Short of reading the whole original literature themselves—which Kamin actually recommends—readers are at a loss to know which conclusion to draw. If the authors had used systematic procedures for summarizing this literature and examined the disparities between studies in systematic ways by relating them to moderating factors, this would have improved the situation, since the reader would at least be able to relate the difference in conclusions to known aspects of the reviews and draw his or her own conclusions from this.

Meta-Analysis of Differential Rearing of Boys and Girls

The question to what extent parents treat boys and girls differently in their socialization practices is an important one since such differential socialization may be at the root of some or all of the behavioral differences known to exist between males and females. Indeed the theory that sex-typing is brought about through "shaping" by parents is a widely held view. Maccoby and Jacklin (1974), in their monumental work, *The Psychology of Sex Differences*, have given a great deal of space and thought to a review of studies in this domain, but they reviewed research in this area by counting the studies that revealed sex-differentiated practices and those that did not (the "vote-counting" method), and then characterized the set of studies qualitatively. They concluded that parents do, indeed, reinforce sex-typed behavior, narrowly defined, for example, they reward traditionally sex-typed play activities and toy choices. Apart from this, however, they came to the conclusion

that "The reinforcement contingencies for the two sexes appear to be remarkably similar" (p. 342).

This conclusion has been criticized, particularly by Block (e.g., 1976, 1978) in several extensive articles. Block argued that Maccoby and Jacklin underestimated the extent to which parents do apply differential socialization practices to the two sexes and she attributed these, in her opinion misconceived, conclusions to various shortcomings of the Maccoby and Jacklin review, for example, the fact that most of the studies they reviewed concerned children less than 6 years old, and the fact that fathers' treatment was reported and analyzed much less frequently than was mothers'. Maccoby and Jacklin, in fact, themselves recognized these limitations of their review which, however, are essentially limitations of the data base they had to work with. Block also criticized the authors for giving the same weight in their interpretation to each study, regardless of its quality or sample size. Further, she disagreed with the authors' classification of studies under different rubrics. Her opinions, too, however, represent her own subjective judgments and are also based on a qualitative inspection of the studies cited by Maccoby and Jacklin, plus others; she did not attempt to provide a quantitative integration of the domain.

These qualitative reviews agree that parents make differences between boys and girls in some areas of socialization, but disagree in which areas important differences occur or what the extent of the differential treatment is. Their disagreements arise from finely shaded nuances of interpretation of essentially the same studies.

In the last 15 years, a great many more investigations of sex-related differential socialization have appeared. My colleague, David Romney, and I therefore decided the time was ripe for carrying out an updated review of this domain that would take into account the criticisms of the earlier review and that would, in the spirit of the age, proceed by way of a quantitative integration of research that could resolve some of the disputes engendered by narrative reviews. Such a meta-analysis should have a certain conceptual base: It would have to distinguish between different areas of socialization (e.g., warmth, encouragement of achievement, encouragement of sex-typed activities) and it would have to take account of various moderating factors, particularly the child's age. As Block (1978) points out, a conceptual framework of socialization behavior would require distinctions between different ages, as parents will adapt their approach to the emerging capabilities and needs of the growing child.

We expected the quantitative integration to answer clearly two questions:

1. In which areas of socialization do mothers and fathers (analyzed separately when possible) treat sons and daughters differentially and what is the extent of the differences they make?

2. Are there any moderating factors, for example, age of child or design of study or year of publication, that affect the size of the difference?

The meta-analysis would not allow us to decide on the origins of any differential practices found, but it seemed to us important to find out more precisely in which areas, at which ages, and to what extent mother and/or father treat their sons and daughters differentially. In areas where differential treatment is found its implications can be discussed—and they are not always straightforward. In areas where *no* differential treatment is found, parental treatment can then be, at least tentatively, excluded as a cause of possible existing differences between males and females (cf. Lytton & Romney, 1991).

THE STORY OF OUR META-ANALYSIS

I now tell the story of our meta-analysis, as it actually proceeded, and I do not flinch from mentioning the difficulties, confusions, and false starts, the subjective decisions we had to take on various research strategies and how these decisions had to be changed in the light of circumstances.

The main steps to be executed in a meta-analysis are the following:

1. Identify and obtain relevant and usable studies.
2. Code study characteristics.
3. Extract one or more effect sizes from each study.
4. Calculate meta-analytic statistics.

Identifying and Obtaining Relevant and Usable Studies

With the help of a research assistant we searched a number of sources for possibly relevant material. *Relevant* was defined as an article, book chapter, or dissertation that examined the treatment of sons and daughters by mothers and/or fathers and contained data on the *level* of this treatment for the two sexes: say, amount of praise for sons and daughters. The minimum of quantitative information that we could use for inclusion in the main analysis was a p value for the difference between the treatment of the sexes.

We obtained items from Maccoby and Jacklin (1974), from a number of reviews of the field, from my own reference lists, as well as from lists kindly supplied by other researchers in the area. The main source of material was a computer search of *PsycLit* for articles, and of *Dissertation Abstracts* for dissertations, using relevant key words, directed essentially at the domains of socialization and sex-related differences. Additionally, a manual search of *Psychological Abstracts* was undertaken.

Several steps were involved in these searches: First, potential material was

identified in Abstracts of studies, then the actual articles, book chapters, or dissertations were inspected to see if they seemed to contain the essential quantitative information, those that seemed to were xeroxed, and then a second inspection narrowed the selection to those items that did, in fact, contain usable quantitative information. The period covered was from 1950 to 1987 for articles and books, and 1980 to 1987 for dissertations. We looked at least at the Abstracts, and often at the text, of about 1250 such items and ended up with 172 usable studies with quantitative information from North American and other Western countries, which translates into a hit rate of about 14 percent. Occasionally an article that had analyzed parental treatment of boys and girls separately concluded simply with the statement that "no significant sex difference was found" for a given treatment. Twenty-eight such articles were assigned to a separate "No p value" file and used only in a secondary analysis (see later). The selection process was quite arduous and took about five months.

Among the items we first picked up were studies that were not empirical in nature, studies that found sex differences, but did not report any statistical data on them, studies whose measures were not compatible with our categorization scheme (see later), and studies that reported not level of parental treatment, but correlations between parental treatment and child characteristics: All these had to be discarded.

Some subjective decisions inevitably enter into the selection of "relevant" studies. For instance, one study, entitled "Why 'Mom is the greatest': A content analysis of children's Mother's Day letters," found that in "Mother's Day letters," solicited by the author, there was a sex difference in the frequency with which mothers' modeling and teaching were mentioned. Should this weak and circumstantial evidence be included? We decided not. Sometimes, too, it is not easy to see whether the statistics provided in a study are usable for a meta-analysis, because even statistics that are in an unpromising form may often be converted to an effect size (cf. Rosenthal, 1984). In all doubtful cases like these we consulted each other and reached a consensual decision.

What did we miss? In view of all the supplementary searches that we carried out I doubt that, within the specified periods, we missed more than a handful of articles that might have contained quantitative data. On the other hand, there may well be hundreds of studies out there that investigated some topic other than sex-based differential socialization practices, but that in passing looked at such differential treatment and then concluded with a bald statement: "No significant differences due to sex of child were found." We noted some of these, but we might well have missed many. If we had been able to include all the nonsignificant effects, this would only have further decreased our estimates of the average effect sizes of differential treatment

which are already quite low (see later). Their omission is therefore not likely to pose a serious threat to our conclusions.

It will be obvious that our search of English-language publications suffers from linguistic parochialism and does not claim to have covered the literature on socialization worldwide. Not only will there be many studies in French, German, Dutch, Russian, and so on, that we did not touch, but the 17 studies from "other Western countries" in English that we included may quite possibly be unrepresentative of publications in their respective countries, for example, Holland, France, Norway, Israel. We also finally had to discard six studies from Third World countries, first identified, as a separate analysis would have been needed and they formed too small a database for this.

Coding study characteristics

We coded various aspects that characterized a given study and that could represent "moderator variables" which might explain variability in study outcomes, on a coding form that went through several draft versions. Data collection method was coded, ranging from indirect assessment of behavior—report by child—to the most direct access to behavior—observation or experiment. Source of study (journal, dissertation, etc.), social class or education level of parents, age of children, birth order, ethnicity, year of publication, sex of author, and sample size were all coded. Quality of the study was rated by us on the basis of our judgment of the internal validity of the study's procedures, for example, carefulness of data collection, including reliability of measures, and appropriateness of operationalization of constructs. On all these variables, which were later related to study outcomes as moderator variables, we reached a high degree of interrater reliability. Country of the study and sex of parent for whom measures were available were noted and these determined which accumulation of studies the given study was assigned to, as we analyzed separately for North America and other Western countries. We also analyzed separately for mothers and fathers, where possible, as well as jointly.

The question of how to deal with several outcome variables in the same study was a major problem in this research, as most studies of parental socialization contain several indices of socialization practices, such as praise and punishment, or demands and warmth, and each of these may be assessed by several measures. Theory suggests that the differences in socialization pressures between boys and girls may vary in different domains, and we therefore adopted the principle that outcomes should be pooled only when they represent the same construct (cf. Bangert-Drowns, 1986). Hence we established separate categories for different socialization areas, namely:

amount of interaction
encouragement of achievement
warmth, nurturance
encouragement of dependency
restrictiveness/low encouragement of independence
disciplinary strictness
discouragement of aggressiveness
encouragement of sex-typed activities
clarity of communication.

Several of these categories were also divided into subcategories, for example, "disciplinary strictness" was divided into its physical and nonphysical aspects. Each of these was submitted to a separate meta-analysis and therefore no more than one dependent variable appeared in any one analysis. But if a study contained several measures of one socialization area such as warmth or encouragement of sex-typed behavior, these were combined into a composite and one effect size derived for the socialization area (see later for further discussion of this procedure).

The classification of study variables into these socialization areas was no doubt the most difficult and debatable aspect of coding study characteristics. The categories were derived from a knowledge of the literature and from inspection of relevant articles and books. The justification for this particular categorization system is that it is rational and based on an informed view of the literature. Of course, other experts in the area will disagree with specific decisions, but then their categorizations may also be disputed, since no universally agreed classification system of socialization practices exists. Here, then, we have subjective decisions entering into the meta-analysis, but the consumer of the meta-analysis has the protection of seeing the decisions made explicit and public (something which was also done in the more traditional, but careful review of Maccoby and Jacklin). Others are therefore at liberty to repeat the meta-analysis, using their preferred categorizations.

Authors use diverse labels for the parental behavior that they observe or record, and it was our task to fit these labels to our categories, each of which, by the nature of a meta-analysis, had to be applicable to a number of studies.

Fiske (1986) thinks that the unrepeatability and lack of generalizability of knowledge in much of social psychology is due to the difficulty of identifying interchangeable categories of protocol (behavior) from one investigator to another. The meta-analyst has to be convinced, of course, that the different labels the primary investigators used are, in fact, interchangeable names for what, on a theoretically reasonable basis, can be subsumed under one category of behavior. If one or two authors develop a new perspective on child-rearing practices and introduce a measure not used by others and that cannot reasonably be fitted into any existing category, this measure—how-

ever interesting and insightful—by definition is idiosyncratic and cannot be used in the meta-analysis. Several instances of this occurred, for example "complexity of language," or "flexibility in play," or "power," defined as comprising all possible methods of influence; such constructs had to be excluded from the investigation.

Some categories, first established, had to be abandoned later, when we did not find sufficient instances of them in the literature. Thus, "Responsiveness," first a separate category, was combined with "Warmth, Nurturance" for analysis purposes.

The problem of excessive splitting or excessive lumping of constructs besets many investigations. We tried to deal with it by establishing more fine-grained socialization areas in the first place and analyzing these separately, and then lumping several socialization areas together into meaningful "Super" socialization areas and analyzing these as well. The "Super" socialization area of Disciplinary strictness, noted above, is an example. Both individual and "Super" socialization areas were analyzed, in case a subarea should exhibit a significant effect not apparent in the "Super" area. This did, indeed, happen in one instance, but the small number of studies involved in the individual socialization area throws doubt on the generalizability of the finding.

Most measures could be fitted to one or other of our categories, but the question which one a variable should be equated with was the debatable point. We used the authors' own categorization, as indicated by their interpretation of a variable, whenever possible, for example, noncontingent help-giving for young children was included in "Encouragement of dependence," as the author (Rothbart, 1971) related it to dependence. But consistency of classification was all-important to ensure homogeneity of meaning for each category across studies, and this sometimes had to override authors' classifications.

However, we also had to be sensitive to the differing contexts in which similarly labeled behavior occurred; thus helping with school work, as opposed to the help-giving for young children, noted earlier, was classified as "Encouragement of achievement." When touching, holding, and rocking were noted in a study in relation to efforts at arousing infants, it was coded as "Stimulation of motor behavior," but when, in another study, holding was interpreted as a manifestation of closeness it was coded as "Encouragement of dependence." In some cases we could see grounds for classifying a given childrearing behavior in more than one way, and then we based our decision on the most common practice of classifying the behavior. One hard case was demands for help in the house or for orderliness which we classified as "Restrictiveness," in keeping with Maccoby and Jacklin (1974), though we could understand the argument for counting this as "Encouragement of mature behavior." A special trial analysis demonstrated that taking these

items separately made practically no difference to the effect size, which was minimal in any case, so they were kept in the category "Restrictiveness."

While the main responsibility for these classifications was mine, in some doubtful cases a decision was arrived at in consultation with my co-investigator. Independent classification of a subset of studies by the two of us showed high agreement.

Extracting Effect Sizes from Each Study

In this step the meta-analyst is overwhelmed by as many different choices and decisions as in allotting study variables to socialization areas. The decisions made are often crucial to the outcome of the meta-analysis, without there being any clear theoretical or logical rationale for the superiority of one decision over another. I describe the choices that faced us and the decisions we made.

An effect size is the ratio of the difference between the means of two groups to the standard deviation, that is, the difference between the means in standard score form. The first decision, then, was to determine which standard deviation to use. Glass (1977) recommends the use of the standard deviation of the control group. However, in a comparison of the treatment of boys versus girls there is no "control" group; both groups represent a normal population and no difference in status exists between them. Hence, we decided that the pooled standard deviation of the two groups was the more appropriate statistic to use, as other authors also recommend. Hunter et al. (1982) also point out that the pooled standard deviation has the advantage of a smaller sampling error than the control-group standard deviation.

Some studies investigated the same domain and measured the same variables by different methods, for example, by behavior observation and by interview about childrearing attitudes. Then the choice to be made is which comparison is to be included in the meta-analysis, or whether they can be combined to yield a single effect size, since only one effect size from each domain in any study can be included. In one such case we chose to include the data derived from behavior observation and exclude those from the interview, as the former gave rise to more definable constructs that fitted into our scheme of socialization areas. In another case we combined data from an experiment and from home observation, as they dealt with the same socialization area, though they were reported as separate studies in the same paper. Others might make different decisions on these matters.

That such decisions are not trivial in their impact on the outcome has been shown by Matt (1989). He replicated part of Smith, Glass, and Miller's (1980) meta-analysis of psychotherapy studies. Smith et al. (1980) had applied the rule of omitting "conceptually redundant effect sizes" in a given study that is, effect sizes for the same outcome type. Matt selected 25 of the original studies

at random and three coders independently identified and calculated effect sizes. Each of these trained coders identified more than twice as many effect sizes per study as did Smith et al. (1980)—as Orwin and Cordray (1985) had done, too, in an earlier reanalysis of part of Smith et al.'s database. Moreover, the average treatment effects of all effect sizes identified by Matt and his colleagues were about half the magnitude of the smaller number of effect sizes found by Smith et al. (1980). Thus, it seems that Smith et al. (1980), in applying the conceptual redundancy rule, tended, whether consciously or not, to omit smaller effect sizes. Applying other rules of omission, Matt (1989) found, changed effect size estimates in different ways. It is clear, then, that the average effect size found in a meta-analysis may depend heavily on judgments about which comparisons to select.

As I explained earlier, when several measures of the same construct or socialization area were present in any one study we averaged these. Hence, no "redundant" effect sizes, strictly speaking, could occur in our meta-analysis. Nevertheless, picking the appropriate measures for a given theoretical construct involved judgment. The trial analysis of seven studies containing a measure for "Encouragement of mature behavior," mentioned earlier, showed that using a different decision rule from our original one made no difference. In view of the relative nonsignificance of most effects in this meta-analysis this conclusion most probably holds in other cases, too. But, short of redoing the meta-analysis, we have no means of being sure.

It will be understood that means and standard deviations for comparison groups (treatment of sons and daughters, in our case) are the basic grist for the meta-analytic mill. However, many other statistics, such as t, F, χ^2 can be converted to an effect size, following procedures set out by Rosenthal (1984), Glass (1977) and Hedges and Olkin (1985). But it takes some thought to ascertain whether a study's statistics are, in fact, usable. We decided, for instance, after some reflection that even an unstandardized partial regression coefficient, plus its standard error, would yield an effect size via t. In spite of the wide variety of statistics that can be used for deriving effect sizes, many articles, even by distinguished authors, do not contain sufficient information for this purpose.

In some studies results are reported in the form of simple dichotomies, that is, the proportions of boys and girls receiving a given treatment. Though the information is cruder than metric information it is convertible to effect sizes, too, and we did so by calculating the difference between the probability of boys and the probability of girls receiving a certain treatment, divided by the pooled standard deviation.

To give readers an idea of the complications involved in extracting and calculating effect sizes, I present a table from Burton, Maccoby, and Allinsmith (1961) which we used (Table 6.1). The Table presents the *numbers* of boys and girls for whom mothers used certain approaches and techniques and

Table 6.1. Childrearing and Child Behavior Related to Resistance to Temptation

S.A		RESISTANCE TO TEMPTATION: Boys High	Boys Low	Girls High	Girls Low	Significance
50	*Mother's attitude toward cleanliness*					Boys** Girls
	Relaxed	13	2	4	5	
	Concerned	9	16	16	12	
	Does S understand what cheating means?					
	Yes	10	7	5	9	
	No	12	11	15	7	
	Usual technique of discipline					
	Reasoning	0	0	1	1	
	Visible suffering	0	0	0	1	
61	Withdrawal of love	0	0	1	0	
	Isolation	7	8	6	5	
	Deprivation of privilege or object	3	4	0	4	
	Scolding	2	2	7	3	
62	Physical punishment	10	4	4	3	
	Scolding and physical punishment vs. all others: Boys Girls Combined: $\chi^2 = 2.79$ ($p < .10$)					
100	*Frequency of Mother's use of reasoning as teaching technique*					Boys Girls Combined: $\chi^2 = 5.41$*
	High	8	12	5	9	
	Low	14	6	15	8	
	Use of "I'm not going to talk to you until you behave better"					Boys* Girls
	Yes	12	3	5	6	
	No	10	15	15	11	
	Use of deprivation of privileges					Boys Girls*
	High	10	8	8	13	
	Low	10	10	12	3	
30	*Frequency of Mother's playing with S just for her own pleasure*					Boys Girls*
	High	14	8	10	15	
	Low	8	10	9	2	
	S's initial anxiety in the experiment					Boys Girls*
	High	5	9	8	1	
	Moderate	8	7	8	9	
	Low	8	3	4	7	

Note. Throughout this table the variable being considered was divided at the median.
* $p \leq .05$. ** $p \leq .01$.

From "Antecedents to Resistance to Temptation in Four-year-old Children," by R. V. Burton, E. E. Maccoby, & W. Allinsmith, 1961, *Child Development, 32,* 689–710. Reprinted by permission.

to whom they displayed certain attitudes. Note first that it is not stated in Table 6.1 or the surrounding text that the figures displayed here are numbers and not percentages—we had to figure this out by calculations. Having decided that these numbers were convertible to percentages since the total numbers of boys and girls were known, and hence to effect sizes, we first had to identify the measures corresponding to defined socialization areas (SA). These are given in their numerical codes in the left-hand column. The first one (50) represents our category of "Restrictiveness," which fits the study's variable "concern with cleanliness," the second (61) reflects "Nonphysical disciplinary strictness" and for this we had to average the four individual measures bracketed in Table 6.1, which we decided corresponded to the construct. The third (62) was "Physical punishment," labeled as such in the Table. The fourth (100) clearly corresponded to our category "Clarity of communication/reasoning,"

and finally we interpreted "Frequency of Mother's playing with S just for her own pleasure" as "Warmth" (30). To capture these last two categories we utilized only the numbers for "High use" of a practice. The numbers that entered into the meta-analysis are circled in the table.

Note also that the main purpose of the investigation was to relate these techniques/attitudes to the children's resistance to temptation and therefore the numbers are broken down by high and low resistance to temptation. The *Significance* column on the right refers to this contrast. Since we were not interested in this aspect, the first step in the extraction process was to add the numbers for high and low resistance together for boys and girls separately. These are shown in Table 6.2 in the columns "ActB" (Actual boys) and "ActG" (Actual girls). Each line in this worksheet, entered into a microcomputer spreadsheet program, represents one socialization area and shows the statistics that went into the calculation of the chi-square, also displayed in the table. (Many spreadsheets were much more complex still.) The chi-square for each socialization area would at a later stage be converted to an effect size and entered into a meta-analysis of that socialization area. These steps were accomplished by means of a microcomputer program, adapted from Mullen and Rosenthal (1985).

A few studies—not many—reported statistics for subgroups, such as different age groups or social class groups, and many reported statistics for mothers and fathers separately. For any socialization area we computed effect sizes separately for each group, but then also calculated an overall effect size for the total sample.

We carried out various adjustments suggested by meta-analytic statisticians, for example adjusting the error sum of squares in a multiway analysis of variance to derive a univariate F for the sex factor, or adjusting t for unequal n's. It should be realized that the differences in the outcomes produced by such adjustments are typically very small and certainly much smaller than those potentially generated by computational and transcription errors.

I noted earlier that many studies contained several measures of one

Table 6.2. Spreadsheet Calculations Preceding Computation of Effect Sizes

SA	*C*	*S*	*A*	*E*	*B*	*P*	*ActB*	*PosB*	*ActG*	*PosG*	*p*	*q*	*p1*	*p2*	χ^2	*dir*	*N*
30	1	2	1	1	0	1	22	40	25	37	0.610	0.390	0.550	0.676	1.277	−1	77
50	1	2	1	1	0	1	25	40	28	37	0.688	0.312	0.625	0.757	1.555	−1	77
61	1	2	1	1	0	1	26	40	25	37	0.662	0.338	0.650	0.676	0.057	−1	77
62	1	2	1	1	0	1	14	40	7	37	0.273	0.727	0.350	0.189	2.506	+1	77
100	1	2	1	1	0	1	20	40	14	37	0.442	0.558	0.500	0.378	1.153	+1	77

Abbreviations: SA: Socialization area (see text); C: Country (North America); S: Social Class (middle); A: Age (0–5 years); E: Ethnicity (Caucasian); B: Birth Order (not given); P: Parent Sex (Mother); ActB: Actual number of boys; PosB: Possible (total) number of boys; ActG: Actual number of girls; PosG: Possible (total) number of girls; p: Proportion of boys and girls; $q = 1 - p$; p1: proportion of boys; p2: proportion of girls; χ^2:chi-square; dir: direction of effect (1 = boys > girls; −1 = girls > boys); N: number of boys and girls, weight of study.

socialization area. These typically were averaged to form a composite effect size for the socialization area. (In the example in Table 6.1 the *n* 's for "Disciplinary strictness" could simply be added.) One problem that arose in this connection and that, to my knowledge, is not mentioned in the meta-analytic literature, is the fact that quite often a study provided statistics for one measure of a given socialization area (even if it was only $p < .05$) and proceeded to state for quite a number of other measures of the same socialization area that they showed no significant difference for sex, without giving any further statistic. If one simply ignores the nonsignificant measures and calculates an effect size from the (significant) statistic provided, one gives an inflated impression of the real sex difference found in this study for this socialization area. If, on the other hand, one interprets the insufficient information as indicating precisely "no sex difference" and, therefore, counts in a p of .5 for each of these measures in the pooled effect size for this socialization area, one unduly minimizes the sex difference found. When the *direction* of the difference was reported in such cases, we employed the compromise procedure of taking the average between the significant p reported (typically, .05) and a p of .5, that is, a p of .275, for each such measure, and then including these p's in the averaged effect size for the given construct. When "no significant difference" was reported for an independent socialization area as a whole, we used the procedure of assuming a p of .5, that is, a difference of exactly zero, in a secondary analysis, as others have done (cf. Eagly & Crowley, 1986).

As with the coding of socialization areas, we first adopted a fine-grained approach also to the coding of study characteristics, knowing that finer distinctions can later be collapsed. Thus we coded every country in which a study was done separately and coded four age groups and four ethnicity groups. Because of the immense task involved in analyzing about 375 comparisons from 172 studies, we later collapsed these categories to three groups each before the analysis proper started and omitted ethnicity from the analysis.

When one examines a series of studies carefully, as the meta-analyst has to do in order to code the studies and extract appropriate effect sizes, one inevitably uncovers a number of discrepancies, confusions and plain errors in the studies. Such errors, we found, occurred in articles in highly reputable journals, as well as in dissertations. Here are some of them.

Duplicate publication of the same data is something the meta-analyst has be on guard against. However, one detects such duplication easily in sorting studies by author. In one such case two separate papers addressed different aspects of the same investigation, but both papers contained the same data for differential treatment of the two sexes—but with differing *df*!

We sometimes had to reconcile the numbers of subjects given in the description of the sample with the numbers listed in relevant Tables. They

may have differed because an analysis was carried out only with a subset of the total sample, but occasionally there was no such obvious reason for the discrepancy. We chose the numbers in the Tables as a basis for our analysis.

More serious were discrepant statistics (means and variances) for the same data in the same study. This happened either because the data in one table had been regrouped and were recoded, or presumably through simple carelessness. In such a case we had to make a choice between the Tables on some rational basis. And lastly, at times when we recalculated some statistic (e.g., *F*), with the help of means and standard deviations also reported, we discovered that the original statistic was incorrect and had to be corrected.

Calculating meta-analytic statistics

The final outcome of a meta-analysis is the overall effect size, averaged across all studies. In our case it was the overall effect size across all studies of each construct, or socialization area, and we calculated these separately for studies done in North America and in other Western countries (including not only Western Europe, but also Australia).

We also computed separate effect sizes for mothers, fathers, and for parents combined. The last category included studies which examined mothers or fathers alone, as well as studies that reported on both parents, either separately or undifferentiated by sex. Thus there were a number of effect sizes and the one for "Parents Combined" represents the most inclusive effect size in a given socialization area.

The meta-analyst faces the question of whether to report the plain average effect size or whether to weight studies in some way. Each study could, for instance, be weighted by the confidence the researcher has in its internal validity or general trustworthiness, that is, by his assessment of its quality, or it could be weighted by a weight that reflects the reliability of each study's measures. But the most commonly used candidate as a weighting index is sample size.

Because the sample sizes of the studies summarized by our analysis varied considerably—from 11 to 2707—it seemed to us most important that this factor should be reflected in a given study's contribution to the overall average outcome. Some authors use the inverse of the variance of effect sizes, which itself is heavily influenced by sample size, as the weighting factor. However, we followed Hunter et al. (1982) in choosing the straightforward sample size, that is, number of boys plus number of girls, as the appropriate weight. The frequency-weighted average provides a more accurate reflection of the overall effect across all studies than does the unweighted average, so long as sample size is not correlated with the magnitude of the effect size (cf. Hunter et al., 1982)—in our case, it was not. This practice, however, gives considerable weight to very large studies and we therefore carried out

secondary analyses, omitting studies with sample sizes larger than 800, to see if their exclusion changed the overall conclusion. (It did in some cases.)

One problem specific to a meta-analysis of parents' socialization practices was that often both mothers' and fathers' practices toward boys and girls were reported. This meant a doubling of responses for each child in the study and was reflected in the degrees of freedom in the analysis tables. The weight applied to the study effect size nevertheless normally was the number of boys plus the number of girls.

For combining effect sizes from many studies and calculating an overall average frequency-weighted effect size we adapted a computer program (Program K), listed in Mullen and Rosenthal (1985), for use on the Macintosh microcomputer. We followed Hunter et al. (1982) in partitioning the variance into its systematic and unsystematic parts. Hunter et al. (1982) have pointed out that the variance of effect sizes is composed of systematic variation in population effect sizes and variation in sample effect sizes brought about by sampling error. Sampling error across studies behaves like measurement error across persons (cf. Hunter et al., 1982, p. 102). If there is systematic heterogeneity in effect size among studies, it is useful to examine the studies to see whether the variation is accounted for by differences due to "moderator variables," say, age of subjects, or quality of study. However, if most of the actual variance is due to unsystematic sampling error variance, it is not justifiable to search for systematic causes of variation. Hence one first determines how much of the actual variance is accounted for by sampling error variance. We did so by means of this program and found that sampling error in no case exceeded 60% of the actual variance—hence, following Hunter et al.'s (1982) and Peters, Hartke, and Pohlmann's (1985) guidelines, a search for moderator variables was justified.

We first correlated the moderator variables with effect sizes across studies, but soon realized that in a meta-analysis such as ours, in which there are many negative effect sizes, these correlations are quite misleading. We had established the convention that positive effect size would indicate a difference in favor of boys (greater amount of X for boys), and negative effect size a difference in favor of girls. Therefore, a negative correlation, say with age, could mean either (a) that with increasing age the positive effect size decreases, or (b) that with age the negative effect size increases, or (c) that with age the difference changes in direction from being in favor of boys to being in favor of girls. Hence, possible moderator variables were analyzed in relation to effect sizes by defining contrasts between groups for example age groups, to test for linear or quadratic trends—an analogue of analysis of variance, designed for meta-analytic investigations (cf. Hedges & Olkin, 1985; Rosenthal, 1984).

It will be clear by now that the whole process of extracting effect sizes and calculating the various meta-analytic statistics is subject to numerous errors, quite apart from the conscious assumptions and decisions that one makes in a

necessarily somewhat arbitrary way. Threats to accuracy arise from a variety of sources: errors of transcription, errors in calculating effect sizes (though setting up computer programs for various algorithms for computing these reduces mistakes to a minimum), errors in selecting the data from the study and in determining the correct algorithm to apply, errors in establishing the direction of effects and errors in combining subcomponents. It helps if the meta-analyst is a little obsessional, as constant checking of work to guard against these slips is essential. Orwin and Cordray (1985) have pointed out that primary investigators vary considerably in the completeness, accuracy, and clarity with which they report outcomes, as well as study characteristics, such as reliabilities, numbers, and ages of subjects, and that such reporting deficiencies will lead to uncertainty and error in the meta-analysis. It is all the more important to be vigilant about and eliminate errors in the meta-analytic calculations, as we otherwise superimpose these mistakes on the unknown number of errors that will exist in the primary literature.

The process of meta-analysis by its nature is very time-consuming. What takes time, above all, is reading the study, determining the applicable comparisons and outcome measures and deciding on an appropriate method for calculating effect size. Dissertations are more time-demanding than journal articles in this respect. The calculation of effect sizes is an equally protracted task, especially if various subcomponents of a construct have to be averaged, and if means or analysis of variance tables have to be recast in order to extract the requisite standard deviations. The time factor will increase dramatically if, because of errors detected later, various analyses have to be repeated. Including corrections and repeat analyses, the calculation of effect sizes and of summary meta-analytic statistics for our study took over a year, part of this full time.

Salient Findings

Our conceptual model posited that sex-related differential treatment of boys and girls will vary for different areas of socialization. In fact, the only socialization area in which a significant effect size emerged in North American studies was encouragement of sex-typed activities—see Table 6.3. The question asked here was: To what extent do parents encourage conventionally sex-typed play or other activities in their children? For example, how much more do they encourage boys than girls, say, to play with hammers and expect them to shovel the sidewalk, and how much more do they encourage girls than boys, say, to play with dolls and expect them to do housework? The effect size here tells us the amount of the difference parents make between boys and girls in this area, each time in the expected direction.

Although the effect sizes for mothers, fathers, and parents combined are significant in this area their magnitudes range from about one-third to

Table 6.3. Significant Effect Sizes: Parents Combined (North American Studies)

Socialization area	*Studies with known effect sizes*	*Studies of $N < 800$*	Studies with $p = .5$ added
Encourage sex-typed activities	.43*(20)	.58*(19)	.37*(23)
Studies from other Western Countries			
Physical punishment	.37*(4)		.24*(5)
Encourage achievement	−.28*(4)	.15(3)	−.19(6)

Figures in parentheses are number of studies included.
*$p < .05$

one-half of a standard deviation, that is they are rather modest. Because of its inordinate weight, we eliminated one study with a sample size of over 1, 800 in a further analysis. When this was done, the effect was still significant for parents combined (Column 3 of Table 6.3), but no longer for mothers alone. Further, when we added in the studies where we assumed a p of .5 (i.e., a zero difference) in the absence of further information, the effect size still remained significant (Column 4). Thus this effect was quite robust.

We had formulated hypotheses, based on theory and literature regarding the effects of certain moderator variables: (a) differential treatment of boys and girls (i.e., the effect size) would, in general, increase with age; (b) effect sizes would be larger when data were obtained from observation or experiment rather than from parent interview; (c) effect sizes would be larger in published than in unpublished studies; and (d) they would be larger in older than in more recently published studies. The outcomes for these hypotheses for the socialization area "Encouragement of sex-typed activities" were as follows: Only the second hypothesis received some slight support, in that the effect size was larger for data obtained from observation or experiment, but the contrast was nonsignificant. The first hypothesis was refuted: the effect size was more than twice as large at the preschool age (.71) than at the primary school age (.29), though this must be considered with reservations, as the number of studies that included this socialization area for the primary school age was only 2. The third hypothesis was not supported—rather there was a tendency in the opposite direction: unpublished studies had a (nonsignificantly) higher effect size. The last hypothesis also received no support: Studies published since 1973 had a (nonsignificantly) larger effect size than studies published in 1973 or earlier, that is, the studies that were available to Maccoby and Jacklin (1974). Higher quality studies (we had not formed any hypothesis about quality) also produced somewhat higher effect sizes than did lower quality studies. The outcomes for socialization areas overall were very similar. However, it should be noted that no significant effect size emerged for any subgroup (e.g., of high quality studies) where there was no significance for the sample as a whole.

As Table 6.3 shows, in Western countries (other than North America) there

was a significant difference for physical punishment which parents inflicted more on boys than on girls. The result should be treated with some reservation, as the number of studies involved was small, though the significance held up when studies for which we assumed a p of .5 were included.

The finding that in Western countries parents encouraged achievement more in girls than in boys must be regarded as rather unreliable. It is based on a very small number of studies and the difference in the female direction is due to one study alone which contributed 877 subjects to the weighted effect size and was in fact the only study of the four that showed a difference in the female direction. Once this anomalous study is removed, the effect size points in the direction of greater encouragement for boys, but it is nonsignificant (the same result as was found in North American studies). The data of this study—a dissertation—represented answers to a questionnaire, administered to high school students in Great Britain, as well as in the United States (in which the difference, though not so strong, was also in the female direction).

We also correlated some of the moderator study characteristics with each other (for all socialization areas). Some interesting findings emerged: The number of female authors increased with recency of publication, female authors tended to produce studies of higher quality than male authors, and the quality of studies increased in more recent years, but quality decreased with larger sample size. This last finding reflects the likelihood that carefulness of data collection slips, and methods become coarser, when a researcher is faced with a large sample. (For further details see Lytton and Romney, 1991.)

What are the main conclusions we can draw from the significant effect for differential encouragement of sex-typed play and activities? The findings corroborate Maccoby and Jacklin's (1974) conclusions that this is an area in which parents do make differences between boys and girls. It must be borne in mind, however, that such encouragement may build on the children's already existing preferences, that is, it may be the children's genotype which evokes differential responses (the "evocative genotype—environment effect: cf. Scarr & McCartney, 1983). There is, indeed, some evidence that supports such an interpretation: Two investigations (Caldera, Huston, & O'Brien, 1989; Snow, Jacklin, & Maccoby, 1983) discovered that 12-to 18-month-old boys and girls already differ in preferences for toys conventionally associated with their sex, when the opportunity for playing with them is equalized. Parents therefore may simply reinforce these preferences rather than create them.

For other socialization areas no evidence for significantly differential treatment of the sexes emerged. Of course, one can never *prove* the null hypothesis: it is impossible to assert that no sex-differentiated treatment in these areas ever exists. We have not exhausted the universe of samples (a) of non-English-language studies, (b) of possible childrearing variables, (c) of future potential methods of measurements, and the studies summarized have not exhausted the universe of possible subjects.

However, given our relatively exhaustive search of the existing English-language literature within the stated periods, and within the limitations of this literature, we can, I believe, tentatively accept the null hypothesis of no significant sex-differentiated treatment in the stated socialization areas (cf. Greenwald, 1975; for further discussion, see later). This confirms Maccoby and Jacklin's general conclusion that reinforcement is similar for the two sexes. It suggests that the observed sex differences in cognitive or social characteristics, for example in high mathematical or spatial ability, in non-verbal behavior or in aggressiveness—the magnitude of which, in any case, seems to have diminished in more recent studies—are *not* due to parents directly reinforcing stereotypical behavior or abilities in their children.

What other explanations remain for the origins of existing sex differences? First, there may, of course, be some subtle forms of parental treatment that were not captured by our categories and which might conceivably show some differences by sex. Children may also be influenced indirectly by parental encouragement of differential types of play in boys versus girls, if "masculine" play promotes the development of "masculine" attitudes and abilities and vice versa for "feminine" play. Second, children will form cognitive schemas of sex roles and sex membership and on this basis will see same-sex parents, peers and others as models on whom they may pattern their behavior. Third, peers are likely to act as powerful socializers of sex-typed behavior in general, though they, too, like parents, may only reinforce existing tendencies. Fourth, teachers may influence children in stereotypical ways, but the evidence for their deliberately influencing children's cognitive abilities in this manner is rather weak. In high school, for instance, girls get better *grades* in algebra, though they do more poorly in standardized tests, than boys—it does not look as if teachers discouraged girls in this area (cf. Grusec & Lytton, 1988; Kimball, 1989).

Last, and above all, we can no longer close our eyes to the possibility of biological predispositions playing a part. These would seem to provide a plausible part-explanation for existing sex differences, supported by evidence for some behaviors, such as, aggression (cf. Maccoby & Jacklin, 1980) or high mathematical ability (cf. Benbow, 1988). Such biological factors are coming to be recognized more and more in areas such as psychopathology. It must be emphasized that these influences will always interact with, and be amplified by, parents' and other societal, environmental forces. Bidirectional processes are always likely to be operating.

INTERPRETATION, LIMITATIONS, AND CREDIBILITY OF A META-ANALYSIS

The message of a meta-analysis cannot be read off mechanically, like an electricity meter. Its meaning, applicability, and the qualifications that should

be attached to it need interpretation. Generalizability of the conclusions depends heavily on the criteria employed for sampling the literature and the extent to which the stated goal of sampling was achieved, and also on the variations of outcomes within and between studies. The practical applicability of a meta-analysis will depend on considerations inherent in the topic and in the theory and practice related to it. To arrive at a decision on these questions readers will require some information about the nature of the studies integrated, their methods of investigation and limitations. Some of this information can be provided in quantitative form, for example age of subjects, quality estimate of the study, or method of investigation, and such variables will then be used as moderator variables. Other information may be more impressionistic and will be combined with the quantitative information in some rational way in the account of the meta-analysis to provide a basis for interpreting the meta-analytic findings. The qualifications relating to finding of a significant effect in the female direction for encouragement of achievement in Western countries, which we discussed above, illustrate how awareness of the nature of constituent studies may affect the interpretation of the meta-analytic outcomes.

It is, of course, important for the meta-analyst to discuss the distinctions within the set of studies he is synthesizing and to explore the differences in effect size that variation in study characteristics produces. The burden of Slavin's (1984) critique of meta-analytic studies in education is that typically meta-analysts do not do this and he therefore deplores the present practice of meta-analysis in some areas. He argues that "meta-analyses could be used to compute effect sizes for well-defined subsets of studies felt by the reviewer (on firm conceptual . . . grounds . . .) to be truly comparable" (p. 12). His case against the oversimplistic interpretation of meta-analyses is a strong one.

Slavin (1984) criticizes Glass, Cahen, Smith, and Filby (1982) for claiming a "remarkably strong" relationship between class size and pupil achievement, and points out that this strong effect exists only if tutoring studies (one-on-one relationships) are included in the "smaller classes," but that if one considers studies that have examined reductions in normal class size, say, from 35 to 25, the beneficial effects vanish. While his criticism of the strong overall conclusion from the study is justified, he misses the point that Glass et al. themselves provided him with the ammunition, since they broke down their results by different class sizes and showed the minimal effect that reductions within the normal range has. This precisely is the advantage of a well-conducted meta-analysis.

A possible obstacle to the acceptance of a finding of essentially "no significant effects" for a set of studies is the presence of great variability among studies. If there is great variability in both the positive and negative direction among studies this will produce a low average, or nonsignificant effect size. But in this case the meta-analysis may simply reflect the state of

research in the area and clearly reveal the stark truth, namely that no reliable results in any direction are available.

However, some of the variability may be explainable by systematic variations among studies. If, after removing the random sampling error (as recommended by Hunter et al., 1982), substantial variation in study outcomes subsists one can explore this variability by partitioning the set of studies into more homogeneous subsets, by means of moderator variables, such as age of subjects, and so on, which may consistently show effect sizes in one direction or the other. If these subgroups still do not yield reliable effects, then the conclusion of "no significant effects" will become more firmly grounded.

A possible source of distortion is the appearance of deviant samples, particularly within a small set of studies. The large-sample study that showed a greater encouragement of achievement for girls and that tipped the scales in this direction for the analysis of Western countries (discussed earlier) is a case in point. Obviously, frequency weighting, although representing more accurately the contribution of component studies to the overall average, has its dangers. This was the reason why we carried out subsidiary analyses, excluding studies with the largest samples.

When results emerge that are counterintuitive and against theoretical expectations the first step is always to check one's calculations and transcriptions, in view of the everpresent possibility of human error. Thus, when on first inspection we once noticed an effect size that was of a different order of magnitude from our usual effect size estimates, we discovered that at some stage one of its components had been multiplied by 10, something that is easily done when one is dealing with quantities expressed only in decimals.

Although meta-analysis was first thought of as a means of resolving discrepancies between narrative reviews with their intuitive interpretations, it has engendered controversies and conflicts of its own. To replicate a meta-analysis is, no doubt, a salutary exercise even when a first meta-analysis has been carried out sensitively and meticulously, in view of all the subjective decisions involved in each study. A case in point is Matt's (1989) (and others') replication of part of Smith et al.'s (1980) meta-analysis of studies of psychotherapy which I have already discussed.

It must be obvious that there are well and poorly executed meta-analyses, just as there are good and poor primary studies—another reason why meta-analyses need to be replicated, too. Conflicts and variation are not uncommon among such quantitative reviews, and hence it is important to understand, explain, and qualify the differing outcomes (cf. Nurius & Yeaton, 1987). Aside from differing assumptions, carelessness may be the most potent cause of discrepancies between meta-analyses.

Take the case of the review of three meta-analyses on the relation between students' ratings of course/instructor and their achievement, reported by

Abrami, Cohen, and D'Apollonia (1988). They comment on wide disagreements on the individual correlation between these variables, the "validity coefficient," extracted from a given primary study. They give as an example the fact that reviewer A cites a correlation of .90 between grade and instructor rating, whereas reviewer B cites a correlation of .136 for the same relation from the same primary study. As I was intrigued by this startling discrepancy, which Abrami et al. (1988) do not explain, I read the primary study involved and the pertinent meta-analyses. I found that Reviewer B's figure of .136 was the correct correlation and he had also cited all other relevant correlations correctly from the quoted source. Although Abrami et al. attribute a cited correlation of .90 to Reviewer A, I could not find a correlation from that primary study cited in any of Reviewer A's published papers on the subject, though his doctoral dissertation—unavailable to me—may have contained one.

Why would Reviewer A attribute a fictitious correlation of .90 to the primary source, if indeed he did so? Such an error is particularly surprising in this case, as the correlation simply had to be copied from the primary source; it would be more understandable if effect sizes had to be calculated from a variety of complex statistics for example from F-tests between different sets of groups (not the contrast groups of interest). The fact is that the three quoted reviews differed sharply on the overall validity of student ratings, with Reviewer A concluding that there was strong support for their validity, as evidenced by the relationship between ratings and student performance, and Reviewer B concluding the opposite. Perhaps Reviewer A's misreading of the correlation contributed to his enthusiasm for the validity of student ratings, but he may also have misread it because it thus could confirm his enthusiasm—the careless error, at any rate, seems to have been in the direction of his predilections.

The moral of the story is: Careless authors produce misleading meta-analyses, and if meta-analyses of meta-analyses perpetuate these errors, and cite them without further inquiry into the matter, they may well conclude that disagreements on the actual values of study outcomes are large: Abrami et al. (1988) report there is only 47% agreement among reviewers on effect sizes. At least such meta-meta-analyses alert us to the possibility of discrepancies, whatever they are due to. Perhaps editors have a part to play here and should demand a demonstration of reliability for the calculation of effect sizes before publication, or arrange for some spot-checking? It may well be that simple clerical or calculation errors are responsible for many, and perhaps most, discrepancies in findings in primary studies and in meta-analyses.

The interesting fact is that meta-analyses can be replicated more exactly than primary studies can, since the meta-analysis' subjects, the studies, are available in their original state for reanalysis, whereas the subjects of a

primary study never are. Further, its assumptions and methods, having been made explicit, can also be replicated precisely, or with stated modifications, though many detailed decisions may still remain hidden.

One of the chief criticism leveled at meta-analysis has been that an average effect size becomes meaningless if it is derived from a heterogeneous sample of studies that differ in populations, settings, methods, or measures. (The studies must address what at least the reviewer can conceptualize as a single *topic*, and I discussed earlier the problem of defining what is or is not the "same" topic.) However, as Nurius and Yeaton (1987) point out, this diversity in *populations* and so on, is also one of the benefits of meta-analysis, if properly used. The reviewer can then attempt to assess the difference in outcome that these differences make, by using these factors (populations, etc.) as stratifying, or moderator variables. If quality varies, the reviewer can take this into account by assigning an appropriate rating.

It is true, the reviewer is restricted by the knowledge base we have, that is, by the topics and constructs that primary researchers have chosen to study. For instance, in our study of socialization practices we were interested in determining whether a sex-related difference emerged for parents' use of Baumrind's "authoritative" or "authoritarian" parenting styles. However, as only one or two studies assessed sex differences for these complex configurations of practices, the number was too small to base an effect size on, though the component practices did appear in a number of investigations and could be analyzed separately. The limitations of the existing literature are, of course, also ever present with narrative reviews, but meta-analyses make the problem more visible.

It is also true that the reviewer cannot remedy the flaws and limitations of the primary literature, and if all the primary studies are weak, the meta-analysis will be based on sand. If this is, indeed, the case, and provided the meta-analysis has sampled the literature appropriately, the only conclusion one can draw is that one does not, and cannot, know anything scientifically about the given topic, since there is no other research to fall back upon.

Do we believe the findings of a meta-analysis, in particular one on sex differences? Or do we hold Virginia Woolf's view: "At any rate, when a subject is highly controversial—and any question about sex is that—one cannot hope to tell the truth. One can only show how one came to hold the opinion one does hold." (*A Room of One's Own*)? Social scientists in the 1980s, to the contrary, would argue that "we now have tools of analysis that allow us to approach the truth about sex differences much more closely than scholars have approached it in the past" (Eagly, 1987, p. 759). As Virginia Woolf rightly states, preconceptions are particularly strong in the area of sex differences. But meta-analytic methods bring some objectivity to the reviewing process. And, while subjective assumptions and decisions are inescapable, as I have repeatedly pointed out, these are not clearly related to, or influenced by, the

reviewer's predilections and the impact of such preconceptions is at least considerably reduced by the open accountability and the statistical inference rules of the meta-analysis. Provided the analysis has been carefully done and the conclusions are based on the evidence, with appropriate qualifications arising from the study characteristics, these conclusions represent the *best evidence we have on the topic at this time.* In order to assert the contrary, critics would have to show that another meta-analysis of the same, or perhaps a broader, set of studies produces a contrary conclusion and explain why the first meta-analysis was mistaken. Controversies may, however, still legitimately arise about the broader implications of the findings of any meta-analysis.

REFERENCES

Abrami, P. C., Cohen, P. A., & d'Apollonia, S. (1988). Implementation problems in meta-analysis. *Review of Educational Research, 58,* 151–179.

Baldwin, A. L., Kalhorn, J., & Breese, F. H. (1949). The appraisal of parent behavior. *Psychological Monographs, 63* (4, Whole No. 299), 1–85.

Bangert-Drowns, R. J. (1986). Review of developments in meta-analytic method. *Psychological Bulletin, 99,* 388–399.

Baumrind, D. (1971). Current patterns of parental authority. *Developmental Psychology Monographs, 4* (No. 1, Pt. 2), 1–103.

Baumrind, D. (1983). Rejoinder to Lewis' reinterpretation of parental firm control effects: Are authoritative families really harmonious? *Psychological Bulletin, 94,* 132–142.

Bell, S. M., & Ainsworth, M. D. S. (1972). Infant crying and maternal responsiveness. *Child Development, 43,* 1171–1190.

Benbow, C. P. (1988). Sex differences in mathematical reasoning ability in intellectually talented preadolescents: Their nature, effects, and possible causes. *Behavioral and Brain Sciences, 11,* 169–183.

Block, J. H. (1976). Issues, problems, and pitfalls in assessing sex differences: A critical review of "The psychology of sex differences." *Merrill-Palmer Quarterly, 22,* 283–308.

Block. J. H. (1978). Another look at sex differentiation in the socialization behaviors of mothers and fathers. In J. A. Sherman & F. L. Denmark (Eds.), *The psychology of women: Future directions in research* (pp. 29–87). New York Psychological Dimensions.

Bradley, R. H., Caldwell, B. M., Rock, S. L., Ramey, C. T., Barnard, K. E., Gray, C., Hammond, M. A., Mitchell, S., Gottfried, A. W., Siegel, L., & Johnson, D. L. (1989). Home environment and cognitive development in the first 3 years of life: A collaborative study involving 6 sites and 3 ethnic groups in North America. *Developmental Psychology, 25,* 217–235.

Burton, R. V., Maccoby, E. E., & Allinsmith, W. (1961). Antecedents to resistance to temptation in four-year-old children. *Child Development, 32,* 689–710.

Caldera, Y. M., Huston, A. C., & O'Brien, M. (1989). Social interactions and play patterns of parents and toddlers with feminine, masculine and neutral toys. *Child Development, 60,* 70–76.

Campbell, D. T. (1986). Science's social system of validity-enhancing collective belief change. In D. W. Fiske & R. A. Shweder (Eds.), *Metatheory in social science* (pp. 108–135) Chicago: University of Chicago Press.

Clarke-Stewart, K. A., VanderStoep, L. P., & Killian, G. A. (1979). Analysis and replication of mother–child relations at two years of age. *Child Development, 50,* 777–793.

Dawes, R. B., Faust, D., & Meehl, P. E. (1989). Clinical versus actuarial judgment. *Science, 243*, 1668–1674.

Eagly, A. H. (1987). On taking research findings seriously. *Contemporary Psychology, 32*, 759–760.

Eagly, A. H., & Carli, L. L. (1981). Sex of researchers amd sex-typed communications as determinants of sex differences in influenceability: A meta-analysis of social influence studies. *Psychological Bulletin, 90*, 1–20.

Eagly, A. H., & Crowley, M. (1986). Gender and helping behavior: A meta-analytic review of the social psychological literature. *Psychological Bulletin, 100*, 283–308.

Epstein, S. (1980). The stability of behavior: II. Implications for psychological research. *American Psychologist, 35*, 790–806.

Fiske, D. W. (1986) Specificity of method and knowledge. In D. W. Fiske & R. A. Shweder (Eds.), *Metatheory in social science* (pp. 61–82). Chicago: University of Chicago Press.

Glass, G. V. (1977). Integrating findings: The meta-analysis of research. *Review of Research in Education, 5*, 351–379.

Glass, G. V., Cahen, L., Smith, M. L., & Filby, N. (1982). *School class size*. Beverly Hills, CA: Sage.

Greenwald, A. (1975). Consequences of prejudice against the null hypothesis. *Psychological Bulletin, 82*, 1–20.

Grusec, J. E., & Lytton, H. (1988). *Social development: History, theory and research*. New York: Springer-Verlag.

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. Orlando, FL: Academic Press.

Hoffman, M. L. (1970). Moral development. In P. H. Mussen (Ed.), *Carmichael's manual of child psychology* (Vol. 2, pp. 261–359). New York: Wiley.

Holden, G. W., & Edwards, L. A. (1989). Parental attitudes toward child rearing: Instruments, issues and implications. *Psychological Bulletin, 106*, 29–58.

Hubbard, F. O. A., & van IJzendoorn, M. H. (1987). Maternal unresponsiveness and infant crying: A critical replication of the Bell and Ainsworth study. In L. W. C. Tavecchio & M. H. van IJzendoorn (Ed.), *Attachment in social networks: Contributions to the Bowlby-Ainsworth attachment theory*. Amsterdam: North-Holland.

Hunter, J. E., Schmidt, F. L., & Jackson, G. B. (1982). *Meta-analysis: Cumulating research findings across studies*. Beverly Hills, CA: Sage Publications.

Kamin, L. J. (1978). Comment on Munsinger's review of adoption studies. *Psychological Bulletin, 85*, 194–201.

Kimball, M. M. (1989). A new perspective on women's math achievement. *Psychological Bulletin, 105*, 198–214.

Kuhn, T. S. (1962). *The structure of scientific revolutions*. Chicago: University of Chicago Press.

Lewis, C. C. (1981). The effects of parental firm control: A reinterpretation of findings. *Psychological Bulletin, 90*, 547–563.

Light, R. J., & Smith, P. V. (1971). Accumulating evidence: Procedures for resolving contradictions among different studies. *Harvard Educational Review, 41*, 429–471.

Lytton, H., & Romney, D. M. (1991). Parents' sex-differentiated socialization of boys and girls: A Meta-analysis. *Psychological Bulletin, 109*.

Maccoby, E. E., & Jacklin, C. N. (1974). *The psychology of sex differences*. Stanford: Stanford University Press.

Maccoby, E. E., & Jacklin, C. N. (1980). Sex differences in aggression: A rejoinder and a reprise. *Child Development, 51*, 964–980.

Maccoby, E. E., & Martin, J. A. (1983). Socialization in the context of the family: Parent-child interaction. In E. M. Hetherington (Ed.), *Handbook of child psychology: Vol. 4. Socialization, personality and social development* (pp. 1–102). New York: Wiley.

Matt, G. (1989). Decision rules for selecting effect sizes in meta-analysis: A review and reanalysis of psychotherapy outcome studies. *Psychological Bulletin, 105*, 106–115.

Mosteller, F., & Bush, R. R. (1954). Selected quantitative techniques. In G. Lindzey (Ed.), *Handbook of Social Psychology: Vol. 1. Theory and method* (pp. 289–334). Cambridge, MA: Addison-Wesley.

Mullen, B., & Rosenthal, R. (1985). *BASIC Meta-analysis: Procedures and programs.* Hillsdale, NJ: Erlbaum.

Munsinger, H. (1974). The adopted child's IQ: A critical review. *Psychological Bulletin, 82,* 623–659.

Munsinger, H. (1978). Reply to Kamin. *Psychological Bulletin, 85,* 202–206.

Nurius, P. S., & Yeaton, W. H. (1987). Research synthesis reviews: An illustrated critique of "hidden" judgments, choices and compromises. *Clinical Psychology Review, 7,* 695–714.

Orwin, R. G., & Cordray, D. S. (1985). Effects of deficient reporting on meta-analysis: A conceptual framework and reanalysis. *Psychological Bulletin, 97,* 134–147.

Peters, L. H., Hartke, D. D., & Pohlmann, J. T. (1985). Fiedler's contingency theory of leadership: An application of the meta-analysis procedures of Schmidt and Hunter. *Psychological Bulletin, 97,* 274–285.

Rosenthal, R. (1984). *Meta-analytic procedures for social and behavioral research.* Beverly Hills, CA: Sage.

Rothbart, M. K. (1971). Birth-order and mother-child interaction in an achievement situation. *Journal of Personality and Social Psychology, 17,* 113–120.

Scarr, S., & McCartney, K. (1983). How people make their own environments: A theory of genotype-environment effects. *Child Development, 54,* 424–435.

Sears, R. R., Maccoby, E. E., & Levin, H. (1957). *Patterns of child rearing.* Evanston, IL:Row, Peterson.

Slavin, R. E. (1984). Meta-analysis in education: How has it been used? *Educational Researcher, 13,* 6–15.

Smith, M. L., Glass, G. V., & Miller, T. I. (1980). *The benefits of psychotherapy.* Baltimore: John Hopkins University Press.

Snow, M. E., Jacklin, C. N., & Maccoby, E. E. (1983). Sex-of-child differences in father-child interaction at one year of age. *Child Development, 54,* 227–232.

Yarrow, M. R., Campbell, J. D., & Burton, R. V. (1968). *Child rearing.* San Francisco: Jossey-Bass.

CHAPTER 7

From Virtual to Systematic Replications*

Jurg Siegfried
Department of Psychology
University of North Carolina at Chapel Hill

The demand for replications in social sciences is widespread (American Psychological Association, 1983; Bahr, Caplow, & Chadwick, 1983; Bulldock & Svyantek, 1985; Campbell & Stanley, 1963; Duncan, 1969; Finifter, 1972; Gasparikova-Krasnec & Ging, 1987; La Sorte, 1972; Smith, 1983; Stouffer, Suchman, DeVinney, Star, & Williams, 1949; White & Pesner, 1983). Many of the authors call replication the core procedure of science. Science is proper, if and only if findings are replicated. At the same time it is asserted that replications in social sciences are very rare (Bahr et al., 1983; Denzin, 1970; Rodrigues, 1982; Smith, 1970, 1983). Why is it that social scientists, knowing that replication is crucial for scientific standards, do not replicate each others' studies more often? Is it true that replications in research are very rare and, if so, what are possible reasons for this? It is these questions that I address in this chapter. First, I outline some differences of replications in natural and social sciences and draw on an example of natural sciences together with examples from psychotherapy research. A brief classification of replication procedures will lead to a 5-year review of replications done in psychological research. In a last section, an attempt is made to outline different procedures of enhancing confidence in research findings in psychology, provided that replications are difficult to perform.

My interest in replications goes back to a huge project of meta-analysis of

*Preparation of this chapter was financed by the Swiss National Science Foundation, Fellowship nr. 83.454.0.87.

I thank my friend Jaan Valsiner, Department of Psychology at University of Chapel Hill, North Carolina, for his constant help, suggestions, and critique of the manuscript.

psychotherapy research in which I was involved (Siegfried, 1985b). In this project the validity of an empirical study in its internal and external aspect was of some importance. Many authors claimed as well that they were replicating an earlier study. Often it was not easy to decide what criteria such authors had in mind for making such statements. It seems that the term *replication* roughly is used for *repeating a study with different subjects and similar procedures*. However, whether the repeated study is indeed a valid second "trial" of the first one is often doubtful. Often procedures are changed quite drastically and still the claim is made that the study is a replication of an earlier one. We began to investigate into the *validity of replications*, thereby expressing the fact that a repeated study may be a valid or not valid experimental trial in an assumed series of trials governed by the same hypothesis and conducted with the same experimental apparatus. It is the validity of replications that mainly distinguishes natural and social sciences.

REPLICATIONS IN NATURAL AND SOCIAL SCIENCES

Many critics of my work disapprove of the position that data of natural and social sciences are totally different. They claim that the scientific study of psychology, for example, is no different from scientific investigations in any of the natural sciences. Some admit that data are weaker and less certain in social sciences. But cumulative evidence will, they maintain, lead to strenghten a number of hypotheses to the point that they safely can be regarded as laws of social sciences. However, this position is not valid if a careful study of the validity of replications in natural and social sciences is performed. I will show this with two examples.

Robert Boyle performed a series of replications of an experiment in 1662 which nicely serve to show the *logic of replications* in natural science experiments. Eager to study the compressing force of the air trapped in a J-shaped tube of glass, he conducted a series of experiments by adding mercury to the longer leg of the tube. In this way he could study how much mercury had to be added in order to reduce the volume of the trapped air inside the closed end of the tube. He realized that the amount of mercury added did not correlate in a linear way with the amount of reduction of the volume of the air. Rather, he found that "the greater the weight is that leans upon the air, the more forcible is its endeavour of dilation and consequently its power of resistance" (cited in Harré, 1981, p. 86).

The number of repeated experiments—in each trial the number of equal spaces at the top of the tube that contained the same parcel of air was varied systematically—led him to the formulation of his hypothesis.

Table 7.1 shows the important features of such experiments:

1. Repeated observations under the same conditions with the same apparatus and the same procedure.

Table 7.1. Boyle's natural science experiment on the resistance of air

A	*B*	*C*	*D*	*E*
48	00		29 3/16	29 3/16
46	01 7/16	A	30 9/16	33 5/16
44	02 13/16	d	31 15/16	31 13/16
42	04 6/16	d	33 9/16	33 1/7
40	06 3/16	e	35 5/16	35
38	07 15/16	d	37	36 15/19
36	10 2/16		39 5/16	38 7/8
34	12 8/16	to	41 10/16	41 7/17
32	15 1/16		44 3/16	43 1/16
30	17 15/16	22	47 1/16	46 3/5
28	21 3/16	1/3	50 5/16	50
26	25 3/16		54 5/16	53 10/13
24	29 11/16	m	58 13/16	58 3/8
22	34 15/16	a	64 1/16	63 9/11
20	41 9/16	k	70 11/16	70
18	48 12/16	e	77 1/16	77 3/3
16	58 8/16	s	87 1/16	87 7/8
14	71 5/16		100 7/16	99 7/7

A: The number of equal spaces in the shorter leg, that contained the same parcel of air diversely extended.

B: The height of the mercurial cylinder in the longer leg, that compressed the air into those dimensions.

C: The height of the mercurial cylinder, that counterbalanced the pressure of the atmosphere.

D: The aggregate of the two last columns B and C, exhibiting the pressure sustained by the included air.

E: What that pressure should be according to the hypothesis, that supposes the pressures and expansions to be in reciprocal proportion.

2. Each trial has to be a valid trial. The added mercury, example, must be the systematically varied amount. If some of the mercury is spilled or the glass tube breaks then the trial is not valid and has to be repeated.

3. The observations and repetitions of the experiment are done in a relatively short time.

4. They still can be performed today, almost 300 years later with the same results. Replications of this experiment are done in numerous schools and universities around the world every day with the same outcome.

In social sciences, as far as they are experimental, trials in experiments are either complex interventions (in the case of a counseling or educational program), tasks to be solved by a subject (a repeated perception task), or the subjects themselves are taken as trials of the experiment (e.g., in a sample 50 subjects gave response X, 10 did not give response X). Klein, Zitrin, Woerner, and Ross (1983), for example, compared three different treatments in three groups of phobic clients. The first group received behavior therapy together with chemotherapy (impiramine). The second group was treated with

behavior therapy and placebo medication. The third group received supportive therapy and medication. No differences were found between the three groups of clients. No superiority of behavioral and supportive therapies was shown.

Clearly, in this study, a complex intervention in the form of a therapy was offered and the subjects served as trials within the different groups. Each of them either improved or did not improve. The individual subject's outcome is an individual trial for each form of therapy leading to the conclusion that therapy as a whole was effective or was not effective.

The important features of such experiments are:

1. Repeated individual trials (i.e., subjects) with 3 different groups of subjects under the same conditions with the same apparatus and the same procedure.
2. Each individual trial must not be a valid trial. Some subjects may have no motivation to participate and are, if motivation is not controlled for, invalid trials.
3. *The individual trials cannot be varied systematically. There is no way* to add some more motivation to some subjects and then see how their behavior is changing.
4. The individual trials cannot be repeated because the therapeutic process is not reversible. A subject is not the same after the treatment as before insofar as the treatment had some effects.
5. The individual repeated trials are of no interest in terms of the result of the study. Therapeutic effects are overall effects of the groups as a whole. Only a replication of such an experiment as a whole could be regarded as a second trial in the sense of the natural sciences.

The difference between the two sorts of experiments can hardly be overestimated, as Table 7.2 shows. Looking first at Table 7.2 in general, we observe that none of the characteristics of the replicated experiment listed in the table are a threat to validity in the natural sciences as long as conditions are equal. The only possible bias is with a change of the experimenter in the whole approach of an experimental set up, as Mitroff (1974a, 1974b) has shown some time ago. The experimenter, it is argued, is not an emotionally free observer of natural phenomena but constructs his observations. Therefore, an experimenter influence must be assumed in natural sciences as well. Generally, however, as long as experimental set ups are identical they will produce identical data (provided that researchers are wearing the same theoretical glasses).

Now consider replications of experiments in social sciences. All the factors listed in the table are possible threats to validity.

Table 7.2. Different experimental features' possible contribution to threats of validity in natural and social sciences

	possibility of threat to internal and/or external validity	
exp. features	*natural sciences*	*social sciences*
identical procedures	no	yes
identical apparatus	no	yes
ident. observations	no	yes
different experimenter	no	yes
different location with equal conditions	no	yes

Let me go through the entries in the table one by one. Consider first the identity of procedures. Taking part in a first study may train therapists to follow certain rules in the procedure. The training effect of the first study is a growth in experience of the therapists. This might have an effect on the outcome of the second study, as shown, for example, by Haccoun and Lavigueur (1979). Therefore, the procedure might, with good reason, not be the same a second time because of the changing experimenters, therapists, and subjects. This is equally the case in experimental psychology. Asking subjects a second time the same question is not the same as asking them the first time. Experience, familiarity with procedures, learning to conform to experimenter expectations are all vital parts in experiments in social sciences and may change outcomes, even if the experimenter considers these changes to be irrelevant.

Inspecting Table 7.2 again the identity of the apparatus used in social scientific experiments shows a similar result. In a replication of the famous Middletown study, Bahr et al. (1983) dealt extensively with the problem of updating questionnaires and interview items. The original study was done in the years 1924–1925 and some of the questions used at that time obviously could not be used anymore in 1977. Yet, to change items may lead to different outcomes and violates the principle of replication in the strict sense (Allerbeck & Hoag, 1984). In another recent replication of Furnham (in press) a questionnaire devised in 1945 was reapplied again in 1988. Results show a clear shift in beliefs in many instances. The statement *Most children are born good* in 1945 was answered by 68 of 141 students in the affirmative whereas in 1988 the same question given to 98 students was answered positively by only 11 students. These examples clearly show some of the problems of replications in social sciences. Problems with questionnaires do not only manifest themselves after such lenghty intervals between replication as in the two mentioned studies, but have to be considered equally for shorter periods.

Again, experimental psychology does not make an exception here. Even if the experimental setup is exactly the same in one and another experiment, subjects will differ which makes them not comparable to a trial in the Boyle experiments earlier. Human subjects have different life histories, represent different communities, times, and so on. All this makes *trials* (equal to a subject's response) in social sciences' experiments much more subject to the arbitrary (Smedslund, 1979) and to historical changes (Gergen, 1973; Sjöberg, 1982; Valsiner, 1985; White & Pesner, 1983) than in natural sciences.

Validity, as indicated in Table 7.2, is threatened again by repeated observations on subjects. Problems of measurement dependency, rater reliability over time (a measure rarely reported in scientific investigations), and irreversible developments or changes within subjects (see Chapter 8) are just a few examples. Such factors are crucial especially in AB designs with replication (Harris & Jenson, 1985) and hard to control.

Going back to the table, a different researcher will have, unlike in natural sciences, a much bigger influence on the outcome of the research. It can be shown that researchers in the majority of virtual replications change some part of the procedure (see later), because they do not agree fully with the categorizations made in the earlier research experiment.

Finally in Table 7.2, the difference of location is mentioned. Cross-cultural research has shown that the location of research in social science is most important. Class differences, differences in education, differences in language categories (Kagan, 1984) as part and parcel of shifting the location of research from one area to another affect validity of research data (Milgram, 1974).

TWO NON-FEASIBLE METHODS TO CONTROL FOR VALIDITY OF REPLICATIONS

All the factors mentioned earlier may, therefore, threaten validity of replications in experimental and—to a much greater extent—nonexperimental social scientific research. An immediate counterargument to these possible threats of validity may be that the nature of a *control group* allows for control over these possible threats of validity. Unfortunately, social science is not in a position to have something like an absolute freezing point. A control group controls only *within the structure of a study and within a particular society at a particular time.* It does not control for effects outside a system. Time differences, social and developmental changes are affecting both control and experimental groups alike and may be responsible for positive and negative effects. Control groups, therefore, are not absolutely at zero and do not constitute an absolute and lasting baseline. Two different control groups, one in a replication and one in the original study, do not control for the validity of

the replication but only for experimental effects in each of the two studies. Therefore, control groups do not control for threats of validity of replications.

Piaget (1970), on the other hand, forwards a theoretical perspective claiming that not more threats of validity have to be considered in psychology than in the natural sciences. Piaget replicates his findings with what he calls an *epistemic subject*. An epistemic subject is a universal subject, a generalized human knower (Wozniak, 1983, p. 328) who is vital to the constructive process of a theory. The outcome of a replicated task with a subject, conceived of as a generalized knower, is valid at both times even if the outcome is different—because a universal subject is studied. It is not the specific characteristics of the particular subject that is of interest. Therefore, it is assumed that reliable repeated observations, if they are contradictory, are an indication of a different underlying process and are not an indication of a factor threatening validity.

Validity is not, or is hardly ever, considered an issue in repeated epistemic observations. Differences in replications of tasks or observations lead to the construction of new fields in the theory. There is no question of validity of repeated observations because the observations are assumed to be epistemic and therefore unquestioned by nature. Validity is a characteristic of a verifying strategy, epistemic observations are characteristic of a constructive strategy. It therefore is misleading to claim that a trial with an epistemic subject validates a previous trial of another epistemic subject. After briefly having looked at the nature of replications and their validity, let me now proceed with the different types of replications. Because I am interested here in true validating strategies of psychological knowledge, I do not deal further with replications in epistemic observations but proceed directly to classes of replications in validating strategies.

CLASSES OF REPLICATIONS

The question of replication is dealt with much more frequently in sociology than in psychology. One of the reasons may be that many programs designed by sociologists are very costly and time consuming. For example a program of health education in a country may involve a huge body of personnel in order to reach the target community. It may be for this reason that the need and the strategies of replication have drawn considerable attention in sociology. I basically follow the classification of replications offered by Finifter (1972), but will extend it at various points.

Virtual Replications

Reseach is virtually replicated if an original study is repeated after some time by another research group at a different place (the characteristics of the other

research group and other place were not included by Finifter, 1972) with no or only minor changes in the design. The purpose of conducting it is to confirm or refute a previously obtained result. It can also be performed in order to assess the contribution of the research procedure to the findings of the original study. In order to demonstrate this, a condition of the original study is altered on purpose.

A hypothetical example of a virtual replication is offered by White and Pesner (1983). Although their analysis is restricted to what they call *nonexperimental research*, I think it is valid for most of the research in social sciences.

In both the original study and replication there exists a positive correlation between trait A and trait B. If we assume that the two trials are two draws of one population, the question of virtual replication is now, whether the second draw of a sample of a population can reasonably be assumed to belong to the same population or to a different one. Given a positive correlation in both trials and the marginal totals it can be calculated that 554, 400 ways of drawing the observations in the replication from the population in the original study [(n – m)! m! for each cell to cell comparison. For further details see White & Pesuer, 1983, p. 29.] are possible. Only four different 4-cell samples with the same marginal totals and a positive correlation as in the replication shown above are possible. Calculating the probability of occurrence for each of the four 4-cell samples leaves a probability of $p = .33$ for the replication sample shown in Table 7.3. The probability of the observed distribution or ones that are equally of less likely to occur by chance is .42. This is the nondirectional cumulative probability. The higher it is the more likely is replication from the same population. With a probability of .42 it must be assumed that the replication sample is not very likely a replication of the original study, but rather a sample from a different population (for details of the calculation of this generalization of Fisher's exact test see White & Pesner, 1983, p. 29).

Systematic Replication

Another class of replications can be called systematic replications (Finifter, 1972). The aim of such studies is to extend the original studies and to test further hypotheses that can be thought of on the basis of the original studies. Replication and extension, therefore, is the main feature of such studies.

Considering the method aspect of systematic replications, it is obvious that systematic changes in designs and tasks or procedures may enhance the credibility in the underlying theory or the effects found in the original study (Hunt, 1975). However, from a methodological point of view they are not replications. Rather, systematic replications follow the logic of theory construction outlined above. They are not replications in the methodological sense. It is important, however, to distinguish systematic replications from

Table 7.3. A hypothetical example of results of an original and a replicated study

		Trait A	
ORIGINAL STUDY		*absent*	*present*
Trait	present	3	8
B	absent	11	5
		Trait A	
REPLICATION		*absent*	*present*
Trait	present	1	5
B	absent	7	2

other forms of replication because of their close link to an original study and their aim to generate confidence in the findings.

Pseudoreplications

The most common form of replications are pseudoreplications. In order to understand pseudoreplications, a special interpretation of significance tests is needed. Finifter (1972) asserts that significance testing relies "on mathematically derived probability values rather than empirical events. The hypothetical nature of significance tests is manifested in the imaginary status of the repeated studies as simulations of identical replication and substitutes for virtual replication" (p. 162). The procedure of pseudoreplication repeats a study on certain subsets of an available total body of real data; it may repeat real data on artificial data sets which are intended to simulate the real data; and it may generate repeatedly completely artificial data sets according to an experimental prescription (p. 124).

A short hypothetical example of pseudoreplication can be the following: Assume in an epidemiologic study that the presence of schizophrenia in a community (for example, mental clinics in a certain area) is to be assessed. Three groups of researchers go to different places and count the number of schizophrenics and nonschizophrenics in these places.

Table 7.4 shows that the total number of schizophrenics in this hypothetical population of, for example, three mental clinics is assumed to be 78%. We want to know how reliable such an estimation based on three samples is. The

Table 7.4. Hypothetical example of repeated measurement in subsamples of a population

	Group 1	*Group 2*	*Group 3*	*Total*
(a)number of schizophrenics and non-schizophrenics	34	3	14	51
(b)number of schizophrenics	30	2	8	40
$f = (b/a)$	.88	.66	.57	.78

estimated standard error for $f = (40/51)$ can be computed with the formula (Finifter, 1972, p. 145):

$$f = \frac{{}^{f}\text{max} - {}^{f}\text{min}}{k} = \frac{0.88 - 0.57}{3} = 0.10$$

where the numerator is the difference between the maximum and the minimum subsample estimate and k is the number of subsamples.

In this example the pseudoreplication in the three random subsamples allows for an estimate of the presence of schizophrenia for the whole community on which the research was performed. One random sample in a community is studied. Another random sample—a replication of the first sample of the same community is then studied again, and so on until an estimate of the whole community on the studied variables can be made. This method, therefore, pseudoreplicates the observations in one subsample in a second one at the same time in order to estimate the whole community on the observed variables.

Comparisons between control groups and experimental groups belong to pseudoreplications as well, because there again the question of whether the experimental group is a replication of the control group, according to a hypothetically assumed distribution, is crucial for statistical comparisons. Here we find, however, a complication: Most of the time we have a system of variables that accounts for possible differences between experimental and control groups. Methods to deal with systems of variables in comparative pseudoreplications are again offered by White and Pesner (1983) and cannot be elaborated here.

So far I have mentioned 3 classes or replications: virtual, systematic, and pseudoreplications. Of these three classes only virtual and systematic replications confirm or disconfirm the conclusions of a study conducted at an earlier time. In the remainder of the chapter I proceed to a review of replications of experiments or studies done in psychology and make an attempt to reformulate the question of enhancing confidence in results of psychology. Pseudoreplications will not be included in this review. What we are interested in is virtual and systematic replications as they are done frequently in natural sciences.

REVIEW OF REPLICATIONS OVER THE YEARS 1983–1988

For this review 277 virtual and systematic replication studies, published in the major psychological journals, were retrieved from the database Psyclit, covering a time range between January 1983 to June 1988. Similar to the procedure of Bahr et al. (1983), the studies were selected on the basis of the word *replication* in the title of the study. It was assumed that close replications of original studies would make a reference to this procedure in the title.

Random inspections of the database without any key word at all showed that 95% of what safely can be called a replication could be obtained in this way.

Bahr et al. (1983) found over 300 replicative studies for the years 1973–1981. He offered a typology of replications based on time, place, method, and subject difference, which I do not follow here. The reason is that in order to assess the amount of virtual replications as defined above the study must be done by another research group. This category, however, was not included in the typology of Bahr et al..

The 277 virtual and systematic replications were distributed as shown in Table 7.5.

According to our criteria, only a little over 10 virtual and systematic replications can be expected in experimental and clinical psychology in one year (77 and 67 studies respectively, divided by 6 years). In child psychology, the expectation rate drops to about 5 per year, and about 7 out of the 8, 000 tests in active use (Shelley & Cohen, 1986, p. 7) may be expected to be reevaluated per year.

In clinical psychology, according to estimations by Grawe, Siegfried, Donati, and Bernauer (in preparation), Grawe, Siegfried, Donati, and Bernauer (1986), Grawe, Siegfried, Bernauer, Donati, and Louis (1987), and Siegfried (1985a, 1985b, 1986), about 400 controlled psychotherapy studies per year are published. The chances of a replication, therefore, are about 3% for each study per year according to our data. It is safe to say that even without discrimination of different classes of replication procedures virtual and systematic replications are very rare in psychology.

Dividing replications in clinical psychology further into different diagnostic categories leads to the distribution shown in Table 7.6.

Table 7.5. Frequency of replications in different areas of psychology

Psychological field	*N*
Experimental psychology	67
(includes perception, memory, communication, IQ, attitude, identity, perception of others)	
Clinical psychology	77
treatment	17
disorders	32
mental retardation	5
other	23
(survey, predictors, cross-cultural comparisons, reviews)	
Child psychology	29
Validation of measurements	39
other	58
(politics, education, medical studies, parapsychology, animal research)	
unclassified	7
TOTAL	277

Table 7.6. Breakdown of replicated clinical studies according to diagnostic categories

Disorder and author	*N and replicated finding*
SCHIZOPHRENIA	*8*
	Family
Strachan et al. (1986)	Families with high expression of emotions express more emotionally negative statements.
Vaughn et al. (1982)	Expression of excessive critique and emotional over-involvement lead more likely to relapse.
Vaughn et al. (1984)	Relapse best predicted by family members' expressed emotion.
Marcus et al. (1985)	Increased vulnerability for offsprings of schizophrenics.
	life events
Caton & Fraccon (1985)	More life events in schizophenics.
	Basic issues
Levinson et al. (1985)	Measurement of habituation in schizophrenics.
Wells & Leventhal (1984)	Schizophrenics fail to utilize gestalt grouping principles.
Shelly & Goldstein (1983)	Discrimination between brain-injured and chronic schizophrenic patients.
ALCOHOL	*4*
	treatment
Bell, Walsh, & Barnes (1985)	Acceptance of illness nature of alcoholism greater after treatment.
Duxbury (1983)	Ambivalence to treat alcoholic patients.
	relapse
Sandahl (1984)	Determinants of relapse (impulsiveness, adjustment of personal opinions, somatization).
	personality
Steer, McElroy, & Beck (1982)	Intrapunitive Self-Attitudes characterize alcoholics.
OBSESSIVE-COMPULSION	2
Grayson et al. (1986)	Exposure to most feared contaminant leads to heart rate reduction and subjective anxiety.
Salkovskis & Harrison (1984)	Characteristics of intruisive cognitions.
INSOMNIA	2
Hauri et al. (1982)	Two different forms of feedback benefited for tense and relaxed insomniacs.
Haynes et al. (1985)	Stress is a mediator of latency for return to sleep.
TICS	2
Miltenberger et al. (1985)	Successful application of two habit reversal programs.
Finney et al. (1983)	Habit reversal is an effective treatment.
MARITAL CONFLICTS	2
Scabini (1983)	Classifying couples is done on three dimensions: task leadership, conflict, and affect.
Elwood & Jacobson (1982)	Confirmation of low reliability of spouse observation.

Excluded from this table were categories with only one replication per disorder (firesetting, violent offenders, panic attacks, agoraphobia, headache, epilepsy, erectile dysfunction, suicide, neurology, delinquency, autism, test anxiety, depression, and behavior disorders). The categories of disorders were not inspected for their definition by the authors of the studies.

Table 7.6 shows that all together, 20 studies replicate research findings in various clinically relevant disorders more than once in the past 5 years. It can safely be concluded, therefore, that research data in clinical psychology are hardly ever replicated. Originality and novelty are still a major demand for publication of the research and replications are therefore scarce.

Dividing replications now into virtual and systematic replications and analyzing them for their outcome shows how little virtual replication is done in psychology in general.

In Table 7.7, it is analyzed how many replications were conducted supporting scientific theories on human beings, treatments, and disorders. Excluded from this analysis, therefore, were replications on measurements, animals, cross-cultural differences, parapsychology, metaanalysis, reviews, and studies in which it was unclear whether they confirm or not confirm original studies.

Table 7.7 shows that almost half of the analyzed studies are done by the same authors or research group. Systematic replication of previous research findings closely replicates studies done on the same hypothesis. In this way confidence in the finding of the original study is enhanced. Confidence in the findings of the original study is not enhanced by doing different experiments based on the same hypothesis, as Hunt (1975) tried to argue. Replications serve to enhance confidence in the research finding in the same way as replications in the natural sciences (Table 7.1). Validity, on the other hand, is tested by doing different experiments based on the same hypothesis in order to see how far the hypothesis can be generalized and what conditions have to be satisfied for a confirmation of the hypothesis.

Virtual replications in Table 7.7 are differentiated between identical, identical and different, and different outcome. This differentiation is not made in systematic replication. It proved that almost all systematic replications led to

Table 7.7. Number of studies in systematic and virtual replications

SYSTEMATIC REPLICATION	*N*
Same authors (often extensions of original studies)	63
VIRTUAL REPLICATION	*N*
Different authors, identical outcome in replication	39
Different authors, different outcome in replication	23
Different authors, identical and different outcome in replication	4
STUDIES ANALYZED	129

a confirmation of the previous findings. Systematic replications are also mostly done by at least one author who was also involved in the original study.

The frequencies in virtual replications reveal about a *30% chance of a replication to have a completely different result* than the original study. Compared to natural sciences (Schlenker, 1974), social sciences make virtually replicated observations as Boyle and all his followers have made them under each condition in only about a little more than 50% of the repeated trials. I wonder whether Boyle would have been able to arrive at a law about the resistance of air with such a small amount of few repeated observations.

The review of replications from 1983–1988 should have shown that virtual replication in psychology is almost totally absent. Secondly, it should have shown that analyzing an area of psychology like clinical psychology, and further subdividing it according to the various research areas, narrows the number of systematic and virtual replications again to a minimal amount. Even when replications are done, they are conducted in an unsystematic way (i.e., testing parts of a theory here and there). The result is highly contradictory and inconsistent research findings. Thirdly, our analysis should have shown that most replications are done as comparisons to a hypothetically existing population and are pseudoreplications.

The concluding part of this chapter is devoted to the causes of the scarcity of virtual replications. What causes can there be for the scarcity of virtual replications in psychology?

The questions of whether there have been enough (Hunt, 1975) or too few (Bahr et al., 1983) replications in psychology covers the more important question of whether replication at all is a goal in psychology, and why it is done after all. I argue that two rather unarticulated motives may prove to be key questions in future replication research.

One such unarticulated motive in psychological research is that research is not a kind of a natural science and, therefore, replications function not in the same way as in natural sciences. I have dealt with the problems of replications in natural and social sciences rather extensively earlier. In social sciences the meaning of replication is different from the natural sciences.

I think that many researchers and publishers in the social sciences may have implicitly admitted that fact by stressing originality and nonrepetitiveness of research. If this fact is accepted, it would make good sense to refrain from virtually replicating research in psychology.

A second motive of not replicating research in psychology is again a not outspoken knowledge of the hypostatic nature of much of psychology's theories and terms. This point needs clarification. Hyman and Wright (1967) have been concerned with the "deceptive" nature of the term *program* in sociology for a long time. All too often, the authors assert, a program is "simply a statement on paper of what the planners in an agency hoped to do that has never been fully translated into action by the field staff" (p. 744).

Without a careful evaluation, a program is deemed to remain a well sounding exaggeration of a possibility of action that never materialized. That is what is meant by the German term *hypostatisch* (hypostatic): The term promises too much and an idea of a really existing thing is formed on the basis of this term, whereas in reality not much is there. The same exaggeration of promising terms may be found in psychology. There are plenty of terms in psychology that seem to promise explanations or topics of interest in research. Looking, however, at the operationalization of these terms in each piece of research, more than often they represent rather raw clusters of everyday prejudices. A nice-sounding term like "behavior therapy" may mean anything from giving tokens up to teaching problem-solving skills to a patient. Such terms as behavior therapy are only a hypostatic object of research. In fact, what we should be interested in is which exact form of treatment with which patient under which circumstances leads to which outcomes (Kiesler, 1971). The contribution of research on behavior therapy, on self-assertiveness, on learned helplessness, on stages of moral judgment, and on anxiety can be judged only on the level of operationalization. It is only there that it is possible to see what actually has been done. Treatment of anxiety, for example, makes sense if and only if the anxieties involved are specified and the treatment for the specific anxieties is described. An intervention program must be described exactly, lest it will remain a hypostatic idea. A questionnaire assessing anxiety is important in the treatment of anxiety. Statements such as *I can't go out at night. I am afraid to speak up. I am afraid of bridges.* and *I am afraid of elevators* will make a difference in the treatment of anxiety and the methods chosen to treat it, even if all these treatments are labeled *behavior therapy*. Research on hypostaticly labeled *action or treatment programs*, and replications of such programs confirming or disproving hypostatic labels, will necessarily lead to much contradiction because a wide variety of definitions, leading to different operationalizations and different outcomes are subsumed under such labels. Knowing this danger, researchers decide not to replicate research in order to avoid contradiction. When investigations are described exactly, however, many contradictions in research results may be explained. But detailed explanation, taking into account much of the possible conditions and relevant factors, open such a wide variety of influences which differ in each study that virtual replication in the sense of the natural sciences (holding relevant conditions constant) is impossible.

TOWARD THEORETICALLY GUIDED SYSTEMATIC REPLICATION

Summarizing the results of this investigation of replication research in psychology, two conclusions can be drawn:

1. There are good reasons in psychology for not doing replications more frequently. Psychology has to acknowledge differences in its scientific endeavor with the natural sciences. More stress should be given to exact descriptions of operationalization procedures instead of inventing new, rather general, labels in order to describe and explain findings.

2. The enhancement of confidence in research results in psychology is not likely to be achieved by replications of studies and experiments (as it is done in natural sciences). Natural sciences' replications—holding conditions, contexts, procedures, independent and dependent measures constant—appear not to be possible in social sciences. How then can confidence be enhanced in social sciences? One route often taken in social sciences is enhancement of confidence by eloquent rhetorics. However, a well forwarded argument may set the assumptions such that it cannot be criticised or falsified. Science has adopted the principle of falsification as a core concept for its advancement. Rhetorics often do not adopt this principle and are, therefore, not a feasible way of enhancing confidence in the results of social sciences.

How then can we formulate a verification technique for the social sciences? It appears that confidence in research results in psychology can be achieved by conducting well described studies or experiments with as much as possible *identical methods* and assessment strategies under *different conditions*, in *different populations* and *different circumstances*. This would imply choosing methods in a way that they can be applied in different settings without losing any of their validity. However, without carefully *theorizing at the same* time the construction of experiments and their applications under different conditions will remain highly speculative.

What I am trying to suggest here is the view of theoretically guided systematic replication as the only nonpseudoverification technique in social sciences. However, systematic replications have to fulfill two conditions in order to be replications: (a) Method and assessment strategy must remain as constant as possible, and (b) Application of as much as possible identical methods and assessments in different circumstances and different conditions must theoretically be justified. If method and assessment are not identical in the replication, then the "replicated" study is, in fact, just another study hardly related to the original experiment.

If applications of identical methods and assessments in different circumstances or in different conditions are theoretically unjustified, there cannot be a clear hypothesis about the outcome of the replicated finding. The replicated finding may then either be due to the difference in circumstances of conditions, or due to changes in the procedure of the experiment. It is only on theoretical grounds that a clear formulation of expectations of outcome can be formulated.

An experimenter, for example, may want to replicate a study of a successful treatment A for phobic clients. Repeating the study in exactly the same way on a different sample of phobics leaves the result of the replication compared to the original study in any case unexplained. The result, if similar to the original study, can be due to identical or different conditions, leading to exactly the same outcome. If the replicated finding is different from the original study, only provisional suggestions accounting for the difference can be given.

If the replicated study leads to different results than the original study, and the replicated study is performed on a different group or in different circumstances, the differences of group or circumstances must be accounted for in a theory, leading to a specific hypothesis prior to the replication. Replicating the study of phobics for example with agoraphobics alone, therefore, requires first a theory of differences and similarities in phobics and the subgroup of agoraphobics which in its turn leads to a specific hypothesis about the effects of the replicated study on agoraphobics alone. This hypothesis is formulated on the basis of the first study and leads to a prediction of the replication outcome. Only with this theoretically formulated hypothesis can confidence in the result of the replicated outcome be enhanced.

If I am correct with my claims, a difference between virtual and systematic replication in social sciences can now be formulated as follows:

In virtual replications the interfering factors are hardly known. In fact it would need a theory of situation, contexts, and social changes in order to explain similarities or differences of repeated observations in social sciences. Because we do not have such a theory at this time, virtual replication is hard to do. If it can be done at all, at least statistical procedures as suggested above should be performed. The mere hypothesis that the repeated observation will be the same as the original one does not help here. On which grounds would such a claim be made? How could it be justified?

As outlined earlier, in theoretically guided systematic replications, keeping methods and assessment strategies as identical as possible, conducting the replication in different circumstances or different conditions and predicting the outcome of the replication on theoretical grounds, is like first introducing an *artificial* difference between the original study and the replicated study. This artificial difference (for example, studying population A in the original study versus population B in the replicated study) allows for a theoretical formulation of a specific hypothesis for the outcome of the replicated study. If this hypothesis is confirmed, confidence in the result of the original, as well as the replicated study, is enhanced.

In a systematic replication in social sciences, in my view, therefore, first the theoretical level (the rationale of the experiment or study and the rationale of replicating the earlier finding on a different population or in different circum

stances), second, the relation of theory to operationalization of its concepts, and third, its experimental reapplication in a specific setting must be considered. Replications in social sciences—different from natural sciences—have to be embedded in the whole frame of experimental construction because unlike repetitions of trials in natural sciences, social sciences repetitions are part of an irreversible ongoing process of social interaction and change.

REFERENCES

Allerbeck, K. R., & Hoag, W. J. (1984). Umfragereplikation als Messung sozialen Wandels (Survey-replication as a measurement of social change). *Kölner Zeitschrift für Soziologie und Sozialpsychologie, 36,* 755–772.

American Psychological Association. (1983). *Publication Manual of the American Psychological Association* (3rd ed.). Washington, DC: Author.

Bahr, H. M., Caplow, T., & Chadwick, B. (1983). Middletown III: Problems of replication, longitudinal measurement, and triangulation. *Annual Review of Sociology, 9,* 243–264.

Bell, A. H., Walsh, D., Barnes, T. (1985). Replication of alcoholism acceptance scores as a function of treatment. *Psychological Reports, 56* (1), 31–36.

Bulldock, R. J., & Svyantek, D. J. (1985). Analyzing meta-analysis: Potential problems, an unsuccessful replication, and evaluation criteria. *Journal of Applied Psychology, 70* (1), 108–115.

Campbell, D. R., & Stanley, J. C. (1963). *Experimental and quasi-experimental design for research.* Chicago: Rand-McNally.

Caton, G., & Fraccon, I. G. (1985). Life events and schizophrenia: A replication. *Acta Psychiatrica Scandinavica, 71* (3), 211–216.

Denzin, N. K. (1970). *The research act: A theoretical introduction to sociological methods.* Chicago: Aldine.

Duncan, O. D. (1969). *Towards social reporting: Next steps.* New York: Russel Sage Foundations.

Duxbury, R. A. (1983). Willingness of graduate social work students to treat alcoholics: A replication study. *Journal of Studies on Alcohol, 44* (4), 748–753.

Elwood, R. W., & Jacobson, N. S. (1982). Spouses' agreement in reporting their behavioral interactions: A clinical replication. *Journal of Consulting and Clinical Psychology, 50* (5), 783–784.

Finifter, B. M. (1972). The generation of confidence: Evaluation research findings by random subsample replication. In H. L. Costner (Ed.), *Sociological methodology* (pp. 112–175). San Francisco: Jossey-Bass.

Finney, J. W., Rapoff, M. A., Hall, C. L., & Christophersen, E. R. (1983). Replication and social validation of habit reversal treatment for tics. *Behavior Therapy, 14* (1), 116–126.

Furnham, A. (in press). Beliefs concerning human nature. Comparing current responses with those gathered in 1945 and 1956. *British Journal of Educational Psychology.*

Gasparikova-Krasnec, M., & Ging, N. (1987). Experimental replication and professional cooperation. *American Psychologist, 42,* 266–267.

Gergen, K. J. (1973). Social psychology as history. *Journal of Personality and Social Psychology, 26,* 309–320.

Grawe, K., Siegfried, J., Bernauer, F., Donati, R., & Louis, C. (1987). Qualitaetsanalyse der Psychotherapieforschung. Eine empirische Untersuchung [Qualtitative analysis of psychotherapy research. An empirical analysis]. *Schweizerische Zeitschrift für Psychologie, 46* (3/4), 259–266.

Grawe, K., Siegfried, J., Donati, R., & Bernauer, F. (1986, September). *The Bernese project of a differential analysis of psychotherapeutic outcome studies.* Paper presented at the EABT-Congress, Lausanne, France.

Grawe, K., Siegfried, J., Donati, R., & Bernauer, F. (in preparation). *Meta-analysis of controlled psychotherapy studies.* Bern: University of Bern.

Grayson, J. B., Foa, E. B., & Steketee, G. S. (1986). Exposure in vivo of obsessive-compulsives under distracting and attention-focusing conditions: Replication and extension. *Behaviour Research and Therapy, 24* (4), 475–479.

Haccoun, D. M., & Lavigueur, H. (1979). Effects of clinical experience and client emotion on therapists responses. *Journal of Consulting and Clinical Psychology, 47*, 416–418.

Harré, R. (1981). *Great scientific experiments.* Oxford: Phaidon.

Harris, F. N., & Jenson, W. R. (1985). Comparisons of multiple-baseline across persons designs and AB designs with replication: Issues and confusions. *Behavioral Assessment, 7*, 121–127.

Hauri, P. J., Percy, L., Hellekson, C., Hartmann, E., & Russ, D. (1982). The treatment of psychophysiologic insomnia with biofeedback: A replication study. *Biofeedback and Self Regulation, 7* (2), 223–235.

Haynes, S. N., Fitzgerald, S. G., Shute, G. E., & O'Meary, M. (1985). Responses or psychophysiologic and subjective insomniacs to auditory stimuli during sleep: A replication and extension. *Journal of Abnormal Psychology, 94* (3), 338–345.

Hunt, K. (1975). Do we really need more replications? *Psychological Reports, 36*, 587–593.

Hyman, H. H., & Wright, C. R. (1967). Evaluating social action programs. In P. F. Lazarsfeld, W. H. Sewell, & H. L. Wilensky (Eds.), *The uses of sociology* (pp. 741–782). New York: Basic Books.

Kagan, J. (1984). The idea of emotion in human development. In C. E. Izard, J. Kagan, & R. B. Zajonc (Eds.), *Emotions, cognition and behavior.* Cambridge, UK: Cambridge University Press.

Kiesler, D. J. (1971). Experimental designs in psychotherapy research. In A. E. Bergin & S. L. Garfield (Eds.), *Handbook of psychotherapy and behavior change: An empirical analysis* (pp. 36–74). New York: Wiley.

Klein, D. F., Zitrin, C. M., Woerner, M. G., & Ross, D. C. (1983). Treatment of phobias: Behavior therapy and supportive psychotherapy: Are there any special ingredients? *Archives of General Psychiatry, 40*, 139–145.

La Sorte, M. A. (1972). Replication as a verification technique in survey research: A paradigm. *The Sociological Quarterly, 13*, 218–227.

Levinson, D. F., Edelberg, R., & Maricq, H. R. (1985). The skin conductance orienting response in neuroleptic-free schizophrenics: Replication of the scoring criteria effect. *Biological Psychiatry, 20* (6), 646–653.

Marcus, J., Hans, S. L. Mednick, S. A., Schulsin, F., & Michelse, N. (1985). Neurological dysfunctioning in offsping of schizophrenics in Israel and Denmark: A replication analysis. *Archives of General Psychiatry, 42* (8), 753–761.

Milgram, S. (1974). *Obedience to authority.* New York: Harper & Row.

Miltenberger, R. G., Fuqua, R. W., & McKinley, T. (1985). Habit reversal with muscle tics: Replication and component analysis. *Behavior Therapy, 16* (1), 39–50.

Mitroff, I. I. (1974a). Norms and counter-norms in a select group of the Apollo moon scientists: A case study of the ambivalence of scientists. *American Sociological Review, 39*, 579–595.

Mitroff, I. I. (1974b). *The subjective side of science.* Amsterdam: Elsevier.

Piaget, J. (1970). *Genetic epistemology.* New York: Columbia University Press.

Rodrigues, A. (1982). Replication: A neglected type of research in social psychology. *Revista-Interamericana-de Psicologia, 16* (2), 91–109.

Salkovskis, R. M., & Harrison, J. (1984). Abnormal and normal obsessions: A replication. *Behaviour Research and Therapy, 22* (5), 549–552.

Sandahl, C. (1984). Determinants of relapse among alcoholics: A cross-cultural replication study. *International Journal of the Addictions, 19* (8), 833–848.

Scabini, E. (1983). Typology of marital interactions: A cross-national replication. *Archivio di Psicologia, Neurologia e Psichiatria, 44* (1), 7–32.

Schlenker, B. R. (1974). Social psychology and science. *Journal of Personality and Social Psychology, 29* (1), 1–15.

Shelley, D., & Cohen, D. (1986). *Testing psychological tests.* New York: St. Martin's Press.

Shelly, C., & Goldstein, G. (1983). Discrimination of chronic schizophrenia and brain damage with the Luria-Nebraska Battery: A partially successful replication. *Clinical Neuropsychology, 5* (2), 82–85.

Siegfried, J. (1985a, June). *An evaluation system for the assessment of controlled psychotherapy studies.* Paper presented at the EABT-Congress, Munich.

Siegfried, J. (1985b). The quality of psychotherapy research—some prominent studies revisited. In W. Huber (Ed.), *Progress in psychotherapy research.* Louvain-la-Neuve: Presse Louvain-la-Neuve.

Siegfried, J. (1986, August). *Analysis of psychotherapy research: Empirical results for behavior therapy.* Paper presented at the SPR-Congress, Ulm, Germany.

Sjöberg, L. (1982). Logical versus psychological necessity: A discussion of the role of common sense in psychological theory. *Scandinavian Journal of Psychology, 23,* 65–78.

Smedslund, J. (1979). Between the analytic and the arbitrary: A case study of psychological research. *Scandinavian Journal of Psychology, 20,* 129–140.

Smith, G. P. (1983). The problems of reduction and replication in the practice of the scientific method. *Annals of the New York Academy of Sciences, 406,* 1–4.

Smith, N. C. (1970). Replication studies: A neglected aspect of psychological research. *American Psychologist, 25,* 970–975.

Steer, R. A., McElroy, M. G., & Beck, A. T. (1982). Structure of depression in alcoholic men: A partial replication. *Psychological Reports, 50* (3), 723–728.

Stouffer, S. A. Suchman, E. A., DeVinney, L. E., Star, S. A., & Williams, R. M. (1949). *The American Soldier: Adjustment during army life.* Princeton, NJ: Princeton University Press.

Strachan, A. M., Leff, J. P., Goldstein, M. J., Doane, J. A., & Burtt, C. (1986). Emotional attitudes and direct communication in the families of schizophrenics: A cross-national replication. *British Journal of Psychiatry, 149,* 279–287.

Valsiner, J. (1985). Common sense and psychological theories: The historical nature of logical necessity. *Scandinavian Journal of Psychology, 26,* 97–109.

Vaughn, C. E., et al. (1982). Family factors in schizophrenic relapse: A replication. *Schizophrenia Bulletin, 8* (2), 425–426.

Vaughn, C. E., et al. (1984). Family factors in schizophrenic relapse: Replication in California of British research on expressed emotion. *Archives of General Psychiatry, 41* (12), 1169–1177.

Wells, D. S., & Leventhal, D. (1984). Perceptual grouping in schizophrenia: Replication of Place and Gilmore. *Journal of Abnormal Psychology, 93* (2), 231–234.

White, D. R., & Pesner, R. (1983). Internal replication, the systems concept, and sources of validity in nonexperimental research. *Behavior Science Research, 18* (1), 26–44.

Wozniak, R. H. (1983). Is genetic epistemology of psychology possible? *Cahiers de la Fondation Archives Jean Piaget, 4,* 323–338.

PART III

Replication In Practice: The Interdependence Of Theories And Empirical Context

Part III of this volume includes particular efforts of replication of specific studies that are of crucial relevance for contemporary research on human development. Three content domains were selected to be included here; (a) the "activity theory" focus on memory organization of children, (b) the "cultural-historical theory" and its focus on the use of external mediating devices in internalization of higher psychological functions, and (c) the "attachment theory" in its efforts to predict the future outcomes of child development.

Chapters 8–10 are devoted to the replication of the work of Zinaida Istomina on children's mnemonic functions. Her study is a rare empirically sophisticated work in the history of Soviet "activity theoretic" research, with recording and reporting of relevant descriptive details of the experimental setting that was meant to show the emergence of memory functions in the context of meaningful activity. In Chapter 8, Jaan Valsiner provides a background description of Istomina's original study. This overview is followed by Wolfgang Schneider and Markus Hasselhorn (Chapter 9) who outline the results of their systematic research program in which they have attempted to replicate Istomina's experiment in a careful, well-controlled manner. As the readers can see, their efforts have resulted in both partial replication and nonreplication of Istomina's results. Trisha Folds-Bennett (Chapter 10) analyzes the issue of comparability of the experimental methods that all existing efforts to replicate Istomina's work have used. Her final conclusions should lead us to a new way of looking at replicability—given the cultural (personal-experiential) differences between the children in USSR, Germany, and the United States (where Istomina's work, and its replication efforts have taken

place), exact reconstruction of functionally similar experimental settings may be an impossible task for any experimenter. Where would that realization take our discussion of replicability as a desired state of affairs (and a canon of scientific validity)? Some answers to that difficult questions will be reflected in the Epilogue to this volume.

The replication efforts in the domain of "cultural-historical theory" of Lev S. Vygotsky are represented here by the efforts of René van der Veer (Chapter 11) to replicate the empirical work of Alexei N. Leontiev from the times when the latter was still a disciple of Vygotsky's. The "Forbidden Colors Game" was one of the experimental methods that was used in late 1920s to demonstrate the use of external "stimulus means" (mediating devices) in the internalization process. However, the replication efforts described in Chapter 11 make it doubtful that this basic tenet of the "cultural-historical theory" is actually verified by the particular experimental method.

Finally, in Chapter 12, Frans Hubbard and Rien van IJzendoorn report an effort to replicate a study by Bell and Ainsworth (1972) on the linkages between infant crying and maternal responsiveness, and the predictions from these. In the present-day emphasis on "responsive parenting" in most affluent First World societies, the realm of "attachment theory" that propagates the relevance of secure emotional ties with the children, has gained ground. The original findings of Bell and Ainsworth have become a convenient "myth-truth" in contemporary child psychology. It is therefore quite important that replication efforts be performed on that study, as the relevance of these efforts (especially if some inadequacies of the "target study" are overcome) is substantial. All in all, Part III should elaborate the general ideas discussed previously among the wealth of empirical detail.

REFERENCE

Bell, S., & Ainsworth, M.D.S. (1972). Infant crying and maternal responsiveness. *Child Development, 43*, 1171–1190.

CHAPTER 8

Replicability In Context: The Problem of Generalization

Jaan Valsiner
Department of Psychology
University of North Carolina at Chapel Hill

The coverage in this book from now on moves from the general issues of replicability and its methods to the analyses of examples of replication efforts. In these efforts, the meaning of "replicability" in science begins to acquire some—indeed paradoxical—specificity. When the question of replicability is asked with respect to a particular empirical study that has been available in the literature, the general imperatives shared by scientists in their collective culture show the need for clarification—*what* is it exactly that we try to replicate, and *how* can replication of whatever we try to replicate be possible under conditions that vary from the original study? Finally, the issue of whether any result (success or lack of it) of our efforts to replicate has any substantial impact upon our scientific knowledge remains luring us on the horizons of our "scientific visions."

THE ISSUE OF REPLICABILITY AS TRANSFER

I argue here that the problem of replicability in science has the same underlying epistemological structure that the question of transfer (of skills or knowledge) has within human development. In the latter domain, novel psychological capabilities that emerge in a specific context need to be transferred to other (novel) settings, if these capabilities are of serious consequences for development. The issue of replicability in science is similar—certain findings obtained in a specific context (experimental setting) reflect knowledge about the causal mechanisms of the phenomena under investiga-

tion. The adequacy of that knowledge is to be tested in contexts other than the original one, thus implying transfer.

Similarly to the issues reflected in the case of transfer, the question of replication can be addressed at either the "horizontal" or "vertical" plane. In the former, the issues of "virtual" or "exact" replication are made to be of central importance—a particular empirical finding of a "target study" is attempted to be replicated by trying to repeat the exact conditions under which the first finding was obtained. As should be obvious from the previous chapters in this volume, the framing of the issue of replicability in its "horizontal plane" leads to practical difficulties on the one hand, and to devaluation of science as an enterprise of general knowledge, on the other.

Given the unfeasibility to view replication as a "horizontal" transfer, we are left with the possibility of viewing replicability in science in ways analoguous to "vertical" transfer. A certain finding obtained in the "target study" allowed the scientists to posit the existence of an explanatory model (=generalization), the adequacy of which needs to be tested in other empirical contexts (=particularization). These other contexts may be created to be similar to those of the "target study," but can also differ substantially (since it is the generalized model, and not the particular findings, that is being replicated). As will become clear from the next chapters, the virtual reconstruction of the setting of the "target study" may turn out to be an impossible task for psychological research where the meaning of the settings in which subjects are in is co-determined by the subjects. Hence, even the most meticulous *formal* effort to repeat the "target study" in detail may fail to do so—all the more if the replication efforts take place in other countries and historical time periods than the original.

RE-VIEWING ISTOMINA: A CROSS-NATIONAL EXPERIMENT IN REPLICATIONS

The next two chapters are devoted to the issue of replicability in a very concrete sense—that of replication of the results of a particular study on children's memory (Istomina, 1948). It would look like a rather ordinary effort to replicate a study of children's memory, were it not the case that these replication efforts cross the boundaries of historical time (i.e., replicating a study published in 1948 in the 1980s) and countries. All the presently existing replication attempts have been made in Western Europe and North America, while the "target" of replication originated from the Soviet Union. This creates an additional complication for all the replicators, who have to transcend their time and scientific socialization in order to be able to address the replicability issue.

The "target" of the replications—the work of Istomina (1948, 1977)—is not

a "random choice" from the vast existing literature on children's memory development. Its main focus—that of the context-embeddedness of developing psychological functions—is very "modern" in the child psychologies of the 1980s. The bifurcation of approaches to memory development that would result from either replication of Istomina's work, or from consistent failure to replicate it, is of fundamental nature. In case Istomina is proven unreplicable, research on children's memory development can continue along the traditions of the rest of memory research, in which the phenomenon (mnemonic functions) is separated from its wider context (motivation of subjects and the meanings of the experimental situation for them), and then subjected to experimental study. The results of such study should be interpretable purely in terms of any theory of memory and its development—without any worries about "other factors" (of the context) interfering with the purity of the experimental data.

In contrast, if Istomina's work is replicated, then most of the research on children's memory development has to be reoriented in its methodology. No longer would the "purity of the experiment" be usable as interpretable in terms of theories of memory per se, because the "contextual factors" would be the main agents that guide the development of memory. In this case, the object of investigation would have to be transposed from studying memory in itself to the research on the ways in which the "context" interacts with the mnemonic processes. That transposition would require a major change in methodology ("classically pure" application of the experimental method becomes highly complicated), and an equally complicated breakthrough in the theoretical sphere ("theories of memory" will become "theories of child–environment relationships" that are explicated on the basis of the study of mnemonic functions). Thus, Istomina's work constitutes an *experimentum crucis* for a vital tradition in our contemporary developmental psychology of memory.

WHAT DID ISTOMINA SET OUT TO ACCOMPLISH?

The particular memory study conducted by Istomina (1948) stands at the crossroads of different perspectives on memory development that emerged in Soviet psychology from the late 1920s onwards. On the one hand, the "cultural-historical approach" of Lev Vygotsky emphasized the relevance of the emergence of "higher" (voluntary) forms of psychological functions that were seen as qualitatively new forms emerging from the "lower" (involuntary) functions. On the other hand, the "Kharkov school" of Soviet psychology (A.N. Leontiev, P.I. Zinchenko, and others) followed the line of the general *Zeitgeist* of the 1930s and built their psychological theories around the idea of activity of persons in their environmental settings. In that focus,

the "Kharkov school" continued the long tradition of emphasizing the role of active person-in-environment in Russian psychology (Lazurskii, 1916; Basov, 1929).

By emphasizing the role of activity in the development of psychological functions, the research orientations emerging from the "Kharkov school" needed to include the context in which that activity took place. In line with the traditions of experimental psychology, the first step towards "taking context into account" was to vary it by some manipulation by the experimenter, and then demonstrate that this manipulation has an effect on the psychological functions under study. This "effect" can be conceptualized in two different ways. First, it can be demonstrated on the basis of the *outcomes* of the psychological functions in question, without any reference to the processes that led to these (affected) outcomes. Secondly, it can be demonstrated by looking at the ways in which the processes (that generate outcomes) change themselves, when the context of their functioning is changed. Obviously, the change in the functioning of processes is more basic than a mere change in the outcomes.

Istomina's experiment was set up explicitly under the general ethos of the "Kharkov school," especially in continuation of the line of thought that emerged in the work of P.I. Zinchenko on involuntary memory (Istomina, 1948, p. 52, 1977, p. 102). Summarizing the focus of her predecessors, Istomina set up the theoretical background of her study:

> It was demonstrated that involuntary memory is a process mediated ["obuslovlennyi" in Russian] in a complex manner by the content and structure of the particular, always meaningful ["osmyslennyi" in Russian] activity into which it is included and on the development of which it depends. Its difference from voluntary memory is not constituted by its own nature, but determined by the task to which it corresponds. If retention becomes subservient [Russian: "stanovitsia podchinennym"] to the special task "to remember," and does not emerge in the process that corresponds to some other task, then it becomes voluntary retention. In that process it becomes a special kind of internal act, i.e., being oriented towards a definite goal that is in some way motivated and can be attained with the help of definite means-operations ["sposoby-operatsii" in Russian]. Istomina, 1948, p. 52; see also Istomina, 1977, p. 102)

Following this line of reasoning about task relatedness of the development of voluntary memory, Istomina proceeded to set up her specific "working hypothesis" for the investigation of the emergence of voluntary memory:

> We assumed that at first, in the earliest and the pre-preschool age, the processes of memory (memorizing ["zapominanie"], recall ["pripominanie"] are non-independent (merely belonging to the structure of one or another activity) and involuntary processes. Later, specifically in the middle-preschool age, these

> processes transform into special internal acts, that is—they become consciously goal-oriented and voluntary. In other words, we assumed that the child's transition to voluntary memory constitutes a result of differentiation of a special kind of acts in his activity—the goal of which is "to remember" or "to recall." Consequently, this transition presumes, first, the presence of motives that can make these "mnemonic" goals meaningful for the child. Secondly, it presumes that the distinguishing of these goals emerges in reality, i.e., that there exists the process of formation of conscious mnemonic intention, and [the process of formation] of the acts of remembering and recalling. Thirdly, and finally—it is necessary that the child also mastered some means of execution of voluntary acts of remembering and recalling. These acts, of course, do not emerge in the child immediately in their developed forms, but are formed gradually, by emergence of different forms that depend upon the factors described above. (Istomina, 1948, p. 53; see also Istomina, 1977, pp. 102–103)

Thus, it becomes evident that Istomina's "working hypothesis" was one of developmental differentiation of the memory functions, the emergence of differentiated "mnemonic goals" from the less differentiated embeddedness of memory in activity context, and the emergence of conscious control over the "mnemonic goals" and "mnemonic means" (acts that lead to the attainment of these goals) that have become internalized. Istomina's reference frame was explicitly developmental in the more restricted sense of this term—it dealt with the processes of emergence of novel psychological functions on the basis (and assistance) of earlier ones. In harmony with this reference frame, Istomina set herself two specific goals for her experimental study: (a) to find out the conditions under which the child differentiates the goals of "to remember" and "to recall" for the first time, and (b) to reveal the forms in which the acts of remembering and recalling first emerge, and to find out how they at first function (i.e., to reveal the nature of memory operations). The comparison of different contexts in which remembering and recalling take place was thus a means to the end of understanding of the development of novel mnemonic functions for Istomina, rather than a goal in itself.

HOW DID ISTOMINA STUDY THE MNEMONIC PROCESSES?

Istomina (like the majority of investigators before and after her) could not escape the methodological peril of developmental psychology—that of the necessity to study development through a cross-sectional design. Although her "working hypothesis" of differentiation of the structure of activity and mnemonic processes calls for the use of a longitudinal design, the age range covered (3 to 7 years) was sufficiently large to render the use of longitudinal design practically impossible. Furthermore, it could be assumed that the need for a longitudinal design was not clearly recognized by Istomina, since most of

the research on child development that was close to hers (e.g., Piaget, the "Kharkov School") did not emphasize nor use the longitudinal design either. Instead, the emphasis on the differences *between* contexts of activity begins to dominate Istomina's reasoning as she proceeds to the empirical setup of the study (Istomina, 1948, p. 53, 1977, pp. 103–104). In fact, the *ontogenetic* framework of the "working hypothesis" became translated into a *microgenetic* framework of the empirical study settings, in which the emergence of novel psychological organization was hoped to be triggered and followed:

> With that goal in mind, in addition to the series of experiments that were carried out along the customary lines of laboratory experiment on remembering and recalling of meaningful words (we carried those out with children as special kinds of "lessons" ["zaniatia"]), a special game of the type of ordinary role-play was organized. That game entailed a specific meaningful ["soderzhatel'nyi"] *situation that created the motive for remembering and recalling., i.e., created* the conditions for the distinguishing and conscious focusing on the goal of remembering and recalling by the child. For the main series of experiments we took two similar to each other simplest scripts of games that frequently occur in ordinary games of preschoolers: the play of "store" and the play of "kindergarten." We connected these two into one united, general script.
>
> For playing "store" we gave the children the following play materials: a toy cash register and scales, sticks, sand, clay, small thematic toys, paper, etc. For *the play of "kindergarten"—a toy stove, dishes, dolls, etc. Besides, we also* prepared beforehand the "admission slips" ["propuska"] to the "store" with the names of "buys" for reasons that will become clear later.
>
> Six children were engaged in the play at the same time. Three of them were given the roles of the employees of the "store" ("salesperson," "cashier," "doorman" [Russian: "shveitsar", or "guard"—as translated in Istomina, 1977]), the other three—the roles of the employees of the kindergarten ("teacher," "cook," "supplies' manager" ["zavhoz"]). The experimenter and her assistant *obligatorily participated in the play. One of them took upon her the role of the* "store manager", the other—that of the "director of the kindergarten."
>
> The game was played in two adjoining rooms. In one of the rooms the setting of the "kindergarten" game was organized. In the beginning of the game the experimenter turned to her subjects with the suggestion "Children, let us play together. We will have a kindergarten." This was followed by the distribution of roles between the participants and explanation of what each of them will be doing. The children arranged the dolls and their belongings, set up the stove, *reorganized the dishes etc. The life of a kindergarten began. The "teacher" was* planning to take "children" for a walk, the "cook" was busy preparing the meal, the "supplies manager" carried out the orders of the "director," giving out groceries and bringing in firewood on a truck.
>
> At the same time in the other room the game of "store" was being created. The experimenter's assistant, who took upon her the role of "store manager," discussed with the children who of them will be the "salesperson," who—the "cashier," and who—the "doorman" ["guard"]. The duties of each of the

participants were clarified, and the set-up of the "store" started. Some goods were displayed on the counter, others were distributed behind it. The scales with weights, bags, and wrapping paper were placed on the counter. The "cashier" prepared the "receipts" and "money". Finally, everything was ready. The "manager" sent the "doorman" to the "kindergarten" to announce that the "store" is open.

In response to that announcement, the "director" of the "kindergarten" called upon a child who was participating in the game and gave the task of going to the store to buy things for the kindergarten. The task was always phrased identically: "Here is the *propusk* to the store, go and buy . . ." (the "director" slowly and clearly named a number of objects). After that, the child-subject was given the "propusk," "money," and a basket for the purchases.

In the "store," the subject showed his "propusk" by the demand of the "doorman," and turned to the "manager." The "manager," taking the "propusk" away from the buyer, asked "What have you been ordered to buy?" The "manager" then listened to what the buyer told her and at the same time checked off the named objects on the "propusk," marked their recall order by numbers, and added items that the child mentioned although these were not included in the list of words-items. If the child forgot any of the items, he was asked an additional question "But what else do you need to buy?" After that, the buyer paid the "money" to the "cashier", received the "receipt," went to the "salesperson" who gave him the goods, and then returned to the "kindergarten." The "propusk" to the "store" was handed back to the experimenter-"director."

The game continued until all the children in the "kindergarten" had been to the "store." Then the "store manager" announced that all goods were sold out and the "store" is closing. That was the end of the game. (Istomina, 1948, pp. 53-55; see also Istomina, 1977, pp. 104-106)

The recording of the data took place in two parallel forms. First, the *propusk* (admission slip) constituted the list of words that the experimenter read to the child, that the child took to the experimenter's assistant who recorded the recall of the words on the slip. Secondly, both the experimenter and the assistant (separately) recorded their observations of the child's actions in the process of performing the tasks. If the data from the admission slips constituted the basis for the quantitative analyses, then the recordings of the actual actions were the bases for the qualitative analyses that Istomina reported.

The second setting of the experiment was made as close to the "lesson" as possible:

> The child was called by the experimenter to her "for a teaching/learning session" ["dlia zaniatii"], and the experimenter set up the task for the child to attentively listen to the words, try to remember them, in order to later tell which words were read to him. The words—analogical in their meanings to the ones

used in the game instruction-task—were named at about 3-second intervals. After a short pause (60–90 seconds) the child was to recall, if possible, the whole list of words; if the subject did not name all words in the list, the experimenter asked him: "And what else did you have to remember, which other words?" The records of the data were kept in a standard way.

In one respect these experiments were sharply different from the experiments with game-based task: the remembering in these was very differently motivated. The goal of remembering was given to the child by the experimenter, whereas in the game-based task instruction the child had to distinguish that goal by himself. In all other respects both kinds of experiments were balanced with one another, in both cases the task consisted of voluntary remembering and recalling. The materials to be remembered were similar both qualitatively and quantitatively (lists of 5, or in a second series 8, items were given in both conditions), in a similar vein additional questions were given if necessary, and, finally, the time distribution of the whole process—the interval between the named words, and that between the last word and the beginning of the recalling, were the same. (Istomina, 1948, pp. 55–56; see also Istomina, 1977, p. 107)

Thus, Istomina created a contrast between externally (contextually) given conditions for differentiating the mnemonic task from the ongoing activities. This contrast could be expected to demonstrate the process of differentiation of the mnemonic task (in the social—"kindergarten" play—setting), while in the "lesson" setting the memorization goals were purposefully set up in the extreme of experimenter-dominated "demands" that could not have linked at all with the child's own understanding of the need for memorization and recall. Thus, we could view Istomina's experiment as including an "experimental condition" (the game-setting) that is being viewed on the background of the "lesson setting" (which serves as a "control condition"). This asymmetric distribution of the "roles" of these settings seems to follow from the theoretical underpinnings of Istomina's study (it should be remembered that Istomina herself was first and foremost an experimenter, rather than a theoretician). As we will see in the next two chapters, the particular focus of theoretical interests that different users of settings similar to Istomina's have used, flavors the process and outcomes of the replications efforts.

FROM ISTOMINA'S STUDY TO ITS REPLICATIONS

The next two chapters will provide an in-depth elaboration of all the replication efforts of Istomina's study that are known in international research literature so far (mid-1989). First, the chapter by Schneider and Hasselhorn guides the readers systematically through the replication efforts that the authors and the existing literature have reported before. Their chapter

includes also empirical data on their new replication efforts that have not been published previously. In a systematic manner, Schneider and Hasselhorn sieve through the experimental details of different replications, control a number of previously uncontrolled factors, and move towards finding an answer to the replicability issue by a carefully designed set of steps. The chapter by Folds turns the Istomina-replications issue from the focus on outcome data to the analyses of the procedural aspects of the replication efforts. She shows that all the replication efforts differ in procedural details (sometimes intentionally, but often unintentionally) from the original procedure, which—in an interesting conjecture that is of relevance for our theoretical understanding of replicability—might be in principle not repeatable because of its cultural and historically unique everyday life settings in the Soviet Union prior to 1948. These chapters will elaborate the "vertical transfer" notion of replicability as I have outlined here.

REFERENCES

Basov, M.I. (1929). Ocherednyie problemy psikhologii. *Estestvoznanie i marksizm*, No. 3, 54–82.

Istomina, Z.M. (1948). Razvitie proizvol'noi pamiati v doshkol'nom vozraste. *Izvestia Akademii Pedagogicheskikh Nauk RSFSR, 14*, 51–88.

Istomina, Z.M. (1977). The development of voluntary memory in preschool-age children. In M. Cole (Ed.), *Soviet developmental psychology: an anthology* (pp. 100–159). White Plains, NY: M. Sharpe.

Lazurski, A.F. (1916). Lichnost' i vospitanie. *Vestnik psikhologii, kriminal'noi antropologii i pedologii, 12* (2–3), 4–13.

CHAPTER 9

Situational Context Features And Early Memory Development: Insights From Replications Of Istomina's Experiment

Wolfgang Schneider and Marcus Hasselhorn

Although the need for replication of empirical studies has been emphasized by numerous researchers in various fields, the total number of true replication studies seems comparably small, at least in the case of experimental studies conducted in the social sciences. How can we account for this obvious discrepancy? In our view, several factors contribute to this unsatisfactory situation. First, conducting a true replication study does not seem very attractive to many because the benefit is small at least. If the replication study confirms the findings of the original study, this result gives further credit to the author of the original study. As the author of a replication study usually does not invest much of his or her creativity in the study but simply tries to follow the original design as closely as possible, the successful replication does not enhance the researcher's reputation other than illustrating that he or she is principally able to carry out experimental research properly. If a replication is unsuccessful, the case is even worse: Given the fact that in most cases it is impossible to exactly reconstruct the conditions of the original experiment, the author of the replication study may have a very difficult time convincing colleagues or reviewers that the deviating findings of the replication study do indeed point to flaws in the original study and cannot be explained by deviations in the experimental procedure of the replication study or by differences in the recruitment of subjects, and so on.

As a consequence, it appears that authors of replication studies should try to reconstruct the experimental conditions of the original study as closely as possible. In most cases, this may lead to additional problems because the detailed information needed for such a purpose is usually not provided in published materials. From our experience, getting access to more detailed

information means a considerable investment of both time and energy, going far beyond the amount of effort typically invested in experimental studies.

A third and even more important problem is that such an effort is usually not well-received by editors of scientific journals. It appears to us that demonstrations of replicability do not rank particularly high on journal editors' priority lists. Again, the situation seems even worse in the case of failure to replicate because this generally implies nonsignificant results, a finding regarded as uninteresting and of little scientific value by general journal policy (cf. later a recontinuation of this point).

Given these problems, it is not surprising that the number of published replication studies is that small. In our view, this is really unfortunate. As many experimental studies are based on relatively small sample sizes, it is by no means obvious that their findings represent stable and easily replicable effects. Even if facts seem established at a specific point in time, they may only have transient validity in a rapidly changing society.

It is the major purpose of this chapter to demonstrate how our understanding of young children's memory development can be improved by way of replication efforts of classical studies. The example chosen concerns the impact of situational context on young children's memory performance, a topic first systematically addressed in an important study conducted by Z. M. Istomina in 1948. During the past 40 years the results of this study have considerably influenced our view of the impact of motivational and social variables on preschool and kindergarten children's memory development. It is interesting to note that, at least to our knowledge, no replication study was conducted before the mid-1980s. In the remainder of this chapter, we want to address this issue in more detail. After a short description of Istomina's study, we first focus on the question of why there were no replications of Istomina's experiment for almost 40 years. Next, the results of two recent replication studies are discussed in detail, with an emphasis on the reasons provided by the authors for conducting these studies. Although the results of these two studies differ in several aspects, they suggest the conclusion that Istomina's findings cannot be replicated. However, although these two studies told us a lot about possible methodological problems of Istomina's study, findings of a more recent replication study indicate that this may not be the whole story. This replication study, consisting of two related experiments, will be dealt with in more detail in the final part of this chapter. Interestingly enough, the first of these experiments indeed replicated Istomina's findings, although it improved over the methodological weaknesses of the original study. A closer inspection of procedural details revealed that specific problems associated with the context manipulation and not explicitly considered in the former replication studies could have contributed to this outcome. This possibility was systematically explored in a second replication experiment which provided a more complete picture of potential problems with replicating the

Istomina experiment. The results of this study serve as a basis for the discussion of possible implications and for concluding comments on this topic.

ISTOMINA'S THEORETICAL APPROACH

Istomina was primarily interested in the question of when and under which circumstances voluntary, deliberate memorization first emerges. Her working hypothesis was based on Soviet theories about the development of voluntary memory in children (cf. Leontjev, 1928, cited by Leontjev, 1977). According to this theoretical approach, memory processes are not independent but integrated into some other activity during the preschool age period and are thus involuntary. Later in development, these processes are transformed into internal acts, thereby becoming purposeful and voluntary. Istomina assumed that the following preconditions are required for such a transformation: First, there should be motives and/or motivational incentives making mnemonic goals meaningful; second, these mnemonic goals should have some concrete reality; and third, the child should have means (strategies) available supporting voluntary memorization. The aims of Istomina's study were (a) to explore the conditions under which a young child is able to deliberately remember information, and (b) to determine the forms in which the acts of memorization first occur, with a focus on the memory operations by which they are affected.

DESIGN AND RESULTS OF THE ISTOMINA EXPERIMENT

Istomina created two settings—a game and a lesson—for children to remember words. Both conditions of the experiment were similar in form but differed functionally. In the lesson condition, the mnemonic goal was explicitly stated by the experimenter and was the only goal of the task. In the game condition, the children had to set goals by themselves (for details of the procedure, see Part III, Introduction). Istomina's hypothesis was that the game situation would be more familiar and also more motivating, and thus, young children would remember more in the game condition compared to the lesson condition.

Istomina tested a total of 60 children of the four ages 3–4, 4–5, 5–6, and 6–7. The same children participated in both conditions. As Istomina expected, recall improved notably with age, and children's recall in the game condition was superior to lesson recall at every age level. See Table 9.1 for an overview of these findings. Particularly poor memory was observed in the 3-year-olds.

Informal and qualitative observations of the children's behavior led Isto-

Table 9.1. Mean number of items recalled by preschoolers and kindergartners as a function of age and activity (Lesson vs. Game)*

	Activity	
Age (yr)	*Lesson*	*Game*
3–4	0.6	1
4–5	1.5	3
5–6	2.0	3.2
6–7	2.3	3.8

*Reconstructed from "The Development of Voluntary Memory in Preschool-age Children," by Z.M. Istomina, 1977, *Soviet developmental psychology* (M. Cole, Ed.), pp. 109, 111, New York: Sharpe. Reprinted by permission.

mina to conclude that the youngest age group did not understand the goal of remembering the items to buy. The major behavioral difference between 3- and 4-year-olds was that the latter group seemed to make an effort to remember. Regarding the two oldest age groups (5–6 and 6–7 years), Istomina assumed that these children already were aware of the mnemonic goal, and that they were capable of achieving that goal by selecting the strategy of rehearsal. However, although these children behaved similarly in the two experimental conditions, they performed less well in the lesson condition. It should be noted that no statistical tests of the data were provided.

ISTOMINA'S INTERPRETATION OF RESULTS

Istomina assumed that the consistent recall differences favoring the game over the lesson condition were due to differential motivational incentives for the children. Remembering presumably was an intrinsically important goal and had real meaning for the children in the game situation, especially in comparison to the lesson situation. Consequently, they probably showed greater interest in the task. In Istomina's view, differences in motivation could also account for the finding that even for the two oldest age groups better performance was obtained for the game condition as compared to the lesson condition, although attempts to deliberately memorize the items were similarly observed in both conditions. Her explanation was that

> the first awareness of remembering and recall as an explicit goal occurs much more readily if the sense of such a goal follows naturally from the motivating factor intrinsic in the overall situation . . . rather than lying in a more complicated and indirect relationship between goal and incentive, as was the case in the laboratory test experiment. (1975, p. 40)

For the following 40 years, this study has been cited widely as evidence that (a) deliberate remembering is not evident before 4 or 5 years of age, and (b)

preschoolers remember more in the context of play activities carried out in familiar settings where memory is not an explicitly stated goal than they do in situations involving deliberate retention (cf. Weissberg & Paris, 1986).

REPLICATIONS OF ISTOMINA'S EXPERIMENT

Before focusing on those three studies that we conceive of as close replications of Istomina's study, we wish to speculate about possible reasons it took so long before replications of Istomina's experiment were carried out. One obvious reason is that Istomina's study was not known to researchers from Western countries until its English translation was published in 1975. In our view, another important reason is that nobody seriously questioned Istomina's results because they were intuitively appealing. Neisser (1982), in his commentary regarding Istomina's experiment, gives a probably representative characterization in that he found the study to be both charming and convincing. Reviews of Soviet investigations of memory and cognition provided particularly in the seventies by Michael Cole, John Meacham, and James Wertsch (e.g., Cole, 1978; Meacham, 1977; Wertsch, 1981) lead to an increasing popularity of Soviet conceptions of cognitive development among American and European researchers. In particular, the reception of Soviet theories of cognitive development has stimulated new interest in issues like the influence of context on development (e.g., Ceci, Bronfenbrenner, & Baker, 1988), and the importance of ecological validity for our proper understanding of memory phenomena (cf. Perlmutter, 1988). Undoubtedly, the focus on issues like environmental context and ecological validity apparent in many recent studies on memory development has increased our knowledge about memory phenomena considerably (cf. for a review Schneider & Pressley, 1989). Accordingly, there seemed to be no reason to question Istomina's finding that young children's memory processes are influenced by contextual factors.

Given that Istomina's findings square very well with contemporary views of memory development, it seems surprising that replication studies were carried out at all. It is indeed difficult to explain why two research groups in the United States and Germany (Weissberg & Paris, 1986; Schneider & Brun, 1987) independently decided to replicate Istomina's study at about the same time. It was probably more or less by accident that both groups discovered possible methodological problems with Istomina's experiment that seemed serious enough to warrant a replication study.

Before discussing these problems in more detail, we want to make one important point clear: The goal of all replication studies reviewed in the remainder of this chapter was to explore the value of Istomina's findings for current memory research. In our view, it is impossible to replicate Istomina's

experiment in a strict sense, mainly because possible influences of cohort and cultural setting cannot be dealt with adequately in contemporary replication studies. For example, running an errand to a grocery store in order to buy food and other things for the kindergarten group probably had a different meaning for the Soviet children in 1948, than for American and German youngsters of the mid-1980s. Given that access to food was restricted shortly after World War Two (note that children were given "admission slips"), it appears that the shopping task was of great personal value and probably resulted in a feeling of responsibility in the Soviet children that could not be equally induced in American and German children participating in the replication experiments (cf. Folds-Bennett, chapter 10). Thus we should keep in mind that the outcomes of the various replication studies do not tell us much about whether Istomina was right or wrong; they only give information about the value of Istomina's findings for current memory research.

METHODOLOGICAL PROBLEMS WITH ISTOMINA'S STUDY

Both Weissberg and Paris (1986) and Schneider and Brun (1987) emphasized that Istomina's study contains several methodological problems. Although Istomina provided a very detailed description of her study, including a number of highly interesting analyses of individual cases, it is very difficult to reconstruct her study exactly because the procedures were specified vaguely. Weissberg and Paris (1986) inferred that the order of activities was not counterbalanced. This is probably right, since we know from the literature that the need for counterbalancing the order of activities was typically not realized in Soviet studies of memory development conducted in the 1940s and 1950s (cf. Schneider & Pressley, 1989, for a review). However, this cannot be decided on the basis of Istomina's description of the procedure, because the only information about this point is that the two experiments were carried out "simultaneously."

Another methodological problem was that different word lists were used in the two conditions, and that the word lists varied in unspecified ways between subjects. Further, a closer look at the reports of individual subjects clearly shows that the procedure was not strictly standardized. The experimenter occasionally repeated the names of the objects in the game condition but did not do so in the lesson condition. Thus one is inclined to assume that the generally superior recall found for the game condition can at least partly be explained by methodological flaws systematically biased against the lesson condition.

The results from a few experiments not closely replicating Istomina's study seemed to confirm this suspicion. By using a performance prediction paradigm, Wippich (1981) found that the game condition did not yield superior

recall in preschool children, as compared to the lesson condition. In a more recent investigation also based on the performance prediction paradigm, Arbinger and Kubsda (1987) basically replicated Wippich's findings. Further, Wippich (1985) contrasted laboratory with game conditions for various memory paradigms. In general, older children always performed better than younger children in these studies, but most critically, memory context never affected performance. Given these findings, it seemed to make sense to carry out close replications of Istomina's study, mainly because of the influence it has had on current claims concerning young children's memory.

REPLICATION STUDY 1: THE WEISSBERG AND PARIS (1986) EXPERIMENT

The primary goal of this study was to replicate Istomina's experiment and to correct the methodological shortcomings mentioned above. That is, the order of activities was counterbalanced, the same words were used for all children, and lists of items were counterbalanced across game and lesson conditions. A total of 96 children participated in the study. There were four age groups, each with 12 boys and 12 girls. Subjects ranged in age from 3 to 7 years.

Weissberg and Paris' (1986) study differed from the original experiment in that two different game settings and two lesson activities were used. Word lists were presented either as part of shopping or party games or as part of corresponding deliberate lessons. Thus, each child participated in four memory tasks—remembering food items or names as part of a game or a lesson activity—in a 4(age) x 2(sex) x 2(activity) x 2(mnemonic object) design. The experiment was well-controlled: The presentation orders of the four situations were counterbalanced within age groups, and the sequences of mnemonic objects were counterbalanced within activities for each group.

The procedure concerning the shopping situation was modeled closely after Istomina's description (see above, and Chapter 10). However, the scenario differed in several aspects. First, the social component of the original study was missing. That is, while the game was played in two adjoining rooms (a "kindergarten" and a "store") in the Istomina study, children in the Weissberg and Paris study were tested individually in one empty room. Children were instructed to buy a list of items from the play grocery store, but were not told that they should buy items for the kindergarten (as were the subjects in Istomina's study). Second, Paris and Weissberg used a list of six items (as compared to the five items used by Istomina), which was read twice in the case of Weissberg and Paris and once in Istomina's experiment.

In the "party" situation, children were shown seven stuffed animals in a circle. Children were asked to pick the animal they liked the most, which was usually given the subject's first name (e.g., Johnny). Children were then told

that the six remaining animals were having a party, and that Johnny would also like to be at that party. Children were then told that Johnny's problem was that he did not know the names of the other animals. They were then given the names of the animals at the party twice, and were asked to introduce the six animals to Johnny after 60–90 seconds had elapsed.

The lesson counterparts to both games required subjects to recall the same words after equivalent delays for the sole purpose of remembering. Instructions were closely modeled after those of Istomina's study.

As a major result concerning the 2-item lists, Weissberg and Paris reported that children recalled more food items than animal names. In the following, we focus on the food items which represent the replication part of the Weissberg and Paris study. Data from the replication study are shown in Table 9.2. As Weissberg and Paris pointed out, the different patterns of results are apparent immediately. While children in Istomina's experiment recalled more items in the game condition than in the lesson condition, the opposite was true in Weissberg and Paris' study in that recall turned out to be consistently better in the lesson condition. The sex-by-activity interaction was significant in the Weissberg and Paris experiment, indicating that while both boys and girls recalled more in the lesson activities than in the game, boys remembered more than girls in game activities and less in lesson activities.

A surprising finding of Weissberg and Paris' study was the large number of subjects who showed signs of rehearsal. Regarding the food items, more children (about 63%) rehearsed in the lesson condition than in the game condition (24%). Although older children rehearsed more often than the younger subjects, it seems remarkable that nearly half of the 3-year-olds rehearsed words spontaneously in the lesson activity. Weissberg and Paris concluded from their findings that children's recall does vary in different contexts, but that the underlying reasons may reflect task-relevant knowl-

Table 9.2. Comparison of the Weinberg and Paris data with Istomina's data: Mean numbers of words recalled by age and activity

	Istomina's Data				*Food*		
Age	*Lesson*	*Play*	*Ratio**	*Sex*	*Lesson*	*Play*	*Ratio*
3–4	.6	1.0	1.7	Male	3.0	2.4	.8
				Female	2.5	2.3	.9
4–5	1.5	3.0	2.0	Male	3.4	2.8	.8
				Female	3.5	2.7	.8
5–6	2.0	3.2	1.6	Male	3.9	3.5	.9
				Female	4.3	2.8	.7
6–7	2.3	3.8	1.7	Male	4.4	4.1	1.0
				Female	3.8	3.8	1.0
Mean total	1.6	2.8	1.8		3.7	3.1	.9

*The ratio is calculated by dividing the words recalled in play by the number recalled in the lesson condition.

edge and memory skills elicited in appropriate settings rather than global motivational dispositions.

REPLICATION STUDY 2: THE SCHNEIDER AND BRUN (1987) EXPERIMENTS

As noted earlier, Schneider and Brun shared Weissberg and Paris' view that methodological problems could be responsible for the context effect found in Istomina's study. In order to clarify this point, Schneider and Brun first tried to contact Z. M. Istomina directly at the Pedagogical Institute of Moscow University. Unfortunately, there was no reply. Consequently, Schneider and Brun decided to carry out research to replicate Istomina's experiment and to improve on her experimental methods. Although Weissberg and Paris were convinced that there was no counterbalancing in Istomina's study, Schneider and Brun were not equally confident, mainly because Istomina's description of the experimental design was so vague.

Thus they decided to first carry out an experiment which was aimed at reconstructing the experimental conditions presumably used by Istomina including the methodological shortcomings as they perceived them. The expectation was that it would be possible to replicate Istomina's findings under these circumstances. The second experiment was designed to correct shortcomings of the original procedure. That is, the order of play and lesson conditions was counterbalanced, lists of items were counterbalanced across the two experimental conditions, and precautions were taken to provide identical, standardized instructional settings in the two experimental conditions. Accordingly, the second experiment was thought to represent a more appropriate test of the context hypothesis.

Experiment 1. A total of 60 children participated in the experiment. Children were recruited from four kindergartens in the Munich area. There were only two age groups in this experiment, with 15 boys and 15 girls in each. The mean ages of the younger and older groups were about 4 years and 6 years, respectively.

The procedure was closely modeled after that used by Istomina and described above. However, there were deviations from Istomina's design similar to those already noted for the Weissberg and Paris study. That is, the social ("Play kindergarten") component was missing; all children were tested individually in one single room. Further, the two word lists included 8 items each, as compared to the 5-item word lists used by Istomina. Eight items were given in order to avoid ceiling effects; a pilot study with 6-year-olds had shown that shorter lists do not yield much variance in this age group.

In short, the results revealed that Schneider and Brun were able to replicate

Istomina's findings. On average, 6-year-olds recalled more than 4-year-olds, and memory performance in the game condition was significantly better than performance in the lesson condition for both age groups. No further main effects or interactions were found. The liberal instructional setting, that is, the possibility of returning to the experimenter and asking her to repeat the items, appeared to have different impact in the two experimental conditions: While in the lesson condition children did not find it necessary to ask for a repetition, the latter was more common in the game condition. Not surprisingly, those who asked the experimenter to repeat the items also recalled more than those who listened to the stimulus list only once. When initial recall (i.e., prior to any repetition by the experimenter) was used as the dependent variable, the effect of experimental condition was negligible. Schneider and Brun concluded from Experiment 1 that a major weakness of Istomina's experiment concerned the instructional setting: Better recall in the game condition may have been related to additional repetitions of the stimulus list in that condition.

Experiment 2. The only difference between Experiment 1 and Experiment 2 concerned the instructions given to the children. In both conditions of Experiment 2, children were told that they had to remember a list of items which the experimenter would read to them only once. In addition to the instruction given in Experiment 1, the experimenter explicitly informed the subjects that the word lists would not be repeated. Again, a total of 60 children was recruited. Half were about 4 years old, and half of them about 6 years of age. There were equal numbers of boys and girls in each group.

The results showed that only the main effect of age was significant. On average, 6-year-olds recalled more items than 4-year-olds. Although performance tended to be better in the game condition, there was no significant effect of activity. A significant interaction between order of activities and number of items recalled in the two experimental conditions indicated that, regardless of age, better recall was obtained for the experimental task that was presented first.

The data from Experiments 1 and 2 as well as Istomina's data are depicted in Table 9.3. It seems obvious that the strictly standardized instruction used in Experiment 2 affected memory performance particularly in the game condition. As a consequence, the game recall/lesson recall ratios in Experiment 2 were substantially lower than those obtained in Experiment 1 and in Istomina's study. Schneider and Brun (1987) concluded from this that Istomina's findings could not be replicated when stricter experimental controls were introduced. Thus the hypothesis that Istomina's findings were due in part to methodological shortcomings (which was already supported by the outcome of the Weissberg and Paris experiment) was also confirmed by the results of Schneider and Brun's second experiment.

Table 9.3. Comparison of the Schneider and Brun data with Istomina's data: Mean number of words recalled by age and activity

	Istomina's data			*Experiment 1*			*Experiment 2*		
Age	*Lesson*	*Game*	*Ratio*	*Lesson*	*Game*	*Ratio*	*Lesson*	*Game*	*Ratio*
4	1.05	2.00	1.8	1.63	3.56	2.2	1.59	2.08	1.3
6	2.15	3.50	1.7	3.06	5.46	1.8	2.53	2.79	1.1

*Data were collapsed across the 3–4- and 5-year-olds on the one hand and the 5–6- and 6–7-year-olds on the other hand.

A COMPARISON OF THE WEISSBERG AND PARIS (1986) AND SCHNEIDER AND BRUN (1987) EXPERIMENTS

At first glance, the experiments carried out by Weissberg and Paris (1986) and Schneider and Brun (1987) seem to convey the same message, namely, that Istomina's findings obtained in the 1940s are obscured by methodological flaws and at least cannot be generalized for children raised in the 1980s. However, a closer comparison of the two replication studies reveals that their results differed in several aspects. First, Weissberg and Paris found that their subjects generally recalled more items in the lesson condition, as compared to the game condition, whereas no significant difference between conditions was obtained by Schneider and Brun when stricter experimental controls were introduced in Experiment 2. A second difference concerned the frequency of rehearsal. In the Weissberg and Paris study, a surprising finding was the large number of children who rehearsed. On the other hand, only a few children in the Schneider and Brun study showed spontaneous verbal rehearsal. Possibly related to this, the children in the Weissberg and Paris study recalled considerably more than the children in the Schneider and Brun study, regardless of age and condition. This finding seems particularly surprising given that more items per set were used in the Schneider and Brun study.

How can we account for these differences? We can only speculate about the underlying causes. One possibility is that the repeated presentation of the word lists in the Weissberg and Paris study may have contributed to the high frequency of rehearsal observed in their sample. As those children who rehearsed words recalled significantly more words than those who did not rehearse, regardless of age, the overall higher recall observed in the Weissberg and Paris study could be due to the comparably higher incidence of rehearsal. If this holds true, the question remains why Weissberg and Paris's subjects rehearsed more in the lesson condition than they did in the game condition. In Weissberg and Paris's view, this finding can be attributed to the fact that there were more distractions involved in the game condition. Although this explanation seems to make sense, it does not account for

Schneider and Brun's finding that their subjects tended to recall more in the game condition than in the lesson condition.

Taken together, a systematic comparison of the two replication studies reveals a number of differences in findings that are very difficult to explain given the information at hand. While in both studies an attempt was made to replicate Istomina's design, a closer inspection of the procedural details shows that not only did the procedures and materials used in the two replication studies differ from each other but they also differed in several aspects from the original study. As noted above, the social component inherent in Istomina's study (i.e., Playing Kindergarten) was not included in the replications. Although we do not think that the failure to replicate Istomina's results is due to this aspect—in fact it could have increased the number of distractions involved in the game condition—it is certainly unfortunate that the implications of this aspect were not assessed. We are thus left with a dilemma: Although the results from the two replication studies indicate that several methodological problems could have biased Istomina's findings, the extent to which deviations from the original procedures and materials affected the outcomes of the replication studies is impossible to assess.

Given these problems, we decided to conduct another replication study just to explore the significance of other variables that had not been systematically controlled in the two previous replications.

REPLICATION STUDY 3: HASSELHORN AND SCHNEIDER (1990)

A closer inspection of Schneider and Brun's (1987) data reveals that reproduction probabilities for single items varied considerably in both experiments. We were led to the conclusion that this extreme variability could be due to effects of attractiveness or salience of the item materials. Apparently, children remembered the more attractive objects much better. We decided to carry out another replication experiment to explore more systematically the effects of attractiveness of item materials on recall. The question of specific interest was whether Schneider and Brun's (1987; Experiment 2) results (i.e., the failure to replicate Istomina's findings) could be replicated for both attractive and unattractive stimulus materials.

Experiment 1. To assess the attractiveness of item materials, a pilot study was carried out with 40 4- and 6-year-old children, with 20 subjects in each age group. Children were presented with a total of 40 photographs showing various objects. According to our judgment, 20 of these objects were probably attractive for most subjects, whereas the remaining objects were perceived of as rather unattractive. Children were given four sets of 10 items. Each set included five items rated as attractive and five items rated as

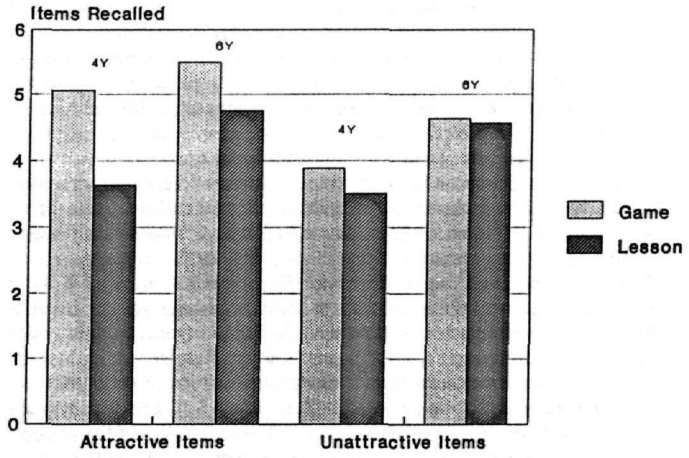
unattractive. The subjects were told to first choose those three items (out of 10) that they considered most attractive ("show me the three pictures that you like most"), and then to select those three items out of the remaining seven that they found least attractive ("show me the three pictures that you like least"). Attractiveness judgments were analyzed separately for the two age groups. The list of unattractive items chosen for Experiment 1 consisted of 8 items that were similarly classified as unattractive by 4- and 6-year-olds. The following items were considered unattractive by most children: toothpaste, vase, handkerchiefs, rock, screw driver, bulb, sponge, and cigarettes (Zahnpasta, Vase, Taschentuch, Stein, Schraubenzieher, Glühbirne, Schwamm, Zigaretten).

The list of attractive items chosen for Experiment 1 included 8 items selected as attractive by the majority of children in both age groups. The following objects were rated as attractive: Toy car, crayons, balloon, teddy bear, chocolate egg, engine, licorice, and sailboat (Spielzeugauto, Malstift, Luftballon, Teddybär, Schokoladenei, Eisenbahn, Lakritze, Segelboot).

A total of 128 children participated in Experiment 1. There were two age groups of children, with 64 children in each. The mean ages of the younger and older groups were about 4 and 6 years, respectively. While there were equal numbers of boys and girls in the younger group, the older group consisted of 24 boys and 40 girls. Children were recruited from kindergartens in the Göttingen West Germany area. With each age group, subjects were randomly assigned to the two activity (game vs. lesson) and attractiveness conditions. The procedure of this experiment differed from that used by Schneider and Brun in that items were not only read to the children but also presented visually at the same time. Another difference concerned the fact that items were given twice in this experiment but only once in the Schneider and Brun study.

Preliminary analyses revealed that there were no effects of sex. Thus data were collapsed across the sex variable. Figure 9.1 shows the mean number of items recalled as a function of age, item attractiveness, and activity. Analyses of variance revealed a significant effect of age ($F(1,120) = 22.2$, $p < .001$). Six-year-olds recalled more items than 4-year-olds. There were also main effects of attractiveness ($F(1,120) = 11.0$, $p < .001$) and activity ($F(1,120) = 13.4$, $p < .001$). However, these effects were qualified by a significant activity by attractiveness interaction ($F(1,120) = 6.0$, $p < .05$). Post-hoc comparisons revealed that better recall in the game condition was restricted to the attractive item materials, regardless of age.

In addition to children's recall, signs of rehearsal were also recorded as a simple dichotomous variable. Children who either showed lip movements or verbally repeated the words were classified as active rehearsers. Based on these selection criteria, about 47 percent of the older children and about 41 percent of the younger children were classified as active rehearsers. There



Note: This experiment used items per set

Figure 9.1. Mean number of items recalled as a function of age, item attractiveness, and activity (data from Hasselhorn & Schneider, 1990; Experiment 1).

were slightly more boys than girls among active rehearsers (50% vs. 38%), and less children rehearsed in the lesson condition than in the game condition. However, these differences did not prove statistically significant. The only effect approaching significance was observed for item attractiveness ($F(1,120) = 3.2$, $p < .10$): Although about 52 % of the children presented with attractive items actively rehearsed the stimuli, only 36 percent of the children presented with unattractive items did so. Point-biserial correlations between rehearsal and recall were nonsignificant for both younger and older children (r's = .04 and − .06, respectively), indicating that the systematic relationship between rehearsal and recall reported by Weissberg and Paris was not found in this experiment.

Taken together, the results of this experiment seem to indicate that the context effect in favor of the game condition reported by Istomina can be replicated even after methodological problems of Istomina's study have been removed. However, the effect may not be as general as originally assumed, that is, restricted to learning materials that seem particularly attractive for young children. Thus the results of Experiment 1 suggest that it does not make much sense to talk about general effects of *the* context because different features of the situational context may turn out to be relevant in memory studies with young children. It seems that not only *objective* environmental features of the learning situation (game vs. lesson) but also *subjec*

tive features like the perceived attractiveness of stimulus materials contribute to this effect. Apparently, different context features interact in influencing young children's memory performance.

If this is so, a reasonable next step seems to explore other features of situational context possibly relevant for memory performance. In order to find out about additional context features, we decided to interview our experimenters of Experiment 1, asking them to reconstruct the procedural details as closely as possible. Further, we analyzed some tape recordings of experimental sessions which originally were produced as instructional materials in developmental psychology classes at the University of Göttingen. Both the discussions with the experimenters and the inspections of tape recordings confirmed our suspicion that the game and lesson conditions differed in several ways. In addition to the environmental context features, the experimenters' behavior also varied across activities. It appeared that the game condition seduced the experimenters to communicate in a more friendly, motivating, and empathic way with the children than they did in the more sterile lesson condition.

There were several signs supporting the assumption that the type of activity influenced the experimenters' behavior. First, it seemed that the experimenters established a more *personal contact* with the subjects in the game condition, as compared to the lesson condition. For example, in the lesson condition the experimenter would start with instruction such as "Let's play a game together" and then explain the procedural details to the child. In the game condition, the experimenter would typically begin with "Let's play doing an errand in a grocery store. You probably have done that before with your mother, haven't you?" In most cases, this question led to a short conversation between child and experimenter, with the child informing the experimenter about his or her experiences in a shopping situation. Such conversations did not take place in the lesson condition.

Further, the tape recordings revealed that the pitch of the experimenters' voices differed between the two conditions. That is, the experimenter's voice tended to be lower and more monotone in the lesson condition, as compared to the game condition. Moreover, there seemed to be systematic differences regarding the way the experimenter verbally reinforced children's responses in the two conditions. These differences were not quantitative but qualitative in nature. For example, in the lesson condition, the experimenter's usual comment on correct responses was "right" (*ja*), as compared to a "great" (*sehr gut*) following a correct response in the game condition. Finally, our interview with the experimenters revealed that they tended to nonverbally encourage children to persist and to try recall of items (e.g., via gestures) in the game condition but not in the lesson condition.

Experiment 2. Given these specific problems, we decided to conduct another replication experiment. The major goal of this experiment was to

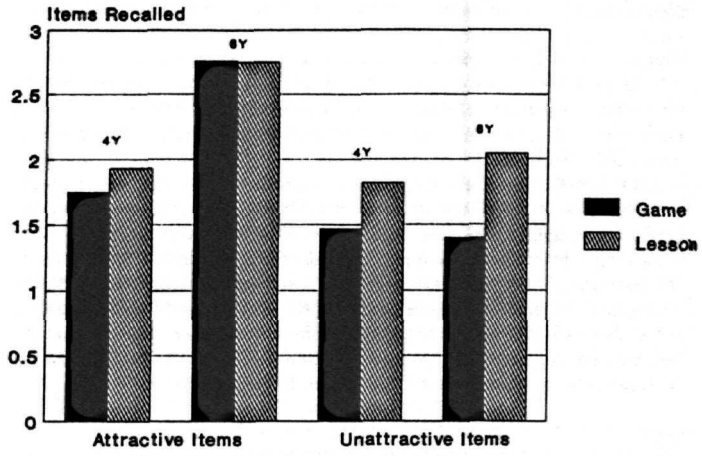
explore whether the context effect found in Experiment 1 with attractive stimulus materials could be replicated with experimenters being informed about the possible influences of environmental context on their behavior in the experimental session. A second aim of Experiment 2 was to explore the impact of item presentation mode (visual vs. acoustic) on memory performance.

Experiment 2 differed from Experiment 1 in two ways. First, the experiments were informed about the influence of the game situation on their behavior and were instructed to behave similarly in the two activity conditions. Second, item attractiveness was varied *intraindividually* and was based on individual attractiveness judgments.

A total of 112 4- and 6-year-old children participated in Experiment 2, with 56 subjects in each age group. There were the same number of boys and girls in each group. Subjects were recruited from 5 kindergartens in the Göttingen and Munich areas. Children were presented with a list of 10 items (5 attractive and 5 unattractive items). In the visual presentation mode condition, children were shown the photographs of the to-be-recalled objects, whereas items were read to the children in the acoustic presentation mode condition.

Again, preliminary analyses revealed no significant effects of sex on recall. Thus, recall data were collapsed across this variable and analyzed separately for attractive and unattractive items (cf. Figure 9.2). Accordingly, a two (age groups) by two (presentation modes) by two (activities) by two (attractiveness levels) design was used, with the last factor repeated within subjects. With the exception of presentation mode, all main effects were significant. Analyses of variance revealed a main effect of activity ($F(1,104) = 7.7, p < .01$), indicating that children recalled more in the lesson condition than in the game condition. The main effects of age ($F(1,104) = 19.3, p < .001$) and item attractiveness ($F(1,104) = 26.8, p < .001$) were qualified by a significant age by attractiveness interaction ($F(1,104) = 12.3, p < .01$). Post-hoc comparisons revealed that the age effect was restricted to attractive items, and that the attractiveness effect held only for the 6-year-olds. Undoubtedly, the most interesting finding of Experiment 2 concerned the *reversed* context effect: It was not in the game condition but in the lesson condition that children recalled more objects, regardless of age. However, this context effect was qualified by a significant three-way interaction between activity, age, and presentation mode ($F(1,104) = 6.0, p < .05$), indicating that older children's better recall in the lesson setting was restricted to the unattractive items. Thus the results of Experiment 2 replicated those obtained by Weissberg and Paris (1986).

As in Experiment 1 children were classified as active rehearsers versus nonrehearsers. While about 48% of the total sample showed signs of rehearsal, a two (age groups) by two (presentation modes) by two (activities) by two (sexes) analysis of variance only revealed a marginal age by sex interaction, indicating that there were more boys among the older rehearsers.



Note: This experiment used 5 items per set

Figure 9.2. Mean number of items recalled as a function of age, item attractiveness, and activity (data from Hasselhorn & Schneider, 1990; Experiment 2).

How can we account for the findings concerning the recall data? As noted earlier, one explanation offered by Weissberg and Paris is that the mnemonic goal of the task can be obscured by several task-irrelevant aspects of the game condition. Our observations of children's behavior in the game condition support such an assumption. For example it appeared that several children did not pay as much attention to the items read or shown to them than to details of the play grocery store ("What kind of stuff do you have in the grocery store?"). Similarly, signs of distractions became also apparent in the retrieval phase of the game condition ("I think some of the things I should buy were not at the grocery store"). Given the limited attentional capacity of these young children, it does not seem surprising that their recall was consistently low in the game condition.

How can we account for the fact that a (restricted) context effect was found in Experiment 1? We think that the problems related to the confounding noted above produced the findings to a considerable extent. Apparently, the experimenters' extremely friendly, motivating, and supportive behavior in the game condition helped to establish the mnemonic goal and more than compensated for the distractions, at least in the case of the attractive items. On the other hand, the findings of Experiment 2 indicate that the more neutral behavior shown by the experimenters had a different effect: Children seemed to distribute their cognitive resources across several goals and activ-

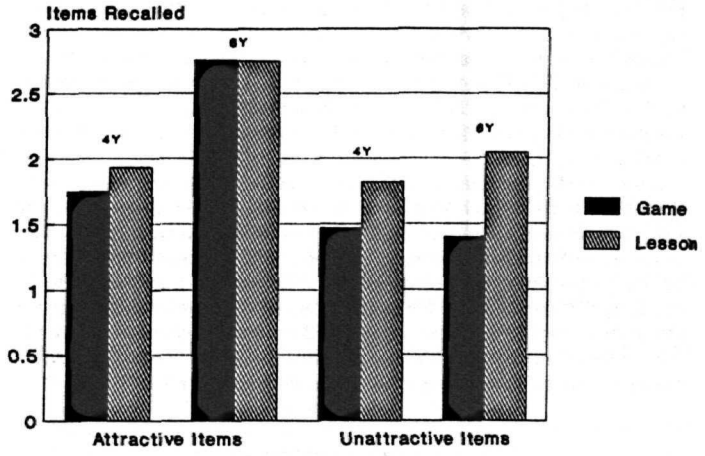
explore whether the context effect found in Experiment 1 with attractive stimulus materials could be replicated with experimenters being informed about the possible influences of environmental context on their behavior in the experimental session. A second aim of Experiment 2 was to explore the impact of item presentation mode (visual vs. acoustic) on memory performance.

Experiment 2 differed from Experiment 1 in two ways. First, the experiments were informed about the influence of the game situation on their behavior and were instructed to behave similarly in the two activity conditions. Second, item attractiveness was varied *intraindividually* and was based on individual attractiveness judgments.

A total of 112 4- and 6-year-old children participated in Experiment 2, with 56 subjects in each age group. There were the same number of boys and girls in each group. Subjects were recruited from 5 kindergartens in the Göttingen and Munich areas. Children were presented with a list of 10 items (5 attractive and 5 unattractive items). In the visual presentation mode condition, children were shown the photographs of the to-be-recalled objects, whereas items were read to the children in the acoustic presentation mode condition.

Again, preliminary analyses revealed no significant effects of sex on recall. Thus, recall data were collapsed across this variable and analyzed separately for attractive and unattractive items (cf. Figure 9.2). Accordingly, a two (age groups) by two (presentation modes) by two (activities) by two (attractiveness levels) design was used, with the last factor repeated within subjects. With the exception of presentation mode, all main effects were significant. Analyses of variance revealed a main effect of activity ($F(1,104) = 7.7, p < .01$), indicating that children recalled more in the lesson condition than in the game condition. The main effects of age ($F(1,104) = 19.3, p < .001$) and item attractiveness ($F(1,104) = 26.8, p < .001$) were qualified by a significant age by attractiveness interaction ($F(1,104) = 12.3, p < .01$). Post-hoc comparisons revealed that the age effect was restricted to attractive items, and that the attractiveness effect held only for the 6-year-olds. Undoubtedly, the most interesting finding of Experiment 2 concerned the *reversed* context effect: It was not in the game condition but in the lesson condition that children recalled more objects, regardless of age. However, this context effect was qualified by a significant three-way interaction between activity, age, and presentation mode ($F(1,104) = 6.0, p < .05$), indicating that older children's better recall in the lesson setting was restricted to the unattractive items. Thus the results of Experiment 2 replicated those obtained by Weissberg and Paris (1986).

As in Experiment 1 children were classified as active rehearsers versus nonrehearsers. While about 48% of the total sample showed signs of rehearsal, a two (age groups) by two (presentation modes) by two (activities) by two (sexes) analysis of variance only revealed a marginal age by sex interaction, indicating that there were more boys among the older rehearsers.



Note: This experiment used 5 items per set

Figure 9.2. Mean number of items recalled as a function of age, item attractiveness, and activity (data from Hasselhorn & Schneider, 1990; Experiment 2).

How can we account for the findings concerning the recall data? As noted earlier, one explanation offered by Weissberg and Paris is that the mnemonic goal of the task can be obscured by several task-irrelevant aspects of the game condition. Our observations of children's behavior in the game condition support such an assumption. For example it appeared that several children did not pay as much attention to the items read or shown to them than to details of the play grocery store ("What kind of stuff do you have in the grocery store?"). Similarly, signs of distractions became also apparent in the retrieval phase of the game condition ("I think some of the things I should buy were not at the grocery store"). Given the limited attentional capacity of these young children, it does not seem surprising that their recall was consistently low in the game condition.

How can we account for the fact that a (restricted) context effect was found in Experiment 1? We think that the problems related to the confounding noted above produced the findings to a considerable extent. Apparently, the experimenters' extremely friendly, motivating, and supportive behavior in the game condition helped to establish the mnemonic goal and more than compensated for the distractions, at least in the case of the attractive items. On the other hand, the findings of Experiment 2 indicate that the more neutral behavior shown by the experimenters had a different effect: Children seemed to distribute their cognitive resources across several goals and activ-

ities, with only one of them being the mnemonic goal implemented by the experimenter. In comparison, children have no problem using most of their cognitive capacity to recall items in the lesson condition, mainly because no secondary goals are available in this situation. Thus, we assume that the information processing capacity interpretation suggested by Weissberg and Paris is the most parsimonious in explaining the reversed context effects found in Experiment 2.

Alternatively, one could argue that it is not the *difference between* conditions but the *structure within* each of them that either gives or not the expective sample differences. So, by way of altering instruction style and triggering individual attractiveness to given objects in children, one can make the lesson condition more effective than the game condition. Likewise, by changing the friendliness of the game partner and giving different objects to remember, one can make the game condition ineffective. All these experimental variations concern the second structure, making the comparisons between settings quite unpredictable from study to study.[1]

GENERAL DISCUSSION

At the very beginning of this chapter we discussed several problems replication studies have to deal with. We noted that experimenters failing to replicate a classic study almost certainly will experience difficulties with convincing the scientific community that the old study was indeed wrong. We think that the description of the three replications of Istomina's study experiment given in this chapter is well suited to illustrate the various problems that can be experienced in such a case.

To summarize, it appeared that the first two replication studies (Weissberg & Paris, 1986; Schneider & Brun, 1987) built a strong case against Istomina's original findings in that both studies were not able to replicate Istomina's results. This was all the more impressive as both studies improved Istomina's methodology and were conducted in different cultural contexts. However, a closer look at the outcomes of the two replication studies revealed that their findings differed in several aspects. Probably most important, Weissberg and Paris found a reversed context effect, indicating that American youngsters performed better in the lesson condition than they did in the game condition, whereas Schneider and Brun's German children still tended to be better (although not significantly) in the game condition than in the lesson condition after strict methodological controls were established. Another major difference between the two replications was that most children in the Weissberg and Paris study used verbal rehearsal, particularly in the lesson condition,

[1]This is the interpretation preferred by Jaan Valsiner. We are grateful for his suggestion.

which helped in improving recall, whereas Schneider and Brun's subjects rarely used this mnemonic strategy.

The third replication study (Hasselhorn & Schneider, 1990) was carried out in order to explore the importance of additional context features not systematically considered in the two previous replications. The first experiment investigated the relevance of item attractiveness for memory performance, yielding the interesting as well as puzzling finding that Istomina's results could be replicated in the case of particularly attractive item materials. However, a closer inspection of procedural details revealed that not physical content per se, but different behavior of the experimenter in the two activity conditions seems to be responsible for the results. This suspicion was confirmed in a subsequent replication study that took care of the experimenter problem: In this experiment, the reversed context effect of Weissberg and Paris could be replicated, whereas the effect of item attractiveness was found to be restricted to the older children, positively affecting performance in both activity conditions.

What did we learn from all these replications?

First of all, it seems that even minimal differences in the experimental design and/or procedure can have dramatic effects on experimental outcomes. Our description of the replication studies revealed that they were not replications *in the strict sense*, but *conceptual* replications (see Finifter, 1975, for a more detailed discussion of this terminological distinction). That is, the major goal of the studies was to replicate Istomina's findings but not to reconstruct the design of the original study in detail. For example, we already referred to the fact that the social component ("Playing kindergarten") of Istomina's experiment was not included in the replications. This is also true for Experiment 1 of Schneider and Brun (1987), which came closest to a replication in the strict sense.

Why has no one tried to carry out a replication in the strict sense? There are several reasons. First, the information concerning procedural details given in Istomina's study are too vague to allow for a replication which matches every detail of the original study. Schneider and Brun's attempts to get more detailed information from Istomina were unsuccessful. Thus it appears that the ideal replication in the strict sense is impossible in the case of Istomina's study. But even if it were possible to exactly reconstruct Istomina's design, the inclusion of the social component would probably cause considerable problems. That is, the inclusion of a "Playing Kindergarten" setting would almost certainly add to the methodological problems because it is probably impossible to carry out a standardized experimental procedure using such a setting. Thus our conclusion is that a replication study in the strict sense is neither possible nor does it make much sense in the case of Istomina's experiment. We think that there is sufficient empirical evidence demonstrating the impact of methodological flaws on the outcomes of Istomina's study. Therefore, it

seems much more important to explore the possibility of conceptual replications.

By and large, the conceptual replications reviewed in this chapter suggest the conclusion that Istomina's findings cannot be generalized to current memory research. That is, those replication experiments which methodologically improved over Istomina's experimental design did not find any evidence for the context effect reported by Istomina. The context effect could only be demonstrated in those two experiments that either intentionally included the methodological flaws observed in Istomina's study (Schneider & Brun, Experiment 1) or that were shown to (unintentionally) include additional methodological problems at least from a strict experimental point of view (Hasselhorn & Schneider, 1990, Experiment 1). The results of these two experiments suggest that the induction of particularly motivating experimental conditions in the game situation (e.g., a strong personal commitment of the experimenters) was responsible for the effect in favor of the game context. These findings are due to the fact that the experimenters behaved differently in the two conditions, and must be interpreted as artifacts from a methodological point of view.

As indicated by the results of the Weissberg and Paris (1986) and Hasselhorn and Schneider (1990, Experiment 2) replications, it is more reasonable to expect a reversed context effect in strictly controlled experiments conducted in the eighties. It seems that there are at least two possible explanations for the fact that better memory can be observed in the lesson condition, as compared to the game condition. First, there is no doubt that the educational environments available to American and German children in the 1980s clearly differ from those typically experienced by Soviet children in the 1940s. Probably due to effects of preschool education, television, and parental influences, even 3- to 4-year-olds are capable of voluntary memory (cf. Baker-Ward, Ornstein, & Holden, 1984; Sodian, Schneider, & Perlmutter, 1986). American and German preschoolers are no longer bewildered by requests to remember lists of words. As indicated by the findings of Weissberg and Paris, they may even know how to use mnemonic strategies efficiently at a very early age.

The second explanation offered by Weissberg and Paris concerns the negative impact of task-irrelevant aspects of the game. That is, several distractors embedded in the game situation (e.g., attractive objects in the grocery store) may obscure the mnemonic goal of the task and instead generate new goals (like playing with those objects) that are detrimental to—if not incompatible with—the mnemonic goal implemented by the experimenter. Our observations (cf. Hasselhorn & Schneider, 1990, Experiment 2) do confirm the conclusion that young children can make more use of their limited information processing capacity in the lesson condition than in the game condition, mainly because there are no secondary goals competing in the lesson condition.

All in all, the conceptual replications of Istomina's study clearly show that young children are better at remembering lists of words in sterile lesson conditions than in more stimulating game conditions, provided that adequate methodological controls are implemented in the experiment. Does this imply that more sterile and—at least to a certain extent—boring learning situations guarantee better memory performance in young children? Is this the message to convey to contemporary memory researchers? We actually do not think so. Let's play the devil's advocate for a moment. We think that the lesson we learned from Istomina's experiment and those replications which corrected its methodological flaws is that the game situation can have enormous motivational effects which in turn seem to facilitate remembering in young children. Apparently, the measures taken to eliminate the experimenter's behavioral differences in the two experimental conditions caused these motivational impacts to disappear. Please note, for example, that in the Hasselhorn and Schneider (Experiment 2) study we did not ask the experimenters to act as enthusiastically in the lesson condition as they already did in the game condition, but instead asked them to slow down and act more neutral in the game condition than they did before. The reason is that we do not know how to elicit the same kind of motivation and a similarly stimulating atmosphere in the lesson condition than is possible in the game condition. We thus cannot exclude the possibility that social context may have a much more tremendous impact on young children's memory than suggested by the replication studies. However, we are left with the dilemma that it is probably impossible to test such an assumption without violating methodological standards typically referred to as internal validity.

In this regard, we would like to note that our inspection of the various replication studies revealed that it is important to distinguish among different groups of context features that may influence remembering. In addition to the environmental features of the context (e.g., game vs. lesson) typically referred to by most researchers discussing the issue of context, it makes sense to include at least two other features, namely features of the learning material (e.g., item attractiveness), and characteristics of the experimenter's behavior as situational features. There is little doubt that these features may influence the outcomes of experimental studies exploring the impact of context in different ways.

In our view, the detailed description of the three replications of Istomina's experiment has revealed that attempts to replicate and validate experimental findings are indeed very important. Given the lack of consistency in findings concerning the effect of context on children's memory, we are afraid that similar problems could arise in replication attempts focusing on other classic developmental studies. Researchers thus are encouraged to try to replicate findings from experimental research. It probably needs several independent replication attempts to get a feeling for the true state-of-the-art concerning

the conceptual issue at stake and to distinguish between obviously solid and more shaky effects. In our case, for example, one of the most solid findings was that the age by activity interaction postulated by Istomina was not obtained in any of the replication experiments. That is, the expectation that recall differences between the game and lesson conditions should be larger for the 4-year-olds than for the 6-year-olds could not be confirmed. Thus the original assumption that a major qualitative change in memory occurs between the ages of 4 and 6 has not been validated by subsequent research.

We think that this is just one illustrative example emphasizing the need for as well as the utility of replication studies in experimental research. We do hope that more researchers and editors of scientific journals will accept the importance of this issue and get more interested in replication studies than they were before. There is no doubt for us that such a revised policy will not cause a delay in scientific progress but instead will not only make possible a faster scientific development but also will add considerably to our knowledge (and not to our prejudices) in various domains.

*We would like to thank Beth Kurtz for her comments on an earlier version of this chapter.

REFERENCES

Arbinger, R., & Kubsda, J. (1987). Situativer Kontext und Verarbeitungstiefe als Bedingungen für Behaltensleistungen im Vorschulalter. *Psychologie in Erziehung und Unterricht, 34*, 196–204.

Baker-Ward, L., Ornstein, P. A., & Holden, D. J. (1984). The expression of memorization in early childhood. *Journal of Experimental Child Psychology, 37*, 555–575.

Ceci, S. J., Bronfenbrenner, U., & Baker, J. (1988). Memory development and ecological complexity: The case of prospective remembering. In F. E. Weinert & M. Perlmutter (Eds.), *Memory Development: Universal Changes and Individual Differences* (pp. 243–256). Hillsdale, NJ: Lawrence Erlbaum.

Cole, M. (Ed.). (1978). *The Selected Writings of A. R. Luria*. New York: Sharpe.

Finifter, B. M. (1975). Replication and extension of social research through secondary analysis. *Social Science Information, 14*, 119–153.

Hasselhorn, M., & Schneider, W. (1990). *External and internal mediators of young children's memory performance*. Unpublished paper. Max Planck Institute for Psychological Research, Munich.

Istomina, Z. M. (1975). The development of voluntary memory in preschool-age children. *Soviet Psychology, 13*, 5–64.

Istomina, Z. M. (1977). The development of voluntary memory in preschool-age children. In M. Cole (Ed.), *Soviet developmental psychology* (pp. 100–159). New York: Sharpe.

Leontjev, A. N. (1977). *Probleme der Entwicklung des Psychischen*. Kronberg, Germany: Athenäum.

Meacham, J. A. (1977). Soviet investigations of memory development. In R. V. Kail & J. W. Hagen (Eds.), *Perspectives on the Development of Memory and Cognition* (pp. 273–295). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Neisser, U. (1982). *Memory observed—Remembering in Natural Contexts*. San Francisco: Freeman.

Perlmutter, M. (1988). Research on memory and its development: Past, present, and future. In F. E. Weinert & M. Perlmutter (Eds.), *Memory Development: Universal Changes and Individual Differences*. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Schneider, W., & Brun, H. (1987). The role of context in young children's memory performance: Istomina revisited. *British Journal of Developmental Psychology, 5*, 333–341.

Schneider, W., & Pressley, M. (1989). *Memory Development between 2 and 20*. New York: Springer-Verlag.

Sodian, B., Schneider, W., & Perlmutter, M. (1986). Recall, clustering, and metamemory in young children. *Journal of Experimental Child Psychology, 41*, 395–410.

Weissberg, J. A., & Paris, S. G. (1986). Young children's remembering in different contexts: A reinterpretation of Istomina's study. *Child Development, 57*, 1123–1129.

Wertsch, J. V. (1981). *The Concept of Activity in Soviet Psychology*. New York: Sharpe.

Wippich, W. (1981). Verbessert eine Einkaufssituation die Vorhersage der eigenen Behaltensleistungen im Vorschulalter? *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, 8*, 280–290.

Wippich, W. (1985). Untersuchungen zur Entwicklung des Meta-Gedächtnisses in verschiedenen Kontexten. In L. Montada (Ed.), *Bericht über die 7. Tagung Entwicklungspsychologie*. Trier, Germany: Universitätsdruck.

CHAPTER 10

Replication: What's The Angle?

Trisha H. Folds-Bennett
Department of Psychology
College of Charleston
Charleston, Sc

INTRODUCTION

Traditionally, a primary aim in the study of psychology has been to produce results that are repeatable given the same research methods. This goal is critical to psychological researchers because one of their central objectives is to understand events that are internal to the individual. Obviously, one cannot directly observe an internal event. Therefore, it is incumbent on the psychological researcher to devise methods of study that are thought to reflect internal activity and then to endeavor to validate these methods. An important step in the validation process is to show that results achieved with a certain method can be reproduced on a new, but similar, sample of individuals. It is typically assumed that the replication carried out must be an exact or algorithmic replication (see Chapter 3), involving a step-by-step precise reproduction of the original method. If, under these circumstances, it is shown that a set of results is not repeatable, then it would be fair to conclude that the original results were unreliable. The lack of repeatability would suggest that the original set of results were limited to a specific sample, in a specific place, and at a specific time, and that the results cannot be generalized to the extent of providing any theoretical insight. This point is critical because a psychological researcher's most explicit goal is to increase understanding about specified phenomena. Given this goal, therefore, it is absolutely necessary that the researcher employ methods that will reliably test hypotheses that he or she has made.

Unfortunately, psychological researchers tend to pay more lip service to the importance of replication than they do to careful execution of replication

studies. This paucity of replications means that the field of psychology often relies on findings that are perhaps less reliable than they are assumed to be, making the understanding of certain psychological phenomena highly method- and sample-bound. One reason for the low number of replications is professional; if a hypothesis has been supported and published using a certain method, the likelihood of publishing an exact replication of the original study is fairly low. Because the professional pay-off for exact replication is decreased, the importance of conducting the replication is obscured. This perceived nonchalant attitude about replication is especially discouraging to young researchers who are still trying to establish a professional reputation. Indeed, they are often persuaded instead to conduct original work rather than to simply imitate an already-published method, even if it is their own.

Another reason for the lack of pure methodological replication (in the sense that an original method is applied directly and exactly) is a pragmatic one; rarely does a researcher document his or her methods to the extent that they can be carried out to the letter by another person. Even if two people who have worked in the same lab and who have used the same methods attempt to produce a replication of some finding, it is nearly impossible to duplicate the original setting. Different laboratory settings, in different locations, with different equipment, different experimenters, and different data coders all can lead to variations in the results of seemingly identical experiments.

Because of the low pay-off for and difficulty in executing exact replication studies, it is often the case that replication studies are carried out in a more indirect manner. Rather than applying the exact methods used by the original author(s), methodological flaws are identified up front and improvements are made when the "replication" is carried out. While this process has higher payoffs both personally and scientifically, if the methods of the original experimenter are not applied directly, it is difficult to compare the original study and the follow-up. If the follow-up study contradicts the original, the tendency is to call into question the original findings. However, because the methods are somewhat different, it is impossible to conclude outright that an already accepted finding has been refuted. Discrepancies between an original set of findings and an attempted replication of the findings could be due to methodological differences in the two studies, or they could speak to more global theoretical issues. As readers of such work, one is often left wondering how the improvements of the replication authors might have changed the whole nature of the experiment. It is nearly impossible in such a case to reach conclusions regarding the methodological implications versus the theoretical implications of a set of findings.

In this chapter, I argue that we need to think through our goals more thoroughly before we pursue replication. If one is more interested in methodological purity, then it obviously makes sense to focus on methodological

flaws of an original work. If this is the stated goal of a replication, then the authors should be cautious about the extent to which they comment on the theoretical implications of their work, especially if their findings contradict the original results. However, if one is more interested in clarifying underlying theoretical assumptions, then it is imperative to give more careful consideration to and analysis of the rationale for designing a study in a particular manner.

It is very easy to slip into an "outcome-oriented stance" when attempting to test a certain set of assumptions. In this case, the results of an original study become more important than the processes underlying the results. I argue in this chapter that there are circumstances when a researcher can produce a more theoretically valid replication by *not* directly applying the previously used method to a new set of subjects. I illustrate what I mean about the importance of considering the underlying theoretical assumption by using the original study by Istomina (described by Valsiner in his introductory remarks), and comparing it to recent replication attempts (see Chapter 9; Schneider & Brun, 1987; Weissberg & Paris, 1986). I argue that often what we may consider replication in the sense that identical methods are used may not be replication in the broader, theoretical sense.

REPLICATION: A CASE STUDY

The replications of Istomina that have been carried out to date have tended toward the "indirect" approach that I described previously. In one study, conducted by Weissberg and Paris (1986), methodological flaws were identified and corrected, resulting in findings contrary to the originally published findings. In a series of replication attempts by Schneider and colleagues (Schneider & Brun, 1987; Chapter 9), similar and additional flaws were identified and improved, resulting in some contradiction of and some support for Istomina's original results.

What should be made of the Istomina replications or lack thereof? The replication authors present their work as reliably calling into question Istomina's original findings. But is this conclusion fair? If we accept the replications and the conclusions that the replication authors reached, are we not likely to doubt the validity of Istomina's underlying theoretical assumptions? In order to better answer these questions we should first explore the theory on which Istomina's study was based and then consider differences and similarities between the methods and results of her study and its replications. A careful analysis of the discrepancies between the studies should shed light on the difficulties of replication and will, perhaps, suggest more conceptually sound replication procedures.

Theoretical Background

Soviet theories of activity. The purpose of Istomina's original study, conducted in 1948, was to provide an illustration of the concepts of activity, prevalent in the work of Soviet psychologists (e.g., Gal'perin, 1968, Istomina, 1975, 1977; Leontiev, 1932, 1978; Zaporozhets, 1977, 1979–80; Zinchenko, 1983–1984). The different concepts of activity are focused primarily on the relationship between the person and his/her environment and specific behaviors that link the two (Wertsch, 1981). The process and result of this person-environment interaction is conceptualized basically as a feedback loop. That is, a person performs a certain action or set of actions within a specific social context. The impact of that action yields some sort of feedback from the environment. The individual responds to this feedback, either actively or passively, and integrates the new information into his or her action system. Activity, specifically, is described in terms of motives that guide an individual to behave in a certain manner and in terms of the object towards which a behavior is directed. Therefore, activity is defined as functional and goal-directed (Leontiev, 1978).

In theories of activity, the concern is not so much with the product of action within a certain context, but instead with the process. In this sense, theories of activity are inherently developmental. That is, the focus is on a microgenetic analysis of how interactions between person and environment lead to certain actions (Werner, 1937), rather than on the final results of the interaction. This point is important when considering empirical support for a theory of activity because often experimental situations are constructed with a certain outcome in mind, rather than reflecting explicit consideration of any underlying processes.

Concept of activity in the context of memory development research. In Istomina's work, the concept of activity is embedded in a focus on memory development. This focus is derived from the work of such Soviet psychologists as Leontiev (1932, 1978) and Zinchenko (1983/84), who described memory development as a progression from involuntary to voluntary memory. The main distinction between these two types of memory is the amount of overt, or deliberate, activity that each involves. Involuntary memory is characterized as an essentially automatic by-product of engagement in a task for which memory is an intrinsic goal. It does not require that the person engaging in the task construct the motivation and the techniques necessary to remember. It is determined by the content and structure of the particular task with which it is linked, and therefore, is not mediated by some conscious activity on the part of the individual. Instead, involuntary memory emerges from the action system of the individual.

It is assumed that involuntary memory emerges first developmentally,

since it involves memory of a mechanical, task-dependent nature (Istomina, 1975). In fact, it is suggested that involuntary memory may be present in a rudimentary form in infancy, as evidenced by the imitation of language and action. However, it is not until 3 or 4 years of age that the existence of involuntary memory becomes obvious. At this point, if the child is engaged in a task for which the goal to remember is so intrinsic to the task that the child is almost "compelled" to remember, then the child should perform relatively well, compared to situations where the goal to remember is extrinsically linked to the task at hand. Leontiev (1978) contends that the child is not yet the master of his or her behavior at this age, but is beginning to be able to unconsciously utilize the task stimuli and context to motivate remembering.

The second type of memory that Soviet psychologists describe is voluntary memory. Voluntary memory is defined as the product of intentional, goal-directed behavior (Zinchenko, 1983–84). It is around the age of 5 or 6 that the Soviet developmental psychologists believe that a child begins to exhibit some form of voluntary memory (Istomina, 1975; Leontiev, 1932). They contend that the emergence of deliberate activity directed at memory is a product of interaction between the child and his or her social world. That is, the child is at first highly dependent on the context in which memory is the goal. The goal to remember is not internal to the child; it is guided by the structure of the task setting. As a child gets older, he or she is exposed to increasingly structured environments (e.g., a school setting), which often require that the child make a deliberate attempt to remember information. Because of interaction in these environments, the child's involuntary memory processes are transformed into purposeful, conscious activity. At first, the child achieves memory goals that are intrinsic to certain tasks. With experience, he or she begins to generalize the goal and techniques for remembering to less supportive contexts. Eventually, the goal to remember is internalized by the child so that memory performance improves even in situations that are relatively meaningless for the child. The child learns to generate his or her own goal of remembering, and it is not necessarily dependent upon task characteristics. Along with generating his or her own memory goals, the child also begins to exhibit intentional activity (e.g., strategies) meant to improve memory.

Istomina's Classical Study of Memory

The design. In the 1940s, Istomina set out to provide empirical support for the developmental transition from involuntary to voluntary memory. Despite her theoretical interest in describing the ontogeny of that transition, she curiously choose a cross-sectional, rather than a longitudinal, design. This decision had critical implications for her interpretation because it limited the extent to which she could draw conclusions about developmental improve-

ments in memory performance. However, Istomina compensated somewhat for the limitation by carefully documenting valuable descriptive data that allow a glimpse into some of the microprocesses that operate in the development of memory skill.

In order to test the hypothesis that memory first emerges in meaningful contexts for which the goal to remember is intrinsic to the task and then generalizes to less familiar contexts for which the goal to remember does not naturally evolve from task activities, Istomina did a within-subject comparison of performance in two task settings. In one task, the goal to remember was intrinsic to the task, and in the other, the child was required to explicitly adopt an experimenter-generated memory goal.

The first task revolved around a "game" setting, in which children were told that they were to pretend to go the grocery store to buy some items requested by the experimenter. Istomina wanted to assure that the game setting "involved a specific, defined situation that created a motivation for remembering" (Istomina, 1975, p. 9), so she strived to make the situation as realistic and familiar to the children as possible. She accomplished this task in several ways.

First, she constructed the grocery store situation to be as ecologically valid as possible (see Chapter 8). All of the features of the context were critical for making the setting as similar as possible to grocery stores that the children experienced in everyday life. The grocery store setting alone, however, did not create a context for which memory was an intrinsic goal. Therefore, Istomina combined the grocery store setting with a real-life kindergarten setting. Children who were "sent" to the grocery store began the experiment in a pretend kindergarten (see Chapter 8). In this setting, the children played the roles of typical adults in a kindergarten (e.g., teacher, teacher's assistants, cook), whereas dolls "played" the roles of kindergarten students. An adult experimenter played the role of kindergarten director, who sent children from the kindergarten one at a time to buy 5 items at the grocery store. Each child was told that the purpose for buying the items was because they were needed for the children (i.e., dolls) in the kindergarten. For example, some of the items were needed for snacktime, while others were necessary for activities, such as drawing and painting, that were typical in a kindergarten classroom. The list of items contained simple, familiar words that could not be readily categorized into groups based on "similarity or habitual association" (Istomina, 1975, p. 11). Unfortunately, from Istomina's documentation of the procedure, it is unclear whether each child received the same words or whether the words varied between children. The latter seems the most likely because Istomina reports a list of 19 sample items (see Table 10.1) from which, she states, the target items were selected.

The creation of a context that provided a reason for the children to buy items at the grocery store was a critical feature of the game situation. By placing the memory goal in a context of meaningful activity, even young

Table 10.1. Stimulus Items used in Istomina's Study and Replications

Study	*Pool of Items*
Istomina	candy, ball, cereal, toys, carrot, milk, socks, paints, cucumber, knife, macaroni, suit, pencils, flowers, cabbage, book, roll, slippers, butter
Schneider & Brun	List 1: apples, pretzels, cheese, salt, toothpaste, pencil, handkerchief, candy
	List 2: carrots, chocolate, bread, sausage, sugar, eraser, laundry soap, cotton balls
Schneider & Hasselhorn	List 1: *unattractive items*: handkerchief, rock, screwdriver, bulb, sponge, cigarette, toothpaste, vase
	List 2: *attractive items*: teddy bear, chocolate egg, engine, licorice, sailboat, toy car, crayons, balloon
Weissberg & Paris	List 1: *Grocery setting*: apples, bread, cookies, lettuce, chicken, eggs
	List 2: *Grocery setting*: crackers, pretzels, carrots, fish, milk, grapes
	List 3: *Party setting*: Norman, Edna, Gus, Bess, Justin, Claire
	List 4: *Party setting*: Dora, Louis, Meg, Walt, Abbey, Hal

children, who were theoretically not yet able to differentiate the memory goal from the task structure itself, were given an opportunity to demonstrate initial stages of memory skill development. Rather than the burden of goal recognition and maintenance being placed on the child, the situation was structured so that it created and sustained the memory goal.

When the children arrived at the store, attempts to maintain the realistic nature of the setting were continued. Each child presented a permission slip to the guard at the door to the store, as a child would be required to do if entering an actual grocery store in Russia of the early 1900s. In the store, the child told the manager what he or she needed to buy, whereupon the grocer gathered the items, the child paid the cashier, and the salesclerk presented the bag of items to the child, who then returned to the classroom. Although not explicitly stated in the procedure section of Istomina's paper, it is apparent from Istomina's description of individual subjects that children were allowed to return to the experimenter for repetition of the items if they could not remember them all (e.g., see Istomina, 1975). The game continued until all the children had participated in the grocery store portion.

Following participation in the game setting, Istomina had each child remember information in the context of a "traditional laboratory setting." Each child sat at a table with the experimenter and was presented with words similar to those in the game setting. (For more detail, see Chapter 8.) The child was simply instructed to remember the words. A critical incongruity between the laboratory setting and the game setting is that children in the lab setting were apparently not given the opportunity to have the words repeated as they were allowed to do in the grocery store setting. Since it was assumed that children would perform less well in the lab situation, this difference could have forced the results in the desired direction.

Beyond this critical methodological difference in the two settings, there still remains Istomina's hypothesized conceptual difference in the two settings. Because the laboratory context was less familiar to the younger children and because it required that the child maintain a focus on the memory goal, rather than the goal being embedded in the task, it was expected that younger children would not perform as well in this setting as in the game setting. In accord with the supposed developmental progression from involuntary to voluntary memory, however, it was expected that older children (5 to 7 years old) would begin to show some understanding of the goal to remember, even in the less meaningful laboratory situation. Therefore, the older children were expected to recall more information than the younger children in the laboratory setting.

The results. Istomina's findings paralleled her predictions. In the game setting 3- and 4-year-olds recalled slightly more items than in the laboratory setting (see Table 10.2). The real differences, however, came with children 4 to 5 years old and above. In both settings, these children recalled more than the younger children. Most importantly, recall scores were higher in the game setting in comparison to the laboratory setting.

If the predominate emphasis in this study were the absolute value of the outcome measures, then one could take issue with the minimal difference between the 1.0 and 0.6 items recalled by the 3–4-year-olds in the game setting and the lesson setting, respectively. In fact, it could be (and has been, Weissberg & Paris, 1986) argued that there was a possible floor effect. That is, the 3- to 4-year-olds recalled so few (or no) items that it is possible that they were not remembering anything in either setting. However, the theory of involuntary/voluntary memory assumes that involuntary memory is just beginning to emerge at around 3 years old. Therefore, it would be predicted that the absolute value of recall still may be quite low at this age. However, it would be assumed that other behaviors may indicate different levels of engagement and understanding depending on the meaningfulness of the context. It would be predicted that because of younger children's better understanding of more meaningful contexts, their overall memory performance should improve incidentally in the game setting. In contrast, in the laboratory setting, it would be predicted that young children should appear almost confused; an indication that they do not explicitly understand the goal of remembering in that context. These predicted differences in the two settings were apparent in the rich observational data that Istomina presented. It is these data that provide the best indication of the processes that mediate performance in each of the settings.

In descriptions of individual children's verbal and physical behavior in the two settings, Istomina was able to demonstrate that while young children seem to have not yet internalized the goal of remembering, the game setting provided a meaningful context from which memory emerged as a result of

Table 10.2. Comparison of results from Istomina and replications

Istomina				
Age of Subject	# of Ss	Lab Recall	Game Recall	Ratio
4*	30	1.05	2.0	1.9
6*	30	2.15	3.5	1.6

Weissberg and Paris							
Age of Subject	# of Ss	Lab Recall		Game Recall		Ratio	
		Food	Names	Food	Names	Food	Names
4	48	3.10	1.40	2.55	1.03	0.8	0.7
6	48	4.10	2.80	3.55	2.30	0.9	0.8

Schneider and Brun: Ex. 1				
Age of Subject	# of Ss	Lab Recall	Game Recall	Ratio
4	30	1.63	3.56	2.2
6	30	3.06	5.46	1.8

Schneider and Brun: Ex. 2				
Age of Subject	# of Ss	Lab Recall	Game Recall	Ratio
4	30	1.59	2.08	1.3
6	30	2.53	2.79	1.1

Hasselhorn and Schneider: Ex. 1				
Age of Subject	# of Ss	Lab Recall	Game Recall	Ratio
4	64	3.48	4.45	1.3
6	64	4.63	5.03	1.1

Hasselhorn and Schneider: Ex. 2				
Age of Subject	# of Ss	Lab Recall	Game Recall	Ratio
4	80	1.68	1.40	0.9
6	80	2.10	1.73	0.8

*The means for 3-year-olds and 4-year-olds have been collapsed into one group and the means for the 5-year-olds and 6-year-olds have been collapsed into one group in order to make the Istomina and the Weissberg and Paris results comparable to those from Schneider's lab, which were reported as two age groups rather than four.

participation in the task. In the game setting, 3-year-olds were extremely motivated to "play the game." For example, Istomina reports that one child got so excited about being told that it was his turn to go to the store, that he ran off to the store without hearing the items to be bought (Istomina, 1975). Another child told the experimenter that she wanted to go to the store to buy ice cream. When she arrived at the store, she "remembered" only ice cream, even though it was not one of the target words. However, many of the 3-year-olds did manage to remember a minimal number of items. This success seemed to be influenced merely by participation in the task, and not because the 3-year-olds were able to generate an internalized memory goal. This interpretation is based on the fact that 3-year-olds seemed fairly clear about the purpose of the task, but could not, when asked, provide an explanation for why or how they remembered. Furthermore, no obvious strategies to aid memory (e.g., rehearsal of the items) were observed. The children were more focused on the mechanics of the task, and from that, a modest level of recall

resulted. Istomina concluded (understandably!) that the youngest children in her study were more focused on being shoppers than on the real purpose for the errand. However, she pointed out that the goal to remember was an internal aspect of the task, rather than an externally prescribed goal. Therefore, mere participation in the task led to a minimal level of recall.

In contrast to the game setting, 3-year-olds appeared unable to understand and adopt the experimenter-articulated memory goal in the lab situation. Possibly, this stemmed from their developmental inability to differentiate a task goal from a task setting. Because the goal was not embedded in the task structure, in which case achievement of that goal would have been a natural result of engagement in the task, young children were unable to make the link between the goal and task activities. Since the memory goal in the laboratory setting was an externally imposed goal, Istomina had predicted that the youngest children would not be able to adopt the goal. This lack of understanding was evident in the case of one 3-year-old who, when asked to remember the items in the lab situation, said, "Where are they?" (p. 36). Another 3-year-old in the lab setting did not even attempt to remember the items, but talked instead about his cat at home. These responses within the lab setting suggest that the younger children did not understand the goal of the task at all. Although many of the 3-year-olds did not display higher levels of recall in the game setting than in the lab setting, they at least demonstrated more understanding about what was expected of them in the game setting than in the lab setting.

Older children (4 to 7 years old) displayed a different pattern of results. As a whole, these children remembered more than the younger children in both settings. However, the 4- to 5-year-olds were differentiated from the 5- to 7-year-olds on the basis of the descriptive data. That is, the older children exhibited behaviors in both settings indicating that they were very deliberate in their attempts to remember the information. Specifically, 5-year-olds and above tended to use overt verbal rehearsal of to-be-remembered items in both settings. This behavior indicated a clear understanding of the goal to remember and techniques subordinate to that goal.

Observed behaviors in the 4-year-old group were less consistent, suggesting that this age group represented a transitional period in the progression from involuntary to voluntary memory processes. Some of the 4-year-olds displayed verbal rehearsal in the game setting, but it was rare in the laboratory setting. In the game setting, it was very evident that the 4-year-olds understood that the primary goal of the task was to remember. Istomina reports that most of the 4-year-olds listened intently until all of the items had been read and were impatient with any attempts to interrupt them between the kindergarten room and the grocery store. Istomina concluded that these children were able to explicitly adopt a memory goal, but only when it was embedded in a meaningful task setting. In the laboratory situation, the

4-year-olds seemed to realize the importance of listening to the items, but when it was time to remember them, they were for the most part unsuccessful. They understood that the task was a memory task because the experimenter had told them that, but they were unable to explicitly set the goal for themselves and to pursue activities relevant to that goal.

In conclusion, Istomina suggested that her results supported the hypothesis that younger children first demonstrate the ability to remember in contexts for which the goal to remember is intrinsic to the task at hand. She stated that younger children understand what it means to remember, but only insofar as they experience that goal as a by-product of their motivation to engage in a task. The goal must be internally rather than externally linked to the task, as was the case in the game setting. In the laboratory situation, the only link between the goal and the task was the experimenter. For younger children, who did not yet have the skills to utilize this externally directed goal, the task was nearly impossible. For older children, however, who had begun to understand externally set goals, it was possible to observe intentional behavior directed toward the memory goal.

Weissberg and Paris' Replication (1986)

The design. Although Weissberg and Paris conducted their study more than 40 years after Istomina completed hers, it was their objective to adopt her basic design and to test the replicability of her results. While the basic design as far as the game versus laboratory situation comparison was retained, Weissberg and Paris identified some design and procedural flaws in Istomina's work and, in their replication, corrected those problems.

The subjects included in the Weissberg and Paris study were similar to those in Istomina's study as far as age (3–6-year-olds). However, the subjects differed in terms of their cultural backgrounds as the Weissberg and Paris study was conducted in the United States. The two experimental situations that Istomina used were basically retained. There was a grocery store setting, complete with store counter and cash register, toy food items, and other items relevant to a grocery store. As in Istomina's study, each child went to the store to "buy" the to-be-remembered items. However, unlike Istomina, Weissberg and Paris did not include the kindergarten aspect of the grocery store setting; instead, children were brought one at a time to an empty room at the school where testing took place and were told that they would be playing some games. For the grocery store setting, each child was told to "go" to the grocery store to buy some items. For the lab setting, the child sat at a table with the experimenter and listened to and then recalled the items.

From that point, Weissberg and Paris's study deviated somewhat from Istomina's. They identified two critical design flaws that they corrected in their replication. First, they were concerned that Istomina did not counter-

balance the order of the two settings. Therefore, half of their subjects participated in the grocery store setting followed by the lab setting, and the other half received the reverse order of the activities. Second, they did not feel that Istomina had sufficiently controlled the items presented to the children for remembering. Thus, rather than allowing the items to vary across subjects, they constructed two word lists, which were counterbalanced across setting, and which all children received. In contrast to Istomina, Paris and Weissberg's lists contained all food items, therefore presenting the children with sets of items that were more semantically related to each other than the words used by Istomina (see Table 10.1). Although it is unclear which items Istomina presented to her subjects, she states that the lists were constructed so that no two words on the list could be grouped into semantic categories. In the list of possible words she reports, there are food, clothes, and school supply items. Therefore, it is feasible that the items presented to Istomina's subjects did not belong to similar semantic categories. Memory development research in the past 10 years has demonstrated that children, especially very young children, show recall improvements when presented with semantically related to-be-remembered items (e.g., Best & Ornstein, 1986; Corsale, 1978; Folds, Ornstein, & Bjorklund, 1988), in contrast to sets of unrelated items. Therefore, because Istomina purportedly used unrelated items, while Paris and Weissberg used more related items, it would be expected that the absolute value of the recall by Paris and Weissberg's subjects would be higher than that of Istomina's participants.

Beyond these specific differences between the original and the replication, Weissberg and Paris seemed to have more general control over the experiences of their subjects. For example, in both settings, the words were read twice to the children and in neither setting did an experimenter repeat the items after recall was attempted. In Istomina's description of the game setting, it is apparent that if a child forgot some items he or she was free to return to the kindergarten to hear the list again. This, of course, represents a very critical difference between the game and laboratory settings in Istomina's study, which most assuredly resulted in an unbalanced effect on recall.

In addition to the two settings used by Istomina, Weissberg and Paris had their subjects participate in a third setting, a party setting that involved the recall of the "names" of stuffed animals who were at a birthday party together. The rationale behind including this setting was to record performance in a situation which was perhaps a little more familiar to the children. In this setting, each child was shown stuffed animals sitting in a circle and told that an animal he or she was holding could not join the party until he or she could remember all the animals' names. The child was told the names and then asked to introduce the animal he or she was holding to the animals at the party. To parallel the party situation, children also were asked to recall names

in a laboratory setting so that a direct comparison of the party setting to a lab setting could be made.

The results. While the differences in the methodologies of the Weissberg and Paris and Istomina studies seem fairly subtle, the differences in the results are quite compelling. In fact, they completely contradict each other. In the Weissberg and Paris study, it is the laboratory settings rather than the grocery store or party setting that yield the highest levels of recall (see Table 10.2). This difference is true for both the older and the younger children. Furthermore, Weissberg and Paris's subjects displayed higher overall recall than Istomina's, regardless of the setting. Interestingly, some rehearsal data that Weissberg and Paris report also contradict Istomina's, indicating that children in all age groups tended to rehearse aloud significantly more in the lab setting than in the game setting.

These results suggest that perhaps the lab setting meant something entirely different to the American children than it did to the Soviet children. Weissberg and Paris concluded that cultural differences may have affected their ability to replicate Istomina's results. They suggested that American children may be more comfortable in situations for which recall is explicitly organized by some externally set goal rather than situations in which recall is simply driven by some motivational disposition engendered by the task at hand. This conclusion, of course, calls into question the generalizability of Istomina's beliefs about the emergence and development of memory. The question that we consider later in the chapter is whether Weissberg and Paris are justified in making this strong conceptual statement based on the study that they conducted. Before we do that though, we consider the set of replication attempts conducted by Wolfgang Schneider and his colleagues.

Schneider and Brun's Replications (1987)

Design. Schneider and Brun designed two experiments that were intended to serve the purpose of replicating Istomina's original study. As did Weissberg and Paris, they identified major design flaws and corrected them. However, in Experiment 1, like Istomina, they allowed their subjects to return to the experimenter for repetition of the to-be-remembered items. Their purpose in retaining this flaw was to allow a more direct comparison between their results (derived from a more sound experimental design than Istomina's) and those obtained by Istomina.

Therefore, in Experiment 1, a group of 4-year-olds and a group of 6-year-olds participated in two memory tasks; one that was part of a grocery store setting and one that was part of a laboratory setting (see Chapter 9). Although the objective was to completely retain the features of Istomina's original

study, like Weissberg and Paris, Schneider and Brun did not provide a kindergarten context as the reason for each participating child to "buy" items at the grocery store. The utilized word lists seemed to be more similar to Istomina's than were Weissberg and Paris's because they contained items other than food. However, while Istomina was unclear about the items with which her subjects were presented, Schneider and Brun opted to have two standard lists, which were counterbalanced across the two settings.

In both the grocery store and lab setting, the children were told that they would only hear the words once, although, as in Istomina's study, children were permitted to return to the experimenter for repetition of the to-be-remembered items if they could not recall them all. Unlike Istomina, children could ask for repetition in both the grocery store and the lab setting. However, Schneider and Brun reported that more children asked for repetition in the game setting than in the lab setting.

The result of this study was a replication of Istomina's findings (see Table 10.2). In fact, even the ratios between recall in the game and the lab setting were almost identical to Istomina's. Therefore, in contrast to Weissberg and Paris who did not allow repetition of the items, Schneider and Brun were able to repeat Istomina's original findings. Although Schneider and Brun were certainly pleased that they successfully replicated Istomina, they were skeptical about the validity of their findings for testing performance differences in more and less meaningful settings since the children asked for repetition of the items far less in the lab setting than in the grocery store setting.

In order to explore the possibility that the repetition of items in the grocery store setting created a critical incongruity between it and the lab setting, Schneider and Brun pursued a second experiment in which more of Istomina's design flaws were corrected. As in Weissberg and Paris, the game and lab setting were counterbalanced within subjects and the word lists were counterbalanced across settings. In contrast to Weissberg and Paris, however, children in the game setting were told to listen carefully to the words because they would not be read again and, unlike Experiment 1, even if the child returned to the experimenter the words were not repeated. The results from this experiment indicated the same trend in the means as Istomina achieved; however, the means in the game setting were deflated making the differences between recall in the two settings nonsignificant.

Schneider and Brun concluded that the children recalled more in the game setting than in the lab setting in Experiment 1 because they heard the words more in the game setting. Their failure to replicate Istomina's results in Experiment 2 led them to suggest that Istomina's conclusions about the importance of a motivating context for early memory success are not valid. They argued instead that the most important determinants of memory performance are technical aspects of the memory context, such as the availability of auditory and pictorial cues and the length of exposure to to-

be-remembered information. This conclusion seems somewhat ironic, however. Schneider and Brun argued that by altering the "technical" aspects of the experiment, they demonstrated deviations from the original findings. Their claim was that they cast doubt on the theory that children's memory emerges first in meaningful contexts. However, their design variations, in fact, altered not only specific aspects of the context, but also the more general context and its meaningfulness for the child. Therefore, they may have indirectly supported Istomina's contention that the more meaningful the context, the more likely younger children will be able to recall information presented in the context. To clarify, Schneider and Brun made the game setting more stilted by telling the children specifically that they would hear the items only once. In the real world, one would rarely be told, "Now, listen carefully; I am not going to tell you items you need to buy more than once." By placing this limit on the child, some of the meaning of the setting was removed. Schneider and Brun's results in Experiment 2 may have emerged because of this alteration in the meaning of the game context. They did not find that recall in the lab setting was better than in previous experiments; they found that recall in the game setting was depressed. Istomina might argue that by stripping the game setting of some of its meaning, Schneider and Brun have produced a negative effect on recall, and thus have to an extent demonstrated the importance of a meaningful context.

Hasselhorn and Schneider's Replication

After perusing and comparing the results of the replication attempts from their own lab and those produced by Weissberg and Paris, Schneider and his colleagues concluded that the validity of Istomina's theoretical notions had not yet been fully determined. Understandably, they were concerned that not only did the results of their two replication attempts, once methodological corrections had been made, seem unclear in terms of Istomina's results, but also that their findings differed so dramatically from Weissberg and Paris's. In order to clarify the differences, Schneider and his colleague, Marcus Hasselhorn, proceeded to extend Schneider's earlier replication attempts. They were especially concerned about possible methodological differences (e.g., the items used, the number of times the items were presented) in the previous replication attempts, and the impact of these differences on the results. Accordingly, they designed two experiments to address some of their concerns.

Experiment 1: Design. In this experiment, Hasselhorn and Schneider were concerned with an observation from the first set of studies out of their lab; that is, children tended to recall words that appeared to be more attractive to them. Therefore, in Experiment 1, Hasselhorn and Schneider set about to explore the effects of item attractiveness by systematically pre-

senting subjects with items judged by children to be either attractive or unattractive. The procedure used in this experiment was basically identical to the earlier studies (see Chapter 9); however, the design differed in that remembering in different settings was manipulated on a between subjects rather than a within subjects basis.

Results. The results of this experiment followed the same pattern as the earlier Schneider studies; children remembered more in the game setting than in the lab setting. However, this difference was limited to the attractive items. Furthermore, rehearsal was more likely with the attractive items than with the unattractive items.

Hasselhorn and Schneider concluded that despite improvements in the Istomina design, they were able to replicate her findings. They suggested that the context is important, but that we must be careful to identify specific features of the context which influence memory rather than talking about the context in global terms. They suggested further that the game versus lesson nature of the context is an "objective environmental feature" (see Chapter 9) rather than a subjective feature such as the perceived attractiveness of the items. This conclusion seems to contradict Istomina's premise that the effectiveness of the game setting rests in the fact that as children are engaged in that setting they become personally involved in and motivated by aspects of that setting. It is this episodic subjective experience that drives cognitive success in the setting. This point will be elaborated further, but first let us explore the final replication attempt and its implications.

Experiment 2: Design. In this experiment, an attempt was made to further standardize the children's experiences in the two settings. Hasselhorn and Schneider had observed on videotape and from interviews with experimenters that the interpersonal features of their study varied across settings. Therefore, in their second experiment, Hasselhorn and Schneider instructed the experimenters to interact with the children in the same manner in the game setting as they did in the lab setting.

Results. The results from this experiment were exactly opposite those of earlier Schneider studies. Children recalled more in the lab setting than in the game setting. Also in contrast to Experiment 1, the attractiveness by setting interaction did not convey. Instead, the effect of item attractiveness was qualified only by an age effect; older children recalled more attractive items than younger children.

INTERPRETING INCONGRUITIES

Obviously, there are many differences not only between Istomina's original study and the replications, but also between the replications themselves. How

are we, then, to interpret these discrepancies? If we do a very simple meta-analysis (see Chapter 3 and 9) and take a tally of the "pro-game setting" versus "pro-lab setting" findings, we discover that the game setting wins, at least if we are speaking in terms of mean trends (see Table 10.2). Does that mean that we should accept Istomina's findings as the "correct" ones? According to the conclusions of Weissberg and Paris and Schneider and his colleagues, we should not. Both sets of researchers suggest that by instituting methodological improvements in Istomina's design they have successfully cast doubt on her conclusions regarding the development of memory in young children. The question is do we feel comfortable with and justified in making this conclusion? In order to guide a more informed answer to this question and to enlighten us about the process of replication and its implications, I will now analyze in detail important methodological and conceptual considerations in the original Istomina study and follow-up replications.

Methodological Differences

As was discussed previously, both Weissberg and Paris and Schneider and colleagues independently identified critical methodological flaws in Istomina's original study which could have had direct effects on the results obtained. I do not deny the importance of their methodological improvements. It is obvious that Istomina failed to use some design manipulations that are typically employed for the purpose of increasing the internal validity of a study's experimental conditions. Given these improvements by the replication authors, though, we still do not get a clear picture of the influence of context on memory performance. In Table 10.3, I organized a summary of the Istomina-type studies and all of the different design features that have been used.

The confusion that one might experience when trying to decipher this table is quite illustrative of the difficulty in piecing together the implications of Istomina's original study and the various replications. A mere glance at this table makes evident the many subtle design differences between Istomina's study and its replications. Consideration of these differences and the similarities between the studies brings to light possible explanations for the discrepant results that were obtained.

Since Schneider and colleagues were for the most part successful in replicating Istomina, let us first consider their research efforts and results. If we examine the results for their lab setting, we see that in all but one of the Schneider studies, that is, Hasselhorn and Schneider, the recall data in the lab setting are very close to what Istomina obtained. These similarities occurred in spite of some critical design differences, for example, counterbalancing of the setting and lists. In Experiment 1 of Hasselhorn and Schneider's study, recall in the lab situation was higher than that obtained by Istomina and in Schneider's other studies. A possible explanation for this increase in recall is

Table 10.3. Comparison of Methodologies

	Istomina	*Weiss. & Paris*	*Schnei. & Brun:1*	*Schnei. & Brun:2*	*Hassel. & Schnei:1*	*Hassel. & Schnei:2*
No Counterbalancing	X					
Counterbalanced Setting		X	X	X		
Completely Within Subject	X	X	X	X		
Setting Between Subject					X	X
Word lists varied systematically	X					
Used same word lists across subjects		X	X	X	X	X
Counterbalanced lists across setting		X	X	X	X	X
Word Lists contained items other than food	X		X	X	X	X
Repeated items if subject requested (esp. in game setting)	X		X			
Told specifically that words would only be read one time				X		
Repeated items two times for all subjects		X			X	X
Presented items visually as well as auditorially					X	X

that items were presented both visually and auditorially. This explanation makes sense because we see the same increase in recall in the game setting. Therefore, it seems that the "double exposure" in Experiment 1 was possibly responsible for the increased recall. In spite of this one instance of increased recall, though, it is apparent that the methodologically different methods used by Schneider and his colleagues did not affect performance in the laboratory situation.

For the game setting, however, Schneider and his colleagues' methodological changes produced marked differences in recall. The level of recall in Experiment 2 of Schneider and Brun's replications was closest to the level of recall in Istomina's study. However, the ratios between recall in the lab and the game setting were most similar in the Istomina study and Experiment 1 of Schneider and Brun's study. This similarity is, of course, understandable because it is Schneider and Brun's first experiment that most resembles Istomina's original study. Specifically, it was in this study that subjects were allowed to return to the experimenter for repetition of the to-be-remembered items. Although this advantage was allowed in both the lab and the game setting (possibly in contrast to Istomina's only allowing it in the game setting), Schneider and Brun report that children rarely asked for repetition in the lab setting. Therefore, it is probable that recall in the game setting, as in Istomina's study, was enhanced (in fact, even more than in Istomina's study) by

repeated exposure to the items. In Schneider and Brun's second experiment, in which repetition was not allowed, recall in the game setting was depressed relative to Experiment 1. However, the mean trend still suggests that the game setting was a better situation for remembering.

Because Schneider and his colleagues' patterns of findings generally mirrored Istomina's, it is puzzling why Hasselhorn and Schneider's second experiment did not produce the same pattern of results. Instead, in this study the lab setting resulted in superior recall over the game setting. This study was almost identical to Experiment 1 of Hasselhorn and Schneider in terms of its method, except that visual versus auditory presentation was treated as a between subjects rather than a within subjects variable. Furthermore, the experimenters were instructed to equate their behavior across the two settings by being less engaging and interactive in the game setting. It was believed that with this control social interaction in the game setting would more closely resemble that in the lab setting.

Hence, rather than observing a change in recall in the lab setting, Hasselhorn and Schneider demonstrated a significant depression in recall in the game setting to the extent that recall level was lower than in the lab setting. One reason that the effect may have been so extreme is that by including in the game setting an experimenter who was not engaging and with whom the children did not enjoy interacting, that setting was essentially stripped of some of the motivating features that Istomina had predicted would support more successful recall. By making the game setting more like the lab setting, it was less likely that the participating children were actively involved in the "fun" of the game. Therefore, to some extent they had to set their own memory goals rather than experiencing memory as a more natural result of participation in the task. If this is truly the case, then even to a larger extent than the Schneider and Brun replications, Hasselhorn and Schneider's second experiment indirectly demonstrated the importance of a meaningful context for which memory is intrinsic to the designated task.

In spite of the relatively consistent pattern of findings in the set of studies by Schneider and his colleagues, we are still left with the reversal in the pattern of findings in the Weissberg and Paris study. This reverse pattern was true, even though the methods in this study resembled closely the methods in Schneider and Brun-Experiment 2. That is, in both studies the setting was counterbalanced across subjects and the lists were standardized across setting and subjects. One difference was that Weissberg and Paris presented each item two times, in comparison to Schneider and Brun's one time each. Also, they used word lists containing only food items, whereas both Istomina and the Schneider lab used lists containing items other than food. Both of these factors could have contributed to the higher levels of recall in the Weissberg and Paris study. First, more exposure to to-be-remembered items has been shown to enhance recall (e.g., Naus, Ornstein, & Aivano, 1977).

Second, since Weissberg and Paris constructed their word lists of only food items they probably were more semantically related in comparison to the Istomina and Schneider lists. This high semanticality could have led to enhanced recall in their study. Research on memory in children supports this contention as it has been shown that young children will display higher levels of recall if they are presented with to-be-remembered items that are highly associated with each other and which belong to obvious semantic categories (e.g., Bjorklund & Zeman, 1982; Folds, Ornstein, & Bjorklund, 1988). However, these factors do not explain the difference in the pattern of findings produced by Weissberg and Paris in comparison to Schneider and Brun and Istomina; they only speak to differences in the level of recall.

What, then, caused the significantly enhanced recall in Weissberg and Paris's lab setting? One possible explanation is that the subjects who participated in the various studies came from different cultural backgrounds. Weissberg and Paris suggest that the American children who participated in their study were exposed at an early age to situations involving more rote-like memory (e.g., watching Sesame Street), making it probable that the lab setting as opposed to the game setting was more familiar, understandable, and perhaps, motivating. This explanation seems even more likely when you consider the unfamiliarity of the game setting for the American children. In the United States, it would be rare for a child to have the experience of going to a small grocery store where the buyer does not have the freedom to move around the store and pick up to-be-bought items on his or her own. In most cases, the shopper interacts very little with the shopkeeper; usually only briefly when paying for the items. Therefore, not only does the lab setting possibly have more meaning for the American children, the grocery store setting has a lot less meaning.

In contrast, for the Russian children in the Istomina study, it was quite common during the 1940s to go to the market with their mothers, where the two would be greeted by a guard at the door. The guard would direct them to give the grocer their list of needed items, which he would gather and bag for them. Likewise, for the German children, who participated in the Schneider studies, trips to a small grocery store where a grocer (or butcher or baker, as the case may be) is told the items that are to be bought is much more common than in the United States. However, it is also true that German children come from a culture that puts a great deal of emphasis on academic achievement; therefore, the more rotelike memory situation was certainly not unfamiliar to them. Unlike Americans, though, Germans view kindergarten as a time for children to develop good social skills; therefore, kindergarten-aged children spend more time playing than they do in structured group activity. If these cultural differences are real, then it is possible that American children do first experience successful remembering in situations where memory goals are set more explicitly. In contrast, Russian and German children may benefit more

from situations in which a memory goal is intrinsic to the task in which they are engaged. Assuming these conclusions are valid, then it would make sense to question the universality of the Soviet theory of memory development.

While this cultural explanation seems plausible, in many ways it is unsatisfying. First of all, the cultural differences between the United States and modern-day Germany are not that distinct. The philosophies of kindergarten may be different, but Germans put a great deal of emphasis on their children's academic achievement. In fact, German and American children are exposed to many of the same games and television shows. Cross-cultural research (e.g., Schneider, Borkowski, Kurtz, & Kerwin, 1986) has shown that in terms of memory performance, it is German children who first display more spontaneous use of memory strategies in laboratory-like settings compared to American children. Furthermore, although Schneider and his colleagues generally produced the same patterns of results as Istomina, the differences between their game and lab settings were small even when methodological flaws were corrected. Therefore, the question still stands. In what type of context do children first begin to display evidence of developing memory skill?

One critical difference in the original study and the replications makes it nearly impossible in my estimation to come to any conclusions about the effect of context on memory development. The difference lies in the intrinsic features of the two experimental situations as constructed by Istomina and the replication authors. In Istomina's experiment, the grocery store situation was set up so that the situationally intrinsic motivation to perform the goal was optimized. The children started the experiment in a classroom (kindergarten) with other children. Each participant was told when it was time for a snack and that he/she would be responsible for "buying" the snack from the store. Neither the Paris nor the Schneider replications used the kindergarten setting. Instead, one child at a time was taken to a room with an experimenter and told that they were going to play a grocery store game. Subsequently, the child was sent to the "store" where he or she was supposed to "purchase" the targeted items. Presenting the grocery store situation in such a manner makes it more like a laboratory setting. The goal may be more interesting to the children, but it is only superficially intrinsic to the situation.

To give Weissberg and Paris credit, by using the party setting they did attempt to create a new context that they judged to be more meaningful for the American children. However, while emphasizing the meaningfulness of the context they failed to consider the intrinsic nature of the task itself to the context. Yes, American children are very likely to have had a good deal of experience with birthday parties. However, how often do all the children at a party sit in chairs and try to remember everyone's names? Furthermore, although remembering names at parties may be a typical expectation, the process of remembering is usually carried out in a much more enriched

context. That is, there are personalities and past associations attached to most of the names. Even if two children meet for the first time at a party, they immediately begin sizing each other up and interpreting each other's behavior. As such, the name "Tommy" or "Susie" means something. Thus, if a child needed to remember Tommy and Susie, he or she may be able to because Tommy and Susie are salient to him or her and there is a reason for recalling their names. A group of stuffed animals sitting in chairs does not have the same saliency to a child. Therefore, they are in essence remembering a group of essentially random names. Such a task requires a great deal of concentration which is more likely to occur in a structured laboratory situation rather than a more relaxed game setting.

The implications of these conceptual differences are critical to the outcomes of the studies in question and impact considerably on interpretation of the discrepant findings. Istomina's conceptualization of the research question is dependent on the embeddedness of the memory task in the context created for the participating subjects. This means that the memory goal must be intrinsic to the task at hand and should not require that the subject explicitly recognize and accept the goal. The motivation to perform the memory task should not be provided by an external source, but rather should be generated by involvement in the task.

In the replications of Istomina, not only was the game setting possibly unfamiliar to the participating children, but the unfamiliarity may have led to some confusion about expectations for the task. If the memory tasks in the game settings had been better grounded in general task activity, that is, the children were given a concrete context-based reason for remembering information, it is possible that the replication results would have turned out differently. As it stands, the only conclusion we can make is that Paris and Weissberg have demonstrated that American children may be accelerated in their memory development in comparison with Russian children. That is, American children show evidence of memory ability in settings where the memory goal is less embedded in the context at a younger age than Russian children. Since American children are familiar with structured memory situations, it is possible that the lab setting was a meaningful and motivating context for the American children. However, this finding does not suggest that these children would not first demonstrate memory ability in contexts for which memory is intrinsic to the task. That question is not answered by the replications, because they did not provide such a context for their participants.

SUMMARY

After our analysis of these five replication attempts of Istomina's work, what do we conclude? I have pointed out numerous methodological and outcome

inconsistencies. The two sets of replication authors conclude that they have successfully cast doubt on Istomina's original premises, but the evidence does not seem so clear. In fact, we should not feel comfortable at all about reaching any conclusions concerning the validity and reliability of Istomina's study as a foundation for understanding the development of children's memory based on the replication results.

When pursuing replication, there are two general levels of analysis that should be taken into account. First, an original piece should be considered from a methodological perspective. In order that real technical comparisons can be made, any replication should involve as closely as possible a reproduction of every documented aspect of the original study. The replication should obviously also involve a group of subjects who are exposed to an "improved" design (without the original flaws). This group, of course, may be broken into several groups in order to explore different flaws in the original.

Unfortunately, none of the replication attempts considered here provided a direct comparison of Istomina's work. Therefore, putting all theoretical questions aside, it is impossible to make any statements about the methodological problems that plagued Istomina's study. Schneider and Brun in their first experiment came the closest to generating a direct comparison, but because they failed to include the kindergarten setting and because they counterbalanced and standardized the word lists in this study, it is difficult to estimate the features of the design that had the largest effect on their success in replicating Istomina's findings.

Beyond the more technical nature of a methodological analysis of a study, a second level of analysis is theoretical. If a replication author wants to comment on the overall theoretical implications of a work, it is important that he or she consider the theoretical justification for the original design and adjust his or her design to reflect that framework.

In any study, design features are critical to the outcome of the study. Manipulation of design features could yield changes in behavior that might suggest that performance is rather fragile and seemingly superficial. However, every design feature is an integral part of the context that is created in the study. For Istomina, this context was absolutely critical to her theoretical framework. She created a context that in her estimation was highly meaningful for the children in her study and one that she judged as less meaningful. In the replications of her work the authors have voiced concern, and rightly so, over some blatant methodological flaws that plagued Istomina's work. In correcting those flaws, however, they have removed their studies from the conceptual framework that guided the original work. In doing so, they may have created superficial changes in behavior that confuse rather than clarify the concept of activity and intrinsic motivation in memory development.

This confusion speaks very loudly to the issue of how we go about conceptualizing replication. Both Weissberg and Paris and Schneider and colleagues

failed to create the rich context that Istomina created with the kindergarten setting. They choose instead to assert more control over the game setting, thus sterilizing it to the point that it possibly lost its meaning. In this case, then, although methodological improvements were certainly made, the essence of the experiment was obscured. If our goal in psychological investigations is to more thoroughly understand psychological principles, then we should give more careful consideration to the theoretical notions underlying our methods. My point is not to harshly criticize Weissberg and Paris and Schneider and his colleagues, however. Both groups of researchers executed commendable studies. Instead, I am suggesting that beyond purely methodological considerations in replications, we should also contemplate intrinsic features of experimental settings that may influence individuals' behaviors. If cultural, interpersonal, or contextual variables are judged to be central components affecting behavior in a certain study, then those pursuing replications of the work should not hesitate to alter the original methodology to test the importance of these factors on new samples of individuals. By including in any replication both methodological and theoretical levels of analysis, one could more validly reach specific conclusions about the reliablity of the methods of study and more general conclusions about their theoretical implications.

REFERENCES

Best, D. L., & Ornstein, P. A. (1986). Children's generation and communication of mnemonic organizational strategies. *Developmental Psychology, 22,* 854–853.

Bjorklund D. F., & Zeman, B. R. (1982). Children's organization and metamemory awareness in their recall of familiar information. *Child Development, 53,* 799–810.

Corsale, K. (1978). *Factors affecting children's use of organization in recall.* Unpublished doctoral dissertation, University of North Carolina at Chapel Hill.

Folds, T. H., Ornstein, P. A., & Bjorklund, D. F. (1988). *Context specificity of mnemonic strategy use.* Paper presented at the biannual meeting of the Society for Research in Child Development, Kansas City, MO.

Gal'perin, P. I. (1968). Towards research of the intellectual development of the child. *International Journal of Psychology, 3,* 257–271.

Istomina, Z. M. (1975). The development of voluntary memory in preschool-age children. *Soviet Psychology, 13,* 5–64.

Leontiev, A. N. (1932). The development of voluntary attention in the child. *Journal of Genetic Psychology, 40,* 52–83.

Leontiev, A. N. (1978). *Activity, consciousness, and personality.* Englewood Cliffs, NJ: Prentice-Hall.

Naus, M. J., Ornstein, P. A., & Aivano, S. (1977). Developmental changes in memory: The effects of processing time and rehearsal instructions. *Journal of Experimental Child Psychology, 23,* 237–251.

Schneider, W., Borkowski, J. G., Kurtz, B. E., & Kerwin, K. (1986). Metamemory and motivation: A comparison of strategy use and performance in German and American children. *Journal of Cross-Cultural Psychology, 17,* 315–336.

Schneider, W., & Brun. (1987). The role of context in young children's memory performance: Istomina revisited. *British Journal of Developmental Psychology, 5,* 333–341.

Weissberg & Paris. (1986). Young children's remembering in different contexts: A reinterpretation of Istomina's study. *Child Development, 57*, 1123–1129.

Werner, H. (1937). Process and achievement—A basic problem of education and development psychology. *Harvard Educational Review, 7*, 358–368.

Wertsch, J. V. (1981). *The concept of activity in Soviet Psychology*. New York: Sharpe.

Zaporozhets, A. V. (1977). The development of logical thinking in the preschool-age child. *Soviet Psychology, 15*, 60–72.

Zaporozhets, A. V. (1979–1980). Thought and activity in children. *Soviet Psychology, 18*, 9–22.

Zinchenko, P. I. (1983–1984). The problem of involuntary memory. *Soviet Psychology, 22*, 55–111.

CHAPTER 11

The Forbidden Colors Game: An Argument in Favor of Internalization?

René van der Veer
University of Leiden
Leiden, The Netherlands

Cognition rests on the flexible use of mental (meta)-skills that have been mastered in various task settings. One way of looking at these skills is to conceive of them as resting on the internalization of specific tools and instruments that cultures provide. Such a view was defended by the Soviet psychologist Lev Vygotsky, whose ideas have earned him the reputation of a "classic" thinker in selected circles of Western psychology and pedagogics (Bruner, 1984; Wertsch, 1985). More specifically, he argued that children will first make use of the external instruments available in their culture and then shift to internal ones based on speech (e.g., Vygotsky, 1928, 1929a, 1930b, 1931). His collaborator Leont'ev (1932) attempted to corroborate this view in experimental practice through the so-called forbidden colors game. In this chapter Leont'ev's experiment, its replications, and their theoretical implications are discussed.

EXTERNAL MEDIATION

That external instruments or tools are crucial in problem solving is an ancient idea that got new impetus in the beginning of this century through Köhler's (1921) famous chimpanzee research. Köhler's idea was to provide his chimpanzees with various tools (e.g., sticks, boxes, etc.) and to see whether and how they would use them in order to solve relatively complex problems, usually obtaining an object that was attractive, but difficult to reach. For various reasons—for example, its relevance for the comparison of human and animal cognition—Köhler's research greatly impressed the scientific community and

many researchers attempted to replicate his findings with different samples (e.g., the study of chimpanzees and other apes by Guillaume and Meyerson, 1930, 1931; the study of human children by Brainard, 1930; Bühler, 1919).

Among the researchers who were fascinated by Köhler's findings and who used his setup for the study of human subjects we find the Soviet research collective headed by Vygotsky, Luria, and Leont'ev (e.g., Vygotsky, 1929b, 1930a; Vygotsky & Luria, 1930). Inspired by Köhler they presented children of various age groups and mental ability with tasks that could not readily be solved without making use of some tools, or instruments that "happened to be present." Subjects were, for instance, presented a series of pictures once and then asked to reproduce them. The length of the series was chosen so as to make this task too difficult for most children. They were then suggested to make use of the "tools" (or means, or instruments) that were present. Practically anything could serve as a tool: pieces of paper, screws, shot, and so on. The interesting thing was that children were indeed able to use these means to improve their performance: They copied, for instance, the shape of the picture to be remembered with pieces of paper, or "wrote down" the first letter of its name. In doing so they proved they were able to reproduce considerably more pictures of a series than without the use of external means.

It is one thing to demonstrate that children can make use of external aids or tools to improve their performance on some task, but it is quite another thing to show the purported ontogenetic transition from problem solving based on external mediation to problem solving based on internal mediation. In Vygotsky's view such a process of internalization, or "ingrowing," indeed takes place in ontogenesis and various stages in mediation can be established. He claimed—following in the footsteps of a long tradition (e.g., Dewey, 1910)—that speech was the primary internal mediator in problem solving. But how should we conceive the transition from performance primarily based on the use of external means to performance primarily based on internal speech? Can the "ingrowing" of external mediators be experimentally demonstrated? And can different forms of mediation be shown to exist in different stages of ontogeny? It was these and similar questions that Leont'ev's (1931, 1932) experimental research tried to address.

LEONT'EV'S EXPERIMENT

Leont'ev conducted a series of experiments that were meant to demonstrate the use of external means and their consequent internalization. The method he used was that of double stimulation discussed above and his main topics were that of memory and attention. In a series of experiments with hundreds of subjects of varying age and background he attempted to show that young children—up to age 7—were unable to make use of (external) mediators. In

this sense of the word their behavior was "natural." Somewhat older children—age 8 and 13—could make successful use of external means and their mediated performance was, consequently, better than their unmediated, or "natural" performance. Finally, adults refrained from using external mediating devices turning to internal mediators instead. At least, this was what Leont'ev—following Vygotsky—concluded in the book summarizing his experiments (Leont'ev, 1931; see also Van der Veer & Valsiner, 1991).

The Forbidden Colors Task

One of Leont'ev's experiments was published in the *Journal of Genetic Psychology* and is known as the "forbidden colors game." In this experiment Leont'ev (1932) presented his subjects with three or four series of 18 questions, out of which seven concerned the color of things. In each series the 2nd, 4th, 8th, 9th, 12th, 15th, and 17th question required a color answer. The other questions were either arbitrary (e.g. "Do you like reading?"), or in some way prepared the next color question. Thus, one of Leont'ev's questions was "Have you seen the sea?", which was immediately followed by the next question "What color is the sea?" The subjects were instructed that they should answer each question promptly and in one word, especially in the case of colors. The first series of 18 questions was presented without any additional limitations, that is, the subjects were free to answer whatever they liked. In the second series the actual rules of the game were explained: The subject should (a) not repeat the name of one and the same color within one series, and (b) avoid mentioning two specified "forbidden" colors. The third series differed from the second one insofar as the subject was now given a set of nine colored cards and suggested that these might help him to accomplish the task ("they must help you to win"; Leont'ev, 1932, p. 64). A fourth series was only presented in cases where the subject did not show evidence of having found out how to use the cards, or did so only towards the end of the experiment. Apparently, before this fourth series the subjects were told explicitly how to make use of the cards. Before and after each series Leont'ev ascertained by means of special questions to what extent the subject had mastered and remembered the instructions. Before the beginning of the whole experiment the subjects were checked for color blindness and/or knowledge of color names using the set of colored cards mentioned above. From Leont'ev's account it becomes clear that for the subjects the whole setting was that of a "game" where they could either lose or win. The experimenter tried to introduce a relaxed atmosphere and linked the questions together—by means of phrases such as "Tell me!" or "What do you think?"—to stimulate a normal conversation-like situation. Also because the experimental situation was very much like a traditional game played in Russian (and Dutch) families it may be assumed that the subjects felt "at home" during the "game" and "played" at an optimal level.

The Nature of the Questions

Most of Leont'ev's questions were very simple and well within comprehension of his subjects. Examples of his questions were: "Can you draw?"; "Did you ever listen to music?"; "What color is the sea?"; "Do you like dogs?"; "Do you want to be big?", "What colors can leaves be?", and so forth. It should be remarked, however, that not all of the questions could be answered equally easily. To illustrate this, let us look at Leont'ev's second series where the forbidden colors were green and yellow. His seven color questions were, respectively: (2) "What color is your shirt?"; (4) "What color are the railway-carriages?"; (8) "What color is the floor (generally)?"; (9) "and the walls?"; (12) "What color are lilacs?"; (15) "What color can leaves be?"; (17) "What is your favorite color?"

It is clear that some of these questions—for example, 17 and, maybe, 15—leave considerable freedom to the subject, while others—for example 2 and 12—permit only one possible answer, unless the subject is prepared to give an untruthful, arbitrary answer (e.g., "Lilacs are red"). The rules of the game allow for such arbitrary answers, but it is quite likely that young subjects will find it very hard to give them (see later). Unfortunately, Leont'ev said virtually nothing of this varying level of difficulty, mentioning only (Leont'ev, 1932, p. 65) that some questions were more "provocative of error" than others. We also do not know whether in the list given above the second question was meant to provoke an error or not; that is, we do not know whether the experimenter deliberately selected some detail of the subject's clothes that has a "forbidden" color. In the following, however, we have assumed that this was not the case. We may conclude, then, that some questions were more difficult than others in the sense that they—if answered truthfully—required one of the forbidden colors. For other questions there were more degrees of freedom.

Naturally, several of Leont'ev's questions were also highly culture-bound. Thus, we would not readily ask a contemporary American child what color a streetcar has, nor whould we ask a non-Soviet child whether it ever had been at a (political) meeting. Colors of walls, floors, and exercise books also tend to vary across cultures. Consequently, as will be seen below, both Adams et al. (1987) and ourselves had to adjust some of the questions. The four series of questions were presented to the subjects in varying order (counter-balanced) to control for possible differences in difficulty.

The Colored Cards

Leont'ev provided his subjects with nine colored cards of unknown size. His colors were black, white, red, blue, yellow, green, purple (lilac), brown, and gray. The "forbidden" colors he used were green, yellow (in the second

series), blue, red (third series), black, and white (fourth series). The fact that each color was represented only once and that nine cards were available is of great significance for the strategies of card use that may be followed. Thus, quite a few children follow the simple strategy of turning upside down first the two forbidden colors and then, following each color question, successively mention one of the remaining color cards and turn it over. In this way they avoid mentioning both the forbidden colors and the repetition of any color. Having nine cards at one's disposal this strategy can always be employed as there are exactly seven color questions in each series. Unfortunately, Adams, Sciortino-Brudzynski, Bjorn, and Tharp (1987) in their replication of Leont'ev's study used only eight colored cards, which would seem to discourage the strategy just mentioned.

The fact that each color is represented only once in the set of colored cards as well as the fact that it contains all "forbidden" colors is, of course, also of significance. Interesting variations of the game and possibly other forms of card use would arise if this were not the case.

Forms of Card Use

Leont'ev (1932, p. 70) claimed that the various methods of using the cards could be reduced to two different forms. One strategy was to put the forbidden colors out of the range of vision, to exhibit the remainder, and—as the subject was answering the questions—to place the already named cards on one side. This approach was very similar to the one mentioned above, but there is one essential difference. In the strategy we described the subject can "mechanically" read off the remaining colors, which puts virtually no demands on either memory or attention processes. The only thing the subject has to do properly is to turn over the colored cards after their use. In no way, thus, does he or she have to keep track of the colors already mentioned or even to remember the forbidden colors. In the first strategy mentioned by Leont'ev the cards are not turned over—are merely put out of sight—which makes this strategy presumably slightly more difficult to follow. Leont'ev (1932, p. 70) regarded this method as "the least perfect" one, reasoning that the subjects often did not use the colored cards as a real mediating device, and only put the card aside *after* having given their answer. Putting aside the cards, then, was little more than the registering of the spontaneously given answers. Leont'ev (1932, p. 74) was inclined to interpret this strategy as a temporary phenomenon caused by the subjects's enchantment by the method: They are so impressed by the magic power of the cards that they use them irrespective of the questions asked. Such a "formalist" phase in the use of mediating devices is characteristic for both ontogeny and phylogeny Leont'ev (1932, p. 74) speculated:

> Probably it is just this phase of the domination of external psychological mediums, through which the development of the higher instrumented, "significative" acts of behavior pass, that reveals itself in the history of the cultural development of humanity, in those numerous and extremely worked-out systems of external methods of behavior which compose a typical feature of primitive society.

Some pages before Leont'ev (1932, pp. 69–70) had compared the same behavior—putting the forbidden colors out of sight—to the "way . . . an Australian or African savage might act in freeing himself from a dangerous man by destroying his image or symbol." Thus, the fascination with a specific elaborate method and the inability to see its limitations and real value would be characteristic of both children in a certain age period and of humanity in its "primitive" or "magic" period.

In the second strategy discerned by Leont'ev the cards remain in sight—the child may separate them into two rows or columns, for instance—and before each answer the child consults the available cards. In this case, then, the cards are used *before* the answer is given and the subjects' behavior can be called mediated in the real sense of the word. Undoubtedly, the second method is more difficult to follow as now the cards used so far are not simply ignored, but consulted. It is only after this consultation of both the forbidden color cards, the cards used, and the remaining cards that the subjects can give their answer. To Leont'ev this meant that the second strategy reflected a higher form of (mediated) thinking typical of a later stage of mental development.

The Subjects

Thirty subjects participated in Leont'ev's experiment of whom seven of preschool age (5 to 6 years old), fifteen of school age (seven children of 8 to 9 years old and eight children of 10 to 13 years old), and eight adults. No further information (e.g., SES, sex, ethnic background) about these subjects was provided.

The Number of Mistakes and their Interpretation

The most important quantitative results of Leont'ev's experiment were formed by the number of incorrect answers to the color questions in the various age groups. Table 11.1 summarizes his data.

Two things seem immediately evident from Table 11.1. First, the number of incorrect answers (with or without cards) seems to diminish with age. Second, performance with colored cards (series III/IV) seems better than performance without (series II). It was the second fact that was theoretically the most interesting to Leont'ev and a substantial part of his paper was devoted to its

Table 11.1. The average number of incorrect answers (after Leont'ev, 1932, p. 69)

		Mean number of errors		
Age	*N*	*series II (no cards)*	*series III/IV (cards)*	*Differences II-III/IV*
5–6	7	3.9	3.6	0.3
8–9	7	3.3	1.5	1.8
10–13	8	3.1	0.3	2.8
22–27	8	1.4	0.6	0.8

interpretation. As can be seen, the youngest children made virtually the same number of mistakes (3.9 versus 3.6 out of 7 possible mistakes) in both card (series III/IV) and non-card (II) series. For older children, however, the results were decidedly better for the card series (differences between series II and III/IV of, respectively, 1.8 and 2.8). Finally, for adults the results were still in favor of the card series, but the advantage was considerably smaller (a difference between series II and III/IV of only 0.8). Combining these results with his qualitative observations of the subjects' behavior and the answers they gave to his questions Leont'ev came to the following interpretation of his findings.

The youngest children (5–6 years old) gave incorrect answers to slightly more than half of the color questions and the difference between the series with and without cards was very small. Leont'ev explained that these children were very easily distracted from the task and did not discover by themselves how to use the cards. Even after the explicit instruction how to use them that preceeded series IV—sometimes, also, the children were allowed to watch the card use of more able peers—the children remained unable to handle the cards (Leont'ev, 1932, p. 68). At best they superficially imitated the card use of others, but their whole behavior as well as their answers to questions asked by the experimenter demonstrated a complete lack of understanding of their function. In fact, Leont'ev several times observed that the color cards actually hindered the correct performance of children as they were drawn to some cards and kept repeating their attractive colors. In general, then, the youngest children were unable to detect and/or understand the instrumental function of the colored cards. It can be concluded, therefore, that their answers were not mediated by the external stimuli provided, but formed an immediate reaction to the question or to some accidental part of the environment.

Quite another picture can be observed in the case of the older children. These children showed a considerably better result in the series with cards (series III/IV) than in that without (series II). As can be seen in Table 11.1 the differences were, respectively, 1.8 and 2.8. Leont'ev suggested that this improvement form series II to series II/IV was due to the children's better understanding of the functions of the card. The children were now increas-

ingly capable of using the cards, often detecting their correct use spontaneously, that is, not needing the additional instruction or modeling and the extra series IV. Leont'ev concluded that these children understood the instrumental function of the colored cards and that their thinking, thus, was mediated through external means.

The results of the adults were at first sight rather puzzling: Although their number of mistakes was quite low, there was virtually no difference between the card and noncard series. At any rate, the difference between the performance in series II and III/IV was considerably smaller (0.8) than the corresponding difference (1.8 or 2.8) for school children. In fact, the magnitude of the difference was rather similar to that of the preschool children. Moreover, observation pointed out that the adult subjects manipulated the colored cards far less than the school children. It would seem, then, that the behavior of adults was a unmediated or "natural" as that of these youngest children. Leont'ev, however, did not at all believe in such a peculiar form of mental regression. In stead, he argued that "the second series of stimulation (that is, the colored cards; RvdV) gets emancipated from primary external forms. What takes place is what we here call the process of 'ingrowing' of the external means: the external sign turns into an internal one" (Leont'ev, 1932, p. 76).

Leont'ev argued, thus, that the behavior of adult subjects remains mediated, but that the mediation process shifts from external to internal means. He provided two arguments to substantiate his claim. First, he pointed out that the phenomenon of internalization of external means is a very general one as in the case of children who learn to do calculations by heart only after extensive training with paper-and-pencil calculations. It, therefore, seemed unlikely that the adults in the present task would suddenly relapse into "natural," unmediated behavior. Second, several of the adults did make use of the cards, but without manipulating them. Having spread out the cards—usually with the forbidden colors in a special position—they would fixate them after each color question and then give the correct answer. To Leont'ev this strategy formed a perfect illustration of the shift to internal means: While younger children have to physically remove the forbidden colors the adults can carry out this operation "in their heads" using the cards only as a sort of reminder. The observed cases, therefore, demonstrated a mental operation that is halfway its internalization process.

Some Problems of Interpretation

Leont'ev's explanation of his data is consistent, but is not without its moot points. In the first place he claimed that older children made more use of the colored cards provided, but did not substantiate this claim with any data. We do not know, therefore, to what extent the older children actually made more use of the cards and how their use correlated with performance as measured

by the number of mistakes. In other words, it is unclear whether the performance in the series with cards was superior *because* of the card use. It might be argued, for instance that the performance in the series with cards was superior because these were always presented *after* the noncard series. The better performance would then simply reflect a training effect. This possibility is all the more critical as the scores of an unknown fraction of the children reflected their performance in series IV. These children, then, had one more series to get used to the experimental situation and the nature of the task. In this reasoning combining the scores in series III with those in series IV would inflate the difference between card and noncard performance. We have to leave open, then, the possibility that card use is simply irrelevant for performance in the forbidden color task. In fact, it can be argued that such a possibility would be completely consistent with a Vygotskian understanding of mental development (see later). How, then, would a possible alternative explanation of Leont'ev's set of data look like? First, the differences between card and noncard series within one age group would be explained by increasing familiarity with the nature of the task and the experimental situation. Second, the differences between age groups would be explained by the increasing availability of cognitive strategies that enable correct performance of the task. Third, the small difference between the performance in card and noncard series in the adult group would be due to a "ceiling-effect": Performance is already that high in the noncard series that it can hardly be improved. Fourth, the small difference between the performance in noncard and card series in the group of youngest children would be due to their general lack of understanding of the task (which can be improved by no amount of training).

Summarizing, it would seem that there are several debatable points in Leont'ev's interpretation of his data and not a few shortcomings in the design of his study. Several of these shortcomings have been pointed out by Adams et al. (1987), who attempted to replicate Leont'ev's study making use of contemporary methods of data analysis and abiding contemporary standards of scientific rigor. It is to this study that we will now turn.

THE ADAMS ET AL. (1987) REPLICATION

Adams et al. (1987)[1] voiced serious doubts concerning the design and method of Leont'ev's experiment. First, they objected to the retesting of subjects who

[1]Adams et al. consistently refer to Vygotsky (1932), or Luria and Vygotsky (1932), when referring to Leont'ev's investigation. This mistake is probably due to the fact that Luria and Vygotsky were mentioned in the *Journal of Genetic Psychology* as guest editors of a series of publications conducted in the Krupskaja Academy of Communist Education. There is no doubt, however, that Leont'ev was the main experimenter in this and several other experiments.

had difficulty mastering the use of the cards—in series IV—and to the combination of data from series III and IV. Even more unsettling they found the fact that sometimes before this fourth series the children were allowed to watch the card use of more successful children. Second, they found fault with the small and unequal number of subjects that precluded the possibility of proper subject assignment and counterbalancing of conditions (Adams et al., 1987, p. 5). Third, Adams et al. criticized the unclear presentation of the structure of the sentence sets.

In order to correct for these methodological problems they used substantially greater numbers of subjects (counterbalanced by gender), parallel forms of the task stressing either the repetition or forbidden color rules, equivalent test procedures for all subjects on all four series of the test, and sentence sets revised to be structurally equivalent and culturally appropriate for their sample.

The Method Used by Adams et al. (1987)

Subjects Ninety-six subjects, 24 (12 male and 12 female) from each of the four age groups used in Leont'ev's study participated in the experiment: 5–6 years (range 4;10–6;11, M = 5;10), 8–9 years (range 8;01–9;09, M = 8;11), 10–13 years (range 10;0–12;4, M = 10;11), and 22–27 years (range 22;0–27;05, M = 23;08). Subjects represented the ethnically divers population of Hawaii.

Procedure The same as in Leont'ev (1932), with the exception that all subjects were administered the fourth series. Before the third series the subjects were given the colored cards and "the option of using them" (Adams et al., 1987). Thus, no explicit instruction as to their possible use was given. Observable use of the cards was recorded.

Materials. Eight 4″x6″ index cards covered with colored instruction paper were used. The colors were red, blue, yellow, green, brown, white, black, and orange. The question sets given by Leont'ev (1932) were modified so as to ensure parallel structure and culturally appropriate content. Thus, each sentence set consisted of 18 questions, seven of which were appropriately answered with a color term. The structure of each sentence set was as follows: questions 2, 4, and 15 required a forbidden color; question 8 was most likely to elicit a nonforbidden color, while question 9 provoked its repetition; question 12 was again most likely to elicit a nonforbidden color; finally, question 17 was open-ended and, therefore, did not necessarily evoke an error (e.g., "What color can a car be?"). All in all, then, questions 2, 4, 9, and 15 were "difficult," questions 8 and 12 "easy," whereas question 17 was of

intermediate difficulty. A total of four to five errors, therefore, could be expected in case the child did not grasp the nature of the task.

The Results of Adams et al. (1987)

Errors Table 11.2 presents the mean number of errors found by Adams et al. in the same format as Leont'ev's original report.

It can be seen in Table 11.2 that Adams et al., contrary to expectation, found rather small differences between the card (III and IV) and noncard series (II). In fact, only one of the differences (between series III and IV for 8–9-year-old children) proved significant. Because the reduction in the mean number of errors was generally smaller than in Leont'ev's sample (and because the results were not statistically significant) Adams et al. (1987, p. 8) concluded that "this pattern of data clearly indicates a failure to replicate the major findings of the forbidden color experiment." It should be remarked, however, that the general pattern of Leont'ev's findings—no improvement whatsoever from noncard to card series for very young children and adults, plus improvement for the two groups of school children—was found again in the Adams et al. sample.

A remarkable difference between Leont'ev's and Adams et al.'s findings was that the contemporary subjects—except for the youngest age group—performed rather better in the noncard series. According to the authors this initial advantage was lost during the subsequent card series, because most subjects made little if any use of the cards.

Statistical analysis confirmed what seemed obvious from the simple visual inspection of Table 11.2. A six-way ANOVA with three between-subjects variables (age, sex, form) and three within-subjects variables (series 2, 3, and 4) found no significant effects of sex, or test form, but did reveal significant effects of age $F(1, 3) = 45.55$, $p < .001$, and task series, $F(1, 2) = 6.33$, $p < .05$. A subsequent four-way ANOVA with one between-subjects variable (age) and three within-subjects variables (series 2, 3, and 4) found only one

Table 11.2. The average number of incorrect answers (after Adams et al., 1987, p. 21)

		Mean number of errors				
Age	*N*	*series II (no cards)*	*series III (cards 1)*	*series IV (cards 2)*	*Differences II-III*	*III-IV*
5–6	24	4.2	4.2	4.2	0	0
8–9	24	2.6	2.1	1.2	.4	.9*
10–13	24	1.7	1.2	1.1	.5	.1
22–27	24	1.0	.8	1.2	.2	−.4

* $p < .001$

statistically significant contrast, for 8 and 9-year-olds between the third and fourth series, $F(1, 95) = 15.50$, $p < .001$.

Card use Leont'ev argued that the subjects' improved performance in series III/IV was due to their performance becoming mediated by the colored cards. Unfortunately, he gave no numbers to indicate to what extent the subjects actually made use of the cards. To remedy for this shortcoming Adams et al. registered card use, distinguishing "consistent," "inconsistent," and no card use. The card use was called consistent when the subjects were observed to or reported to use the cards six or seven of the seven possible times within one series. It was called inconsistent when the cards were used from one to five times within a series. Table 11.3 summarizes Adams et al.'s findings.

The set of results in Table 11.3 partially confirms Leont'ev's rather vague account. Thus, it can be seen that the youngest children hardly made use of the cards, while card use increased with age. It is true that the rather high percentage of card use by adults forms a problem. Adams et al. take this card use as contradicting Leont'ev's theoretical argument, but this is not altogether clear. According to Leont'ev the *manipulation* of cards would be clearly lower for adult subjects, but he did allow for card use in the form of visual fixation of the (forbidden) color cards (see earlier). As Adams et al. (1987, p. 9) in Table 11.3 combined both observations of card use (presumably mostly manipulations) and self-reports by subjects (which might include use by visual fixation) it is unclear what the percentages for card use exactly mean: It might be argued that adults consistently relied more on visual fixation, while younger children relied on manipulation. This would clearly be compatible with Leont'ev's line of reasoning. The data obtained for school children also

Table 11.3. Percentage of subjects using cards (after Adams et al. 1987, p. 22)

	Pattern of card use		
Age	*None*	*Inconsistent*	*Consistent*
5–6			
series III	83.3	12.5	4.2
series IV	83.3	8.3	8.3
8–9			
series III	45.8	29.2	25
series IV	25	29.2	45.8
10–13			
series III	58.3	16.7	25
series IV	50	16.7	33.3
22–27			
series III	58.3	33.3	8.3
series IV	58.3	37.5	4.2

partially confirmed Leont'ev's findings: both 8–9-year-olds and 10–13-year-olds made ample use of the colored cards. However, this card use was not clearly related with the number of errors made. Thus, combining the findings of Table 11.2 and 11.3 one finds that the substantial card use found for 8–9-year-olds in series III does not lead to an improvement of performance (as compared with series II) in terms of the mean number of errors. Neither was this the case for 10–13-year-olds. Only the increasing card use from series III to IV for 8–9-year-olds corresponded with a statistically significant decrease of the number of errors.

It would seem, then, the pattern of card use found in Adams et al.'s replication partially confirmed Leont'ev's findings, but that it remained far from proven that the use of cards actually contributed to the improvement of performance. Alternative explanations of the set of data, such as the one we provided earlier, cannot be ruled out.

Verbal strategies One reason—other than increased card use—older children may do better than younger ones is that they can circumvent the forbidden color names making use of various sorts of verbal strategies. In order to investigate this possibility Adams et al. (1987) classified the subjects' answers into various categories. Responses were categorized as "acceptable" when they were truthful and stereotypic, and did not violate either of the rules (e.g., "What color is the sea?" "Blue"). An answer was coded "possible" if the color was theoretically possible, but deviated from the accepted cliché (e.g., "What color is the sea?" "Green"). "Arbitrary" answers were answers that did not violate the rules, but were not truthful (e.g., "What color is the sea? "Red"). "Compound" answers were those in which the subject used a modifier such as "light-" in "light-blue" to give a truthful answer. "Related" answers were those in which the subject offered the name of a color related to the truthful response (often a seldom used synonym, such as "azure" for "blue"). Using this classification Adams et al. (1987) found that older subjects make increasing use of the categories other than "acceptable." Consequently, they suggested that age-related improvements in the number of correct answers rest on the subjects' willingness to give patently false ("arbitrary"), or somewhat dubious ("possible") answers, and on the growth of their color vocabulary ("compound" and "related" answers). Adams et al. (1987) did not find any relation between card use and the use of the various response categories.

Adams et al. (1987) concluded that Leont'ev's claims regarding the effectiveness of the colored cards were not supported by their data. It is true that school children made increasing use of the cards, but this use did not demonstrably cause the improved performance. The authors suggested that exposition of the subjects to a model of effective card use (which occasionally took place in the original investigation, see Leont'ev, 1932, p. 68) could

possibly lead to more effective card use and a replication of Leont'ev's results. Adams et al. (1987, p. 15) suggested that a literal notion of "ingrowing," or internalization of auxiliary means is not very fruitful. Instead, they argued that subjects leave the external mediating devices such as cards for subtle verbal strategies such as giving "false" (arbitrary, possible) answers that do not strictly violate the rules of the game and/or making use of words that circumvent the naming of forbidden colors (compound and related answers).

THE PRESENT REPLICATION

Adams et al. (1987) introduced a number of notions and methodological refinements that have proved valuable for the present replication. At the same time it seems clear that their investigation suffered from a number of shortcomings. First, as they themselves have pointed out, the role of explicit instruction or modeling of possible card use was not taken into account. Thus, while Leont'ev (1932) sometimes—in case of inadequate performance—provided explicit instruction or modeling before series IV, Adams et al. (1987) only handed the cards to the subjects with the remark that they might help. The fact that the difference between card and noncard series was less in their sample might be due to this factor. Second, Adams et al. (1987) made no distinction between different forms of card use—such as manipulation and visual fixation—and, consequently could not check Leont'ev's claims regarding age-related changes in this respect. Third, the authors—like Leont'ev (1932)—did not control for or even discuss the possibility that superior performance in later series was due a to a simple training effect (rather than increased card use or superior verbal strategies). While this may be due to the fact that they generally found little or no improvement over series it seems necessary to discount a simple training effect. Fourth, the authors provided no more than eight color cards, which may have influenced their results (see above). In order to clarify these issues the present author conducted a new replication of Leont'ev's experiment.

Method

Subjects Eighty subjects, twenty (10 male and 10 female) from each age group participated in the experiment: 5–6 years (range 5;0–6;10, M = 5;09), 8–9 years (range 8;02–9;10, M = 9;0), 10–13 years (range 10;03–12;0, M = 11;02), and 21–30 years (range 21;07–30;08, M = 25;09). Subjects represented the ethnically homogeneous population of the Netherlands.

Procedure. The procedure was basically similar to that followed by Leont'ev (1932) and Adams et al. (1987), but the number of series and the

instruction differed. Each subject participated in six series of questions. Before the third series the colored cards were introduced with the remark that "these might help in winning the game." Before the fifth series the children were explicitly told how they might use the colored cards and two ways of using them were shown (the experimenter explained that the subject might put the forbidden and/or already used cards to one side or turn them over and illustrated this). In addition, the subjects were told that it was allowed to give "funny" answers, like saying that the sun is blue. In this way, the instruction both demonstrated ways of using the cards and hinted at the use of a verbal strategy. Thus, each subject participated in the following series (see Table 11.4).

Ideally, the results of our series 2 (the noncard series) for each age group should roughly coincide with the corresponding series of Leont'ev (1932) and Adams et al. (1987). Also, the results of series 3 and 4 should be roughly equivalent to those of the corresponding series of Adams et al. (1987), as they simply introduced the colored cards without further instruction or modeling. Finally, series 5 and 6 were added to investigate whether explicit instruction or modeling makes a difference in performance and to check whether this factor might explain the differences between the results found by Leont'ev (1932) and Adams et al. (1987). Thus, the results of our series 5 and 6 should be more similar to the data found by Leont'ev (1932) as he gave some subjects explicit instruction and others not. Of course, in general one would expect a diminishing number of errors over series due to the influence of training.

Materials

The structure of the sentence sets was identical with those of Adams et al. (1987). Nine (instead of eight) colored cards were used (see above).

Table 11.4. The six series of the experiment with the type of intervention by the experimenter

Series 1 (no rules)
INTRODUCTION OF RULES OF THE GAME
Series 2
INTRODUCTION OF CARDS
Series 3
Series 4
EXPLICIT INSTRUCTION
Series 5
Series 6

Results

Errors. The quantative results of the experiment concern the number of errors in the different series, the incidence and frequency of card use, and the distribution of the various response categories. Table 11.5 presents the mean number of errors found in the present replication.

These results are also depicted in Figure 11.1 One sees how the number of errors decreased gradually (virtually linearly for the first three series) with age. A clear gap separates the scores for series 2, 3, and 4 from those of series 5 and 6, at least for the younger age groups. The gap diminishes due to a ceiling effect for 11- to 13-year-olds and adults.

A first visual inspection of Table 11.5 and Figure 11.1 gives rise to the following simple and obvious interpretations. First, in this sample the age effect observed by both Leont'ev (1932) and Adams et al. (1987) is found again: In all but the very last series the mean number of errors diminishes over age. Second, there seems to be only a slight effect of the introduction of the colored cards per se. This would confirm the result found by Adams et al. (1987). Third, the explicit instruction caused a dramatic decrease of the number of errors. This finding most probably corroborates Leont'ev's (1932) results.

Statistical analysis confirmed these first impressions. A MANOVA for repeated measurements found significant effects for both series, $F(73;4) = 14.7$ and $p < .000$, and the interaction between series and age groups, $F(193;12) = 2.0$ and $p = .03$ (see Table 11.6). On further univariate analysis (mixed model) it was found that the contrast between series 2 and all other series was significant for all age groups. The contrast between series 3 and 4 versus series 5 and 6 was significant for all age groups, but the 10- to 13-year-olds. This contrast was most pronounced for 8- and 9-year-olds, and smallest for adults.

Card use Following Adams et al. (1987) we observed the card use of subjects and distinguished between consistent and inconsistent card use. It was found—in particular for the lower age groups—that the introduction of

Table 11.5. The average number of incorrect answers for different series and age groups

		Mean number of errors				
	CARDS	INSTRUCTION				
Age	N	II	III	IV	V	VI
5-6	20	3.4	3.2	3.4	1.9	1.7
8-9	20	2.6	2.4	2.1	.9	.3
10-13	20	2.2	1.6	1.4	.7	.5
21-30	20	1.6	1.1	.9	.7	.6

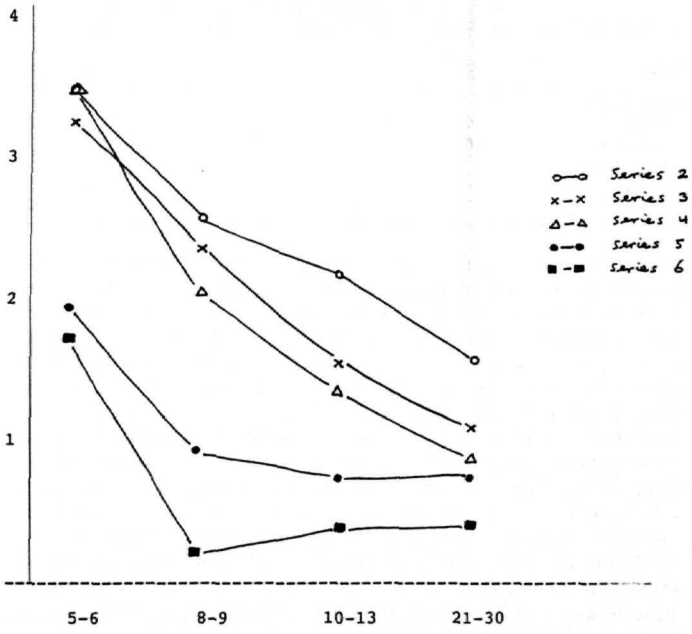


Figure 11.1. The number of errors in the five series for all age groups

the colored cards as such did not lead to their frequent use. The explicit instruction, however, proved highly effective in promoting the use of the cards enhancing the percentage of children using them from practically zero in the lower age groups to a percentage of 50 or more and even raising the percentage of adults using the cards from approximately 40 to 55. Table 11.7 summarizes the results.

Table 11.7 demonstrates that explicit instruction was very effective in enhancing card use: In all age groups there was a dramatic increase in either inconsistent or consistent card use. The effect was smallest for adults as they already before the instruction tended to make use of the cards. It would thus seem that the decreased number of errors is due to the increased frequency of card use. Unlike Adams et al. (1987), we decided to check whether consistent card use was actually related to the number of errors made. The frequency of card use was measured in two ways. Two observers scored either (a) "card manipulation," when subjects touched or manipulated the colored cards, or (b) "card reference" when subjects either looked at the colored cards, or

Table 11.6. Results of MANOVA for present replication

Source	*SS*	*df*	*27 MS*	*F*	*p*
Between Subjects					
Error	290	76	4		
Age groups	168	3	56	14.7	<.000
Within subjects					
Series	.42	73;4		25.2	<.000
Series x Groups	.73	193;12		2.0	.03

touched and manipulated them. The latter variable, thus, measured a broader concept of card use and included the first. These variables were correlated with the variable "error change," which was computed by substracting the number of errors made in series 5 plus 6 from those made in series 3 plus 4. This variable, thus, reflected the effect of instruction on the error rate. It was found that the correlations between "error change" and either "card manipulation" or "card reference" were extremely low and nonsignificant for the group of subjects as a whole. Thus, for the group of all subjects the increasing card use is not related to the decreasing number of errors. Computing these same correlations for each age group separately it was found that only for the group of 8- and 9-year-olds "error change" and the two forms of card use were correlated significantly (card reference, $r = .42, p = .03$; card manipulation, $r = .40, p = .04$). This finding suggests that, at least for the other age groups, factors other than card use—such as the subjects' use of various verbal strategies—are responsible for the increased performance after instruction.

Type of responses. Looking at the different response categories introduced by Adams et al. (1987) we found that the frequency of arbitrary answers increased remarkably after instruction. Other types of responses—except for simple "acceptable" answers—were relatively infrequent. Computing the correlation between "error change" and the frequency of arbitrary answers it was found that this variable correlated negatively with the number of arbitrary answers given in the preinstruction series ($r = -.39, p = .000$) and positively with those given in the postinstruction series ($r = .33, p = .002$). This pattern was found in every age group with the exception of adults where the number of arbitrary answers given in the pre-instruction series was not (negatively) related to "error change," presumably because adults already gave a substantial number of arbitrary answers before instruction. Thus, in general there is a strong relation between instruction and the frequency of arbitrary answers: after instruction there is a marked (and statistically significant) increase in arbitrary answers.

Table 11.7. Percentage of subjects using cards before and after instruction

	Pattern of card use		
Age	*None*	*Inconsistent*	*Consistent*
5–6			
Series III	90	10	—
Series IV	70	30	—
INSTRUCTION			
Series V	10	40	50
Series VI	10	40	50
8–9			
Series III	75	25	—
Series IV	60	35	5
INSTRUCTION			
Series V	—	25	75
Series VI	—	15	85
10–13			
Series III	45	40	15
Series IV	45	20	35
INSTRUCTION			
Series V	—	40	60
Series VI	—	15	85
21–30			
Series III	25	35	40
Series IV	20	35	45
INSTRUCTION			
Series V	5	30	65
Series VI	5	40	55

GENERAL DISCUSSION

The most remarkable finding of our replication study is, we think, that card use is probably not very effective in diminishing the number of errors made. More important seems the use of various verbal strategies, such as the idea of giving arbitrary answers. At any rate, when the instruction involves both the explanation of possible ways of card use and the idea of giving arbitrary answers then the children and adults pick up both elements, but the strategy of giving arbitrary answers is far more effective in diminishing the number of errors. The one exception is the age group of 8–9-year-olds were increased card use is associated with improved performance. This finding is consistent with the results obtained by Adams et al. (1987) and contradicts the expectations of Vygotsky and Leont'ev. In contradiction to Vygotsky and Leont'ev we had to conclude that in the experimental setting card use did not substantially improve performance and that adults did not refrain from using the cards.

Thus, the results of the forbidden color game constituted no proof for the idea that children as they grow older internalize external means, such as colored cards.

The fact that the age group of 8–9-year-olds forms somewhat of an exception confirms Leont'ev's (1932) results. These children after instruction extensively used the colored cards and this use seemed associated with improved performance. We might say, then, following Leont'ev, that children of this age group are particularly fascinated by the use of external means and that their performance tends to be dominated by them.

Adequate performance in the forbidden color game does seems to be related to the ability to make use of various verbal strategies such as the willingness to give "odd," or arbitrary answers. Of course, the possibility of giving arbitrary answers was pointed out to the children—as were two ways of using the colored cards—in the explicit instruction before the fifth series. But the fact, that virtually all children proved able to make use of this strategy points our that children of this age can use various internal means and do not have to rely on external mediating devices.

The original findings of Leont'ev (1932) proved very little. The number of subjects was very low and the design and analysis of his study was not in accordance with contemporary standards of scientific rigor. As a result his findings can easily be interpreted in different ways. The fact, for example, that for adults performance with cards is hardly better than performance without cards can be interpreted as a ceiling effect and does not constitute proof for the internalization hypothesis. A similar ceiling effect was found in our own replication study.

The forbidden colors game formed too crude a test of the internalization hypothesis. It is probably not a very fruitful idea to think that all internal mediating devices necessarily go through an external material phase. That is, one may posit that all internal means originate in the external world, but this does not necessarily imply that they originate as material objects, such as colored cards. The design of the forbidden colors game followed the model of the "Marxist" idea of internalization of material tools, whereas a "Baldwinian" idea of internalization of social conduct and social relations seems more fruitful (Valsiner & Van der Veer, 1988; Van der Veer & Valsiner, 1988). Moreover, the use of a cross-sectional design for the study of a developmental process is unfortunate and made any conclusion about ontogenetical shifts in type of mediation stand on rather weak grounds. Finally, the increase of internal mediation was inferred by Leont'ev on the grounds that external mediation decreased rather than proven by any positive data.

The forbidden colors game was also unfortunate in so far as it suggested that for each specific task one can discern periods of, respectively, no mediation, external mediation, and internal mediation. In reality, it would seem that children confront each new task armed with a whole repertoire of

internal mediating devices that proved effective in solving other tasks. At any rate, it seems a little naive to aspect that children of up to 13 years old had to rely on the external mediating device of colored cards.

It would thus seem that there were good theoretical grounds to doubt Leont'ev's interpretation of his data. The empirical replications by Adams et al. (1987) and ourselves have strengthened these doubts by suggesting that for most subjects the use of colored cards is not effective in task performance and, consequently, Leont'ev was not even studying external mediation. If we accept this empirical finding, then Leont'ev's whole argument about the decrease of external mediation—even if we neglect its theoretical inadequacy—becomes simply irrelevant.

We may conclude, then, that Leont'ev's (1932) original study formed no proof or argument for the internalization hypothesis in any form. The subsequent replication studies by Adams et al. (1987) and ourselves suggest that for the majority of subjects the colored cards did not mediate performance at all and that other, internal, mediating devices play a decisive role in solving the forbidden colors task. It still remains to be seen in what way these internal mediating devices are mastered by the child.

REFERENCES

Adams, A. K., Sciortino-Brudzynski, A. P. Bjorn, K. M., & Tharp, R. G. (1987). *Forbidden colors. Vygotsky's experiment revisited.* Paper presented at the SRCD meeting in Baltimore, MD.

Brainard, P. P. (1930). The mentality of a child compared with that of apes. *Journal of Genetic Psychology, 37,* 268–292.

Bruner, J. (1984). Vygotsky's zone of proximal development: The hidden agenda. In B. Rogoff & J. V. Wertsch (Eds.), *Children's learning in the "Zone of proximal development"* (pp. 93–97). San Francisco: Jossey-Bass.

Bühler, K. (1919). *Abriss der geistigen Entwicklung des Kindes.* Jena: Verlag von Gustav Fischer.

Dewey, J. (1910). How we think. In J. A. Boydston (Ed.), *The middle works of John Dewey. 1899-1924* (Vol. 6, pp. 177–355). Carbondale and Edwardsville: Southern Illinois University Press.

Guillaume, P., & Meyerson, I. (1930). Recherches sur l'usage de l'instrument chez les singes I. Le problem du détour. *Journal de Psychologie, 27,* 177–236.

Guillaume, P., & Meyerson, I. (1931). Recherches sur l'usage de l'instrument chez les singes II. L'intermédiaire lié à l'objet. *Journal de Psychologie, 28,* 481– 555.

Köhler, W. (1921). *Intelligenzprüfungen an Menschenaffen.* Berlin: Julius Springer.

Leont'ev, A. N. (1931). *Razvitie pamjati. Eksperimental'noe issledovanie vysshikh psikhologicheskikh funkcij.* Moscow-Leningrad: Uchpedgiz.

Leont'ev, A. N. (1932). The development of voluntary attention in the child. *Journal of Genetic Psychology, 40,* 52–81.

Valsiner, J., & Van der Veer, R. (1988). Lev Vygotsky and Pierre Janet. On the origin of the concept of sociogenesis. *Development Review, 8,* 52–65.

van der Veer, R., & Valsiner, J. (1988). On the social nature of human cognition. *Journal for the Theory of the Behavioral Sciences,* 18, 177–135.

van der Veer, R., & Valsiner, J. (1991). *Understanding Vygotsky. A quest for synthesis.* Oxford:

Basil Blackwell.
Vygotsky, L. S. (1928). Problema kul'turnogo razvitija rebenka. *Pedologija, 1,* 58–77.
Vygotsky, L. S. (1929a). The problem of the cultural development of the child. *Journal of Genetic Psychology, 36,* 415–434.
Vygotsky, L. S. (1929b). K. voprosu ob intellekte antropoidov v svjazi s rabotami V. Kolera. *Estestvoznanie i marksizm, 2,* 131–153.
Vygotsky, L. S. (1930a). Foreword. In W. Köhler. *Issledovanie intellekta chelovekopodobnykh obez'jan* (pp. I–XXIX). Moscow: Izdatel'stvo Kommunisticheskoj Akademii.
Vygotsky, L. S. (1930b). Socialisticheskaja peredelka cheloveka. *VARNITSO, 3,* 36–44.
Vygotsky, L. S. (1983). Istorija razvitija vysshikh psikhicheskikh funkcij. In L. S. Vygotsky, *Sobranie Sochinenij. Vol. 3. Problemy Razvitija Psikhiki* (pp. 5– 220). Moscow: Pedagogika. (Original work published 1931)
Vygotsky, L. S., & Luria, A. R. (1930). *Etjudy po istorii povedenija. Obez'jana. Primitiv. Rebenok.* Moscow-Leningrad: Gosudarstvennoe Iz datel'stvo
Wertsch, J. V. (1985). *Vygotsky and the social formation of mind.* Cambridge, MA: Harvard University Press.

CHAPTER 12

Does Maternal Responsiveness Increase Infant Crying?*

Replication of the Baltimore Study

Frans O. A. Hubbard and Marinus H. Van IJzendoorn
Center for Child and Family Studies
Leiden University
The Netherlands

Viewed from the perspective of attachment theory, crying is attributed the function of maintaining the baby's proximity to protective caregivers. Bowlby (1971) assumed that crying belonged to the class of preattachment behaviors, such as sucking and smiling, that served the human species in its evolutionary struggle for life. In this cybernetic model of attachment theory it is plausible to consider that crying is decreased by a prompt response from the caregiver, rather than being reinforced (Ainsworth & Bell, 1977; Bell & Ainsworth, 1972). If crying is considered as directed toward the goal of maintaining proximity to the protective caregiver, reaching that goal should be sufficient "reason" to terminate the crying behavior (Ainsworth, Blehar, Waters, & Wall, 1978). Infants are seen as gradually deriving a feeling of competence and mastery over the environment as a result of their perception of a prompt reaction to their crying. If infants really experience danger they anticipate short latencies in the caregiver's response. In mild distress they do not feel urged to alarm the environment instantly by crying (Bell & Ainsworth, 1972); therefore, during the last quarters of the first year of life crying

*Parts of this chapter were presented at the International Conference on Infant Studies, Washington, DC, April 1988, and at the Third International Workshop on Infant Cry Research, July 13–18, 1989, Hanasaari, Espoo, Finland. The research was supported by a grant from the National Foundation for Preventive Studies (Praeventie Fonds, Grant No. 28-859) to Marinus H. van IJzendoorn and Louis W. C. Tavecchio. We would like to thank Louis Tavecchio for his kind support and Agnes van Busschbach, Annet Foortse, Karin Haak, Gidia Jacobs, Ester Kloppers, Johan van Mill, Ina van den Oever, Letty van Oyen, Hansje Planjer, Inge Ruhaak, Marianne Thomeer, Nico Vesseur, Yvonne ten Wolde, André Wortelboer, Jeroen Zomerplaag, for their assistance in collecting, coding, and analyzing data.

behavior would decrease because of the responsive reactions of the caregiver.

The Original Study

Bell and Ainsworth's (1972) ecologically impressive and unique longitudinal study was intended to explore the idea that crying behavior was reduced by prompt reactions from the caregiver. Twenty-six middle-class mother–infant dyads were visited at home for about four hours every three weeks, during the first year of the baby's life. Although observing the mother–infant interactions, observers took notes which were later written out as a narrative account. In the narrative accounts, frequency and duration of all cry events during the home visit were coded, as were the maternal interventions. The frequency and duration measures for crying behavior contained fussing as well as long and intense crying. All measures were aggregated over three monthly intervals. Bell and Ainsworth (1972) concluded that the main results confirmed the hypothesized relation between maternal unresponsiveness and infant crying. First, crying behavior in the first half-year appeared much less stable than maternal unresponsiveness. This would exclude a purely constitutional explanation (e.g., irritability) for the differences in crying behavior at the end of the first year. Second, within-quarter correlations between crying and unresponsiveness indicated a tendency to more frequent and longer periods of crying by infants whose mothers were less responsive. Third, across-quarter correlations indicated that less responsive reactions to crying in the first quarter corresponded with more crying in a later quarter.

Controversies about the Original Study

Although the study seemed to have produced clear-cut results, it has also been criticized sharply on technical grounds, especially by Gewirtz and Boyd (1977a, 1977b). They stated, for example, that in computing the correlations across quarters, important concurrent and antecedent variables were not controlled. In their opinion it was necessary to control for unresponsiveness in the same quarter and infant crying in an earlier quarter in computing the correlations between crying in a later quarter and unresponsiveness in an earlier quarter. Data would allow the use of parametric methods, and thus for partializing variables. Ainsworth and Bell (1977), however, considered their sample too small to allow for parametric analyses such as cross-lag panel analysis or regression techniques. Furthermore, Gewirtz and Boyd (1977a) stated that there was an artificial dependence between infant and mother variables. A mother could neither respond nor ignore her infant's crying until the infant actually cried. Ainsworth and Bell (1977) agreed that within-quarter correlations between crying and unresponsiveness inevitably would be in-

flated because of confounded measures. Gewirtz and Boyd (1977a) also showed that separate analyses on inherently correlated frequency and duration measures for crying and unresponsiveness must have inflated in an unknown way the type I error probabilities. Later findings led Ainsworth and Bell (1977) to attach much more importance to the durational measures than to the frequency measures, because anxiously and securely attached dyads appeared to differ significantly only in the former measures. Lastly, Gewirtz and Boyd (1977a) pointed out that maternal responsiveness is not necessarily the inverse of maternal unresponsiveness as operationalized by Bell and Ainsworth (1972), although their conclusions appeared to imply this inverse relation. Ainsworth and Bell (1977) conceded that Gewirtz and Boyd (1977a) were technically correct, but they emphasized the psychological validity of considering maternal responsiveness the inverse of unresponsiveness. Ainsworth and Bell (1977) stated that the criticisms by Gewirtz and Boyd (1977) could be reduced to a claim of "not proven." They were confident that their conclusions would find confirmation in future research.

Existing Replication Effects

Because of the absence of a strong replication tradition in attachment research in general (van IJzendoorn & Tavecchio, 1987), and because of the very time-consuming nature of the data collection and coding involved in this specific study, few studies have been carried out to replicate even part of Bell and Ainsworth's provocative findings. First, Belsky, Rovine, and Taylor (1984), observing 60 mother-infant dyads for 45-minute periods in the first, third, and ninth month after birth, found that mothers engaging more in reciprocal interaction had babies who cried less later on in their first year of life. Maternal (un-)responsiveness to infant crying, however, was not implied in the reciprocal interaction measure. Second, Crockenberg and Smith's (1982) study involving 54 mother–infant dyads and covering the first three months showed that first-born children and children whose mothers were less responsive in their expressed attitudes fussed and cried more at three months. Maternal responsive behavior, however, did not predict infant crying. In a follow-up study, Crockenberg and McCluskey (1986) found a significantly positive relation between maternal responsiveness at three months and separation crying during the Strange Situation at 12 months, indicating that babies whose mothers had responded less quickly cried longer. It is not clear, however, whether crying during the 3-min separation episodes of the Strange Situation is a valid indicator of crying at home. Finally, Grossmann, Grossmann, Spangler, Suess, and Unzner (1985) observing 49 families for about 1 hour in the second, sixth, and tenth month after the birth of the baby, rated the mothers on a responsiveness scale and found that the more responsive the mothers were, the less their babies cried. This relation, however, is

based only upon tenth-month observations, and interquarter correlations to determine the influence of maternal responsiveness in an earlier quarter on crying behavior later on, were not reported.

Rationale for the Present Replication

We must, therefore, conclude that satisfactory empirical evidence in support of the hypothesized relation between maternal unresponsiveness and infant crying has still to be provided. We report here on a replication study in which some of the technical criticisms raised against the original study were taken into account. First, we present data on a larger sample ($N = 50$) and demonstrate the applicability of parametric analyses to control for antecedent and concurrent variables (Gewirtz & Boyd, 1977a). Second, we only analyze durational measures of infant crying and maternal unresponsiveness, thereby preventing inflation of type I error probabilities (Gewirtz & Boyd, 1977a). Third, because maternal responsiveness indeed is not necessarily inversely related to the durational measure of unresponsiveness (Gewirtz & Boyd, 1977a), we interpret the outcome of this measure only in terms of unresponsiveness. We also use a proportional measure of unresponsiveness, which is a priori the inverse of responsiveness. Lastly, we introduce a new measure for infant crying—"development of infant crying"—which appears not only to be a more valid indicator of the process Bell and Ainsworth (1972) tried to describe but also leads to much less inflated concurrent correlations with unresponsiveness. We leave aside, however, Gewirtz and Boyd's (1977a) criticisms that are based on theoretical considerations from a conditioning perspective. We do not think that studies carried out under an ethological attachment paradigm could be relevant to operant-learning theory (Gewirtz & Boyd, 1977a, 1977b). Bell and Ainsworth's (1972) design, as well as this report of our critical replication, are not intended to demonstrate presence or absence of conditioning contingencies between maternal unresponsiveness and infant crying, because data summarized in rather molar variables preclude functional analyses required by the experimental conditioning paradigm.

METHOD

Subjects

Given the limitations imposed by the restricted availability of four technical equipment sets (event recorder/F.A. audio registration unit) we were obliged to collect data from late summer 1983 to late summer 1984 for the first part of the sample ($n = 28$; a preliminary report on the first part of the study can be

found in Hubbard and Van IJzendoorn, 1987) and from fall 1986 to fall 1987 for the second part of the sample (n = 25). Due to the malfunctioning of technical equipment, three dyads had to be removed from our sample (N = 50). The sample consisted of original Dutch (i.e. nonimmigrant) families living either in, or in the neighborhood of, two Dutch cities, The Hague and Leiden. Eighty percent of the families were located through midwives (private practice), and 20 percent through municipal birth records. All infants were normal, healthy, full-term deliveries (with the exception of one caesarian delivery). The age of the babies ranged from 3 weeks (first observation) to 36 weeks (last observation). Twenty-six of the babies were boys; 24 were girls. The mean educational level for the mothers was 5.3 (SD = 2.4) on a scale ranging from 1 (six years of schooling) to 9 (16 years of schooling). The mean educational level of the fathers was 5.7 (SD = 2.5). The sample could be described as representative of young lower to middle-class families with two parents in which parental roles were traditionally allocated between spouses.

Procedure

All subjects were visited 12 times at home at 3-week intervals during the first nine months. The hypotheses of the study were not revealed. Visits were scheduled at the mother's convenience, the only restriction being that morning, afternoon, and evening observations were needed (unfixed order) to get a representative sample of the baby's crying behavior. Mother–infant pairs were visited regularly by one female observer, except for an occasional joint visit made by two observers for reliability checks. The last five visits on seven families were made by the first author. Visits lasted for 2 hours (first quarter) to 4 hours or more (third quarter). Observers planned to arrive given the latest expectation of the mother half an hour before the baby awoke, to install the technical equipment in a standby operating mode. Usually there was ample opportunity for interviews after the observation period. The observation period started when the baby awoke and finished when the baby fell asleep (or after 3 hours if the baby was longer awake). During the observation period the observer was obliged to play a low profile, semiparticipant role so as to be able to attend continuously to the ongoing stream of behavior recorded on the event recorder. For each quarter data of four successive observation periods was used to compute measures. The mean total observation period for the first quarter was 5.8 hours (SD = 1.6); second quarter: 7.1 (SD = 1.9); third quarter: 8.8 (SD = 2.1).

In order to cope adequately with the complex observational situation, the use of technical equipment was necessary. An event recorder (Epson HX 20 portable mini-computer operating on batteries) was used to code the mother–infant interactions continuously. The coding scheme of the event recorder consisted of three coding types: contextual codes (three digits); behavioral

codes (six digits), and system codes (one or two digits). An example of a typical relevant sequence is shown here in Table 12.1. The sequence starts with a contextual code A4E: the baby is actively awake; no caretaking situation, mother is in another room and can hear the baby cry. Interaction episode 1 (E1) indicates that the baby starts to cry, while mother is busy in the kitchen (ic3mbc); the next episode starts when mother enters the room and the baby still cries (meric3); E4: mother picks up/holds the baby, who is still crying; the next code is contextual, indicating the same behavioral state of the baby and close proximity; E5: the baby initiates this episode by looking at the mother (crying stops), and the mother vocalizes positively. In this way the observation session was segmented into interaction episodes dictated by behavioral changes either on the part of the mother or the baby (or both). The time table shows the start of the interaction episodes. The contextual code does not affect the time table. The task of the observer was reduced to accurately coding behavioral changes using a letter-coding scheme (see for a detailed description of the coding system Hubbard, 1989). In this chapter the focus is on the onset of maternal interventions, and not on the type of intervention.

During the observation period vocalizations of the baby and the mother (if she was in the same room) were recorded with an audio registration unit. This unit contained three components: (a) a wireless FM-transmitter/microphone combination (type Sennheiser SK 1012 (MKE 2012); (b) an Uher report 4400 portable stereo tape recorder; (c) a FM receiver (type Sennheiser/Telefunken EM 1008). In order to synchronize the event recorder and the audio-registration unit on a time-reference axis, a time-code generator was used to write the time (coded by tone pulses) on track one of the tape recorder. The vocalizations of the baby and the mother were recorded on track two. At the beginning of the visit, the time-code generator was connected up to the event recorder in order to match the two internal clocks. The audio-registration unit was placed in the living room, out of sight of the mother. The FM transmitter/microphone combination was always placed in the room where the baby was. The observer could check in the living room if the baby was awake via her earphones. When the baby awoke, the unit was started up, and could operate

Table 12.1. An example of coding interactional sequences

Code type	*Time table*	*Code*	*Behavioral episode*
	hrs:min:sec		
context	11:15:00	A4E	—
interact.	11:15:00	ic3mbc	E1
"	11:17:30	meric3	E2
"	11:18:00	mcpic3	E3
"	11:20:00	mphic3	E4
context	11:21:00	ACF	—
interact	11:21:00	illmvp	E5

for 90 minutes before the tape needed changing. In the meantime the observer was able to code the interactions of mother and baby by means of the portable event recorder.

The technical equipment was chosen to enhance the reliability and the validity of the data. Under all circumstances it could register the infant crying without being intrusive. If the baby was crying out of earshot of the observer (and often of the mother too) it was recorded. If the mother wanted the baby to sleep in the baby's room as usual during the observation session, this was possible. The observer would remain in the living room, checking unobtrusively from time to time through the earphones out of earshot of the mother, to hear if the baby was still awake. The registration of infant crying did not interfere with the mother's perceptions and consequently not with possible interventions. The job of the observer was easier than in the Baltimore study, because of the fact that the timing of relevant events was done by the event recorder and the time-code generator. Bell and Ainsworth (1972, p. 1174) remarked: "observers in their participant roles sometimes found it impossible to time accurately and hence resorted to estimates."

Infant Crying

Bell and Ainsworth (1972) had defined the central unit measure, that is, the crying episode, as any instance of a vocal distress signal (protest, fuss, or full-blown cries) not too brief to be timed (on the spot) and separated by more than a momentary pause from another instance. We took over this definition, except for the rather vague and not replicable value of the two parameters involved, minimal duration and time lag between subsequent crying. Every cry signal separated by a pause of 2 seconds from the next crying instance was coded as a crying episode.

Tape recordings were analyzed with a time decoder which displayed the timetable recorded during the observation. Given that there was synchronization between the timetable of behavioral episodes (event recorder) and the timetable of corresponding tape recordings, it was possible for the coders to make an accurate quantitative analysis of infant crying and the interventions of the mother, using the observers' coded observations as a guideline. Six coders analyzed the vocalizations of the baby either in a couple in the beginning or alone (after some 25 analyses). They were initially trained, using a record of infant crying (Wasz-Höckert, Lind, Vuorenkoski, Partanen, & Valanné, 1968) and our own tape recordings of crying. Vocalizations were analyzed twice before being coded as crying or noncrying. In fact, the analysis of one coder was independently checked by another coder. Disagreements between coders were coded as noncrying. After the check by the second coder the onset and finish of every crying episode was noted.

Mother's Unresponsiveness

The onset and end of every intervention or intervention sequence was noted with respect to every crying episode. Because we used a sensitive wireless microphone the vocalizations of the mother were also available on track two of the taperecording. Given the time table recorded on track one, the onset of verbal interventions was noted from the taperecording. Furthermore, coders regularly inspected the accuracy of the observers with respect to the timing of maternal interventions. These checks revealed that observers were accurate by means of the event recorder (the coding of onset time took a one-button press) and that the average delay was about two seconds. In the rare case that there was a large discrepancy this was corrected by the coders using cues of tape recordings. For example the observer might be less accurate for the intervention "enters room," if a mother rushes to the baby upstairs. The onset of this nonverbal intervention was also available on tape recording because the microphone was in the baby's room. The duration of unresponsiveness equaled the time the baby cried without an intervention of the mother (mother's delay), and equaled the duration of crying if an intervention started later than two seconds after the crying episode stopped. The two-seconds criterion is only relevant for nonverbal interventions and is only a matter of precision. In fact, we adopted the Bell and Ainsworth definition of unresponsiveness. Maternal interventions were, for example, picking baby up, holding; vocalizing; changing position; offering pacifier or toy; removing noxious stimulus; entering room. The focus of this chapter is on the delay of maternal interventions, not on the type of intervention (Bell & Ainsworth, 1972).

Reliability

Intercoder reliability. The time-tabled tape recording of infant crying and maternal vocalizations made it possible to calculate duration of crying and the delay of maternal verbal interventions post hoc in such a way that intraobserver and interobserver error variance due to measurement on the spot (Bell & Ainsworth, 1972) were eliminated. The distinction between crying and noncrying was made by consensus (i.e., if there was disagreement among coders about crying, then vocalizations were excluded from the data). For vocalizations of the baby with a minimal duration of five seconds, agreement percentage for crying was 95 percent for a random sample of 60 visits out of a total of 600 visits. Given the method of consensus reliability of duration of crying is reduced to the reading of onset and finish on a time reference axis. The mean agreement among coders for duration of unresponsiveness verbal interventions was 98.4 percent for a sample of 60 visits. For nonverbal interventions mean agreement percentage was 95 percent (from a sample of 14 visits done by two observers).

Data Analysis

Several different operationalizations of crying episode were computed with a computer program (based on Fortran) by specifying the two parameters: minimal duration of a crying episode and pause between two crying instances (see Hubbard, 1989; Hubbard & Van IJzendoorn, 1987, 1991). In this chapter we focus on duration measures (Ainsworth & Bell, 1977), based on the following definition of a crying episode: a distress signal with a minimal duration of five seconds, separated by four seconds or more from another instance. For each quarter maternal and infant measures were summarized for four successive observations. These measures are supposed to be comparable to Bell and Ainsworth's (1972) measures.

First, univariate descriptive statistics for infant crying behavior and maternal unresponsiveness are described. Second, in order to replicate the results of the original study, interquarter correlations between infant crying and maternal unresponsiveness are presented. Third, cross-lag panel analysis and a partial correlation approach are applied to control for contaminating variables. Last, using a development of duration of infant crying variable, hierarchical multiple regression analyses are carried out to compute standardized regression weights between maternal unresponsiveness and development of infant crying in subsequent quarters. Sample size did not allow for inferential causal modeling by means of LISREL (Boomsma, 1982; Tanaka, 1987).

RESULTS AND DISCUSSION

Developmental Changes in Infant Crying and Maternal Unresponsiveness

First we discuss changes in duration of infant crying and maternal unresponsiveness throughout the first 9 months. We found the same overall reduction in duration of crying as was found in the Baltimore study (see also Wolff, 1987, p. 80 ff.). The median of 5.9 minutes crying per hour in the first quarter is reduced to a median of 2.9 minutes per hour in the third quarter. bell and Ainsworth (1972) reported a median of 7.7 minutes crying per hour in the first quarter and 4.4 minutes per hour in the last quarter, and observed the same wide range of more than 15 minutes per hour to almost no crying at all, with a narrowing of the range toward the end of the first year. Wolff (1987) found crying and fussing about 10% of the time he observed the infants during the first quarter of their life.

The duration of maternal unresponsiveness also decreased in the first nine months from a median of 4.0 minutes per hour (Bell & Ainsworth, 1972; 3.8

minutes per hour) to 1.9 minutes per hour in the third quarter (Bell & Ainsworth, 1972, reported the same median for the fourth quarter). The medians for all measures were close to the means, indicating that the distribution are approximately normal. A repeated measures analysis of variance with age as factor confirmed the trend of decreasing duration of infant crying and maternal unresponsiveness during the first 9 months. For each quarter, analyses of variance with parity and sex as factors and duration of unresponsiveness and crying as dependent variables were computed. These analyses yielded no significant results.

Stability of infant crying and maternal unresponsiveness. A significant (auto-) correlation for duration of crying was found only between the first and second quarter ($r = .64$; $p < .001$). Bell and Ainsworth (1972) reported a significant correlation between the third and fourth quarter. In our case, results pointed toward instability of individual differences in duration of crying after the second quarter. Duration of maternal unresponsiveness showed about the same pattern. Bell and Ainsworth's (1972) data showed stability for the whole year.

Relation between infant crying and maternal unresponsiveness. We also found significant positive interquarter correlations between earlier maternal unresponsiveness and later infant crying. If these correlations were not spurious, they would indicate that babies who cried longer had mothers who were more unresponsive to their infants' crying in earlier quarters. Only the correlation between the first and second quarter ($r = .53$), and the first and third quarter ($r = .28$), appeared to be significant, whereas Bell and Ainsworth (1972) reported also significant figures for the last two interquarter correlations. On the basis of this correlational pattern Bell and Ainsworth (1972) concluded that those mothers who tended to delay response in one quarter would see their baby cry longer in the next one. In our study, this conclusion would certainly apply for the first half year, with the same tendency in the third quarter.

Controlling for antecedent and concurrent variables. Because intraquarter (synchronous) correlations between crying and unresponsiveness were very high, and the autocorrelations for crying and unresponsiveness were also considerable, statistical control for contaminating variables had to be made (Gewirtz & Boyd, 1977a). First, a cross-lag panel analysis was carried out. Our data allowed for cross-lag panel analysis because synchronous (or intraquarter) correlations did not differ significantly, and because a stable difference existed between parallel autocorrelations, indicating a stationary process (Cook & Campbell, 1979; Kenny, 1975). The reliability of synchronous panels was also comparable, because the same procedures were used to

measure unresponsiveness and crying. The largest difference between the three cross-lag correlation pairs was small (.06) and not significant, indicating that the alleged causal relation between unresponsiveness in an earlier quarter and crying in a later quarter is spurious.

Second, crying in an earlier quarter was partialled from the association between crying in a later quarter and unresponsiveness in an earlier quarter, to confirm our cross-lag panel analysis using a different statistical approach. For the first two quarters the partial correlation between unresponsiveness and crying was .03 (n.s.), and for the last two quarters this partial correlation was −.19 (n.s.). The significantly positive bivariate correlations between earlier unresponsiveness and later crying disappeared after controlling for duration of crying in the earlier quarters. Partialing unresponsiveness in the second quarter and in the third quarter also yielded nonsignificant partial correlations between earlier unresponsiveness and later crying. We have to conclude, that Bell and Ainsworth's (1972) hypothesis about the relation between unresponsiveness and crying is not supported by the partial correlations, nor by the results of the cross-lag panel analysis.

Development of crying. Because Bell and Ainsworth's (1972) measure for duration of infant crying did not take into account individual differences in development of crying, we computed a variable "development of duration of infant crying," defined as the log-transformed proportional decrease or increase of crying for two successive quarters.[1] This measure was used as a criterion variable to answer the question whether infants with more unresponsive mothers decrease their crying faster or slower than infants with less unresponsive mothers. In fact, it was this developmental question that Bell and Ainsworth (1972) tried to address in their Baltimore study, because they already had registered how infants in general decreased their crying during

[1]The development measure is not a difference score, but a log-transformation of the quotient of crying during the second quarter divided by crying during the first quarter, and so forth. Change scores have been criticized because of their lesser reliability, their inability to take initial differences into account, and their artificially high negative correlation with initial scores (Lord, 1963; O'Connor, 1970; Visser, 1982). Rogosa, Brandt, and Zimovski (1982; see also Zimmerman & Williams, 1982) showed, however, that under certain conditions this criticism is unwarranted. In our case, the reliability of the original variables is very high, because they were registered in a purely mechanical way on four different occasions per quarter. Statistical regression to the mean would, therefore, occur within those four different times of measurement. Furthermore, the developmental measure constitutes a change score based upon a quotient that takes initial differences into account, and through the log transformation the measure is symmetrically distributed. The correlations between the developmental measure and the initial score were in our sample −.31 ($p < .05$) for the first two quarters, and −.07 (n.s.) for the last two quarters (cf. Zimmerman & Williams, 1982). These correlations are quite modest, they can be considered "facts of life" (Rogosa et al., 1982), and in a multiple regression approach the significant correlation can be controlled for.

the first year. Besides representing the developmental changes more adequately, our developmental measure has the advantage to be correlated less strongly with concurrent maternal unresponsiveness (.43 and .69 for the second and third quarter respectively), thereby avoiding the multicollinearity problem. A hierarchical multiple regression analysis on development of crying in the first half year was carried out. Unresponsiveness in the second quarter and crying in the first quarter were entered into the equation first, and unresponsiveness in the first quarter was entered thereafter.

From this regression it can be derived that controlling for earlier quarter crying and concurrent unresponsiveness (Gewirtz & Boyd, 1977a), resulted in a significantly negative standardized bèta weight for the relation between first quarter unresponsiveness and development of crying in the first half year. This result did not support Bell and Ainsworth's (1972) hypothesis: Mothers who tended to be more unresponsive in the first quarter had babies who decreased their crying relatively more in the first half year. To test whether this outcome was due to the subgroup of infants who cried relatively short (cf. Ainsworth, personal communication; Landau, 1982), the sample was divided in two groups using the median for duration of crying as criterion. In the subgroup of infants crying relatively short ($n = 25$), the bèta weight for unresponsiveness in the second quarter was .69 ($p < .001$), for crying in the first quarter $-.46$ ($p < .01$), and for unresponsiveness in the first quarter $-.35$ ($p < .05$). In the subgroup of infants who cried relatively long ($n = 25$), bèta weights were .92 ($p < .001$), $-.28$ (n.s.); and $-.39$ ($p < .05$) respectively. In both subgroups, therefore, essentially the same relation between first quarter unresponsiveness and development of crying in the first half year was found as in the total group.

A hierarchical multiple regression analysis was also carried out on development of infant crying in the last two quarters. Unresponsiveness in the third quarter, development of crying in the first two quarters, and infant crying in the first quarter were entered into the regression equation first, unresponsiveness in the second quarter was entered thereafter, and unresponsiveness in the first quarter was entered last. Maternal unresponsiveness in the second quarter and the first quarter was not significantly related to development of infant crying during the last two quarters. This result does not support Bell and Ainworth's (1972) hypothesis, but also does not confirm the reverse relation between unresponsiveness and development of crying that appeared to exist in the first half year. Multiple regressions in the two subgroups of infants who cried short or long yielded the same result.

To address the question, raised by Gewirtz and Boyd (1977a), whether responsiveness is inversely related to unresponsiveness, the same analyses were carried out with a proportional measure for unresponsiveness. The percentage duration of crying ignored—which is, of course, the the inverse of the percentage of duration of crying responded to—did not correlate with first

quarter duration of crying ($r = -.01$). Outcomes of multiple regression analyses with this proportional measure for unresponsiveness were parallel to the results we reported upon above: a significantly negative bèta weight ($-.28$; $p < .05$) was found for the relation between first quarter maternal unresponsiveness and development of infant crying during the first half year, indicating that mothers who tended to respond more promptly, had infants who decreased their crying relatively less in the first half year of life. No such relation could be established for the development of infant crying after the second quarter.

CONCLUSIONS

According to Ainsworth and Bell (1977) an independent replication, preserving the intensive naturalistic longitudinal aspects of their exploratory study, was necessary before the dispute with their critics could be settled. The present study is such an effort. The same design has been used in which the mean total duration of the observations in every family is about seven to eight hours per quarter. Furthermore, more reliable and refined methods of observing, coding, and analyzing data in a larger sample have been applied. It is, therefore, interesting to see how far our descriptions of development in infant crying and maternal unresponsiveness converge with those of the original study, in view of the fact that our study has been carried out in another country more than fifteen years after the original one. Although Bell and Ainsworth (1972) collected data via the method of narrative accounts, for which reliability figures were lacking (Lamb, Thompson, Gardner, & Charnov, 1985), descriptive results did not differ markedly from those of our study. It is also evident that our bivariate interquarter correlations between infant crying and maternal unresponsiveness are consistent with the main conclusion of the original study: those mothers who tend to delay their response in one quarter, will see their babies cry longer in the next one (Bell & Ainsworth, 1972).

A cross-lag panel analysis, however, indicated that the correlations between unresponsiveness in an earlier quarter, and crying in a later quarter have to be considered as spurious. Partialling crying in an earlier quarter from the association between crying in a later quarter and unresponsiveness in an earlier quarter also resulted in nonsignificant interquarter correlations. Therefore, we have to conclude that Bell and Ainsworth's (1972) outcome is not replicated in our study. Although Ainsworth and Bell (1972, 1977) interpreted infant crying as a developmental phenomenon, differences in duration of crying at a certain point in time do not appear to reflect adequately differences in development of crying. Therefore, a new measure for development of infant crying was introduced operationalizing differences in individual "growth" of crying behavior. Using this new measure as criterion variable in multiple regression analyses we found a negative relation be-

tween earlier unresponsiveness and later development of crying in the first half year. Antecedent infant crying and concurrent maternal unresponsiveness were simultaneously controlled for. For the development of crying from the second to the third quarter we did not find a relation with earlier unresponsiveness. The data for the first half-year indicate that mothers who tend to delay their response relatively longer in the first quarter have infants who decrease their crying relatively more in the next quarter. This outcome was replicated in the two subgroups of infants who cry relatively long or short. Results were also the same if a somewhat more "precise" measure of crying episodes was used (i.e., every instance of a distress signal separated by two seconds or more from another).

The outcome of our critical replication study makes clear that the relation between maternal unresponsiveness and infant crying is somewhat more complex than originally was suggested. The technical critisms of Gewirtz and Boyd (1977a) on the Baltimore study appear to be supported. The critical replication study does of course not imply a refutation of attachment theory as a research program (Van IJzendoorn & Tavecchio, 1987), but indicates the necessity of adapting the theory to our unexpected outcome. Maybe current attachment theory should take Bowlby's (1971, p. 374) contention more seriously into account that mothers may condition certain infant behaviors. The conditioning paradigm has demonstrated experimentally that mothers are able to reinforce certain infant crying behaviors (Petrovich & Gewirtz, 1985). We propose, therefore, that in future studies the model of "differential responsiveness" should be tested. This model implies that only severe distress vocalizations (e.g., the pain cry) should be conceptualized as evolutionary "biased" attachment behavior, requiring a prompt response. The development of mild distress vocalizations ("operant" or "instrumental" cries) may solely be explained by proximate causes, or in terms of conditioning. Bowlby (1971, p. 347) already considered crying as a graded signal (Murray, 1979), carrying different information depending upon context, intensity, and rhythm (Thompson & Lamb, 1984), and he did not think it necessary for mothers to react promptly to all crying behaviors (cf. Ainsworth, 1973). A delayed response to mild distress may enable the infant to learn to cope with the kind of situations in which mild distress arises (Landau, 1982). Furthermore, simultaneously ignoring fussing behavior but being responsive to other behaviors such as exploratory play might diminish mild distress vocalizations (Lester, 1985). Only a minor percentage of infant crying behavior consists of pain or panic cries (Wolff, 1987) and might be considered attachment behavior. The model of differential responsiveness may, therefore, explain why in our study mothers who delayed their response to infant crying had children who tended to cry less later in the first half year of life. This model, however, can only be tested thoroughly in experimental studies discriminating between different types of infant distress vocalizations.

Our replication of the Baltimore study is, of course, not an exact replication

(see Chapter 3). Our study has been carried out in another culture some decades after the original study. Nevertheless, the raw data of the replication study did converge with those of the Baltimore study. It appears that a secondary analysis of the original study would have resulted in the same conclusions. A replication in another culture at a different point in time, however, also shows the remarkable robustness of the development of infant crying and its relation to maternal unresponsiveness. In this respect, the behavioral sciences do not always have to deal with phenomena that are more elusive than those studied in the natural sciences (see Chapter 5). Although our replication produced an unexpected outcome, it confirms at least the possibility of replicating social phenomena across a considerable cultural and historical span.

REFERENCES

Ainsworth, M. D. S. (1973). Systems for rating maternal care behavior. In E. G. Boyer, A. Simon, & G. R. Karafin (Eds.), *Measures of maturation. An anthology of childhood observation instruments*. Philadelphia: Research for Better Schools.

Ainsworth, M. D. S., & Bell, S. M. (1977). Infant crying and maternal responsiveness: A rejoinder to Gewirtz & Boyd. *Child Development, 48*, 1208–1216.

Ainsworth, M. D. S., Blehar, M. C., Waters, E., & Wall, S. (1978). *Patterns of attachment, a psychological study of the Strange Situation*. Hillsdale, NJ: Erlbaum.

Bell, S. M., & Ainsworth, M. D. S. (1972). Infant crying and maternal responsiveness. *Child Development, 43*, 1171–1190.

Belsky, J., Rovine, M., & Taylor, D. G. (1984). The Pennsylvania Infant and Family Development Project, III: The origins of individual differences in infant-mother attachment: Maternal and infant contribution. *Child Development, 55*, 718–728.

Boomsma, A. (1982). The robustness of LISREL against small sample sizes in factor analysis models. In K. Jöreskog & H. Wold (Eds.), *Systems under indirect observation: Causality structure-prediction* (Part I, pp. 149–173). New York: North-Holland.

Bowlby, J. (1971). *Attachment and loss. Volume I: Attachment*. Harmondsworth, UK: Penguin.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation. Design and analysis issues for field settings*. Chicago: Rand McNally.

Crockenberg, S., & McCluskey, K. (1986). Change in maternal behavior during the baby's first year of life. *Child Development, 57*, 746–753.

Crockenberg, S. B., & Smith, P. (1982). Antecedents of mother-infant interaction and infant irritability in the first 3 months of life. *Infant Behavior and Development, 5*, 105–119.

Gewirtz, J. L. (1961). A learning analysis of the effects of normal stimulation, privation and deprivation on the acquisition of social motivation and attachment. In B. M. Foss (Ed.), *Determinants of infant behavior* (pp. 213–299). London: Methuen.

Gewirtz, J. L., & Boyd, E. F. (1977a). Does maternal responding imply reduced infant crying? A critique to the 1972 Bell & Ainsworth report. *Child Development, 48*, 1200–1207.

Gewirtz, J. L., & Boyd, E. F. (1977b). In reply to the Rejoinder to our critique of the 1972 Bell and Ainsworth report. *Child Development, 48*, 1217–1218.

Grossmann, K., Grossmann, K. E., Spangler, G., Suess, G., & Unzner, L. (1985). Maternal sensitivity and newborns' orientation responses as related to quality of attachment in Northern Germany. In I. Bretherton, & E. Waters (Eds.), *Growing points of attachment theory and research. Monographs of the Society for Research in Child Development, 50*(1-2, Serial No. 209), 233–256.

Hubbard, F. O. A. (1989). *Infant crying and maternal unresponsiveness*. Leiden, The Netherlands: DSWD Press.

Hubbard, F. O. A., & Van IJzendoorn, M. H. (1987). Maternal unresponsiveness and infant crying. A critical replication of the Bell & Ainsworth study. In L. W. C. Tavecchio & M. H. van IJzendoorn (Eds.), *Attachment in social networks. Contributions to the Bowlby-Ainsworth attachment theory* (pp. 334–375). Amsterdam: Elsevier Science Publishers (North Holland).

Hubbard, F. O. A., & Van Yzendoorn, M. H. (1991). Maternal unresponsiveness and infant crying across the first nine months: A naturalistic longitudinal study. *Infant Behavior and Development, 14*, 299–312.

Kenny, D. A. (1975). Cross-lagged panel correlations: A test for spuriousness. *Psychological Bulletin, 82*, 887–903.

Lamb, M. E., Thompson, R. A., Gardner, W., & Charnov, E.L. (1985). *Infant-mother attachment: The origins and developmental significance of individual differences in Strange Situation behavior*. Hillsdale, NJ: Erlbaum.

Landau, R. (1982). Infant crying and fussing. *Journal of Cross-cultural Psychology, 13*, 427–443.

Lester, B. M. (1985). Introduction: There's more to crying than meets the ear. In B. M. Lester & C. F. Z. Boukydis (Eds.), *Infant crying. Theoretical and research perspectives* (pp. 1–28). New York: Plenum Press.

Lord, F. M. (1963). Elementary models for measuring change. In C. W. Harris (Ed.), *Problems in measuring change*. Madison: University of Wisconsin Press.

Murray, A. D. (1979). Infant crying as an elicitor of parental behavior: an examination of two models. *Psychological Bulletin, 86*, 191–215.

O'Connor, E. F. (1970). *Extending classical test theory to the measurement of change*. Unpublished dissertation, University of California, Los Angeles, CA.

Petrovich, S.B., & Gewirtz, J.L. (1985). The attachment learning process and its relation to cultural and biological evolution: Proximate and ultimate considerations. In M. Reite & T. Field (Eds.), *The psychobiology of attachment and separation* (pp. 259–286). New York: Academic Press.

Rogosa, D., Brand, D., & Zimowski, M. (1982). A growth curve approach to the measurement of change. *Psychological Bulletin, 92*, 726–748.

Sroufe, L. A. (1985). Attachment classification from the perspective of infant-caregiver relationships and infant temperament. *Child Development, 56*, 1–14.

Tanaka, J. S. (1987). "How big is big enough?": Sample size and goodness of fit in structural equation models with latent variables. *Child Development, 58*, 134–146.

Thompson, R. A., & Lamb, M. E. (1984). Individual differences in dimensions of socio-emotional development in infancy. In R. Plutchik & H. Kellerman (Eds.), *Emotion: theory, research and experience (Vol. 2). Emotions in early development* (pp. 87–114). New York: Academic Press.

Van IJzendoorn, M. H., & Tavecchio, L. W. C. (1987). The development of attachment theory as a Lakatosian research program: Philosophical and methodological aspects. In L. W. C. Tavecchio & M. H. van IJzendoorn (Eds.), *Attachment in social networks. Contributions to the Bowlby-Ainsworth attachment theory* (pp. 3–31). Amsterdam: Elsevier Science Publishers.

Visser, R. (1982). *On quantitative longitudinal data in psychological research*. Unpublished doctoral dissertation, University of Leiden, The Netherlands.

Wasz-Höckert, O., Lind, J., Vuorenkoski, V., Partanen, T., & Valanné, E. (1968). The infant cry: A spectrographic and auditory analysis. *Clinics in Developmental Medicine, 29*, Lavenham, Suffolk: Spastics International Medical Publications.

Wolff, P. H. (1969). The natural history of crying and other vocalizations in early infancy. In B.M. Foss (Ed.), *Determinants of infant behaviour IV* (pp. 81–109). London: Methuen.

Wolff, P. H. (1987). *The development of behavioral states and the expression of emotions in early infancy. New proposals for investigation*. Chicago: University of Chicago Press.

Zimmerman, D. W., & Williams, R. H. (1982). Gain scores in research can be highly reliable. *Journal of Educational Measurement, 19*, 149–154.

Epilogue

We have reached the end of our book; a book that was meant to investigate one of the canons of traditional research methodology: that experimenters in both the natural and social sciences should strive for replication or replicability. In the course of the book it has become clear—we believe—that the traditional belief that replication involves isomorphic copying of the original studies is ill-founded. The introduction to this book and the chapters in the first section prepared the ground for this conclusion. The contributors to Part I dealt with questions concerning the nature, function, and possibility of replication. It was argued that there are no ahistorical or cross-cultural rules about what constitutes replication and that exact replication is impossible. Rather, researchers should strive for theoretically interesting conceptual replications. Such replications provide the fuel for the never-ending debate about the validity of conclusions drawn on the basis of research data. In Part II the problem of method effects formed a major topic of discussion. It was argued that studies may not be replicable as the results are not method-invariant. Several suggestions for the prevention of method effects were given and a major new instrument to detect method effects—meta-analysis—was described in great detail. The authors agree that exact replication is either impossible or fruitless and that replication efforts should be theory-driven. In Part III several attempts to replicate interesting original studies have been described. The main conclusion of the careful replications of the Istomina study was that it may well be an impossible task for any experimenter to exactly reconstruct functionally similar experimental settings across cultures or historical periods. The replication studies related in the other chapters deliberately introduced variations in the original research

design and called into question the tenability of the conclusions drawn in the original studies. In general this section abundantly demonstrated the difficult problems one has to meet when actually attempting to replicate an experiment.

Several times in this book it was suggested that there may be a fundamental difference between the natural and the social sciences that makes replication more difficult in the latter. Although replication in the natural sciences appears to be far from easy several factors that may differ between the original experiment and its replication logically cannot play a role in the natural sciences and may do so in the social ones. These include factors such as the personality of the experimenter, the subjects studied, the sociohistorical background of the study, and so on (see the General Introduction of this book for a list of possibly contaminating factors). Consequently, it seems indeed easier to replicate the results of the anatomist than to replicate those of the psychologist, educationalist, or anthropologist. It is the impossibility of perfectly reconstructing the historical, cultural, or personal circumstances that lead to a certain type of conduct that determines the boundaries of our efforts to reconstruct the phenomena of the human mind.

All this leads to a major issue that has to be addressed at the end of this volume. There remains a great need to make sense of the concept of replicability, rather than merely reiterate its fundamental importance for science. That constitutes a paradox in scientists' treatment of replicability (which almost all of the contributions in this volume reflected): On the one hand, the canon of replicability as a norm for science is repeatedly emphasized, while on the other hand the realistic applicability of that canon turns out to be highly complicated or even (in many cases) impossible. So we have to demonstrate replicability while at best we can replicate any study in its conceptual rather than exact (or virtual) details. But conceptual replication often leads to the use of new empirical strategies that have little in common with the procedures of the previous investigations. Hence, the difficult question: how can one distinguish the cases of "conceptual replications" from new empirical efforts that use new procedures for testing previous conceptual models?

It seems that this latter question has no possible answer, and that the boundary between "conceptual replicability" of previous studies and "empirical extendability" of previous conceptual models is (and remains) a rather fuzzy one in a scientific discipline that claims to develop new knowledge about the particular class of phenomena. In other terms, science itself is a constantly developing (but not automatically progressing) knowledge construction process for which replicability is a means to the end of developing further, rather than merely documenting that some research tradition of the past had "right" results. The notion of replicability acquires a new meaning

when put into a context of a constructivist perspective on knowledge (see also Chapter 3 by Van IJzendoorn).

The issue of replicability is thus complicated by the developmental nature of knowledge construction in science. However, a similar complication awaits scientists on the side of the object of their investigation in any research on human development (be it psychological, sociological, or anthropological)—the developmental nature of biological, psychological, or social systems renders it in principle impossible to succeed in any effort of virtual replication of previous findings.

DEVELOPMENTAL PHENOMENA IN THE MIRROR OF REPLICATION EFFORTS

We should reemphasize that the present volume concentrated on issues of replicability as it is of relevance to research on human development rather than to social sciences as a whole. The majority of the contributions dealt with one or another issue of development. It must be noted that the developmental perspective in human sciences has not been very clearly emphasized in most of contemporary psychology. Hence it is possible to see an interesting fusion between theoretical perspectives of nondevelopmental kind and some scattered ideas reflecting development in present-day scientific discourse. It is as if the canons of science remain to be accepted as those fitting the assumption of the "true" or "stable" world, while the experience with phenomena reveals that that very world is not stable in any but a very abstract sense. The latter is evident at all levels represented in the present volume: microgenetic (e.g., human beings solving problems in research situations), ontogenetic (children's construction of novel ways of handling life- or experimental situations over time), culture-genetic (societies or cultures changing over the course of history, see Chapter 2), and—last but not least—science-genetic (i.e., the level at which scientific knowledge construction proceeds within the web of constantly changing society).

In the most general form, the developmental nature of phenomena makes it in principle impossible to arrive at any virtual replication in our studies. If procedure X is carried out with a subject (in a general sense; the subject here may be a person, social group, society, or scientific paradigm) at a certain time moment (T_0), and is attempted to be replicated at a later time moment (T_1), the developmental history of the subject between the time moments T_0 and T_1 makes it impossible to use exactly the same procedure and, thus, to find exactly the same results. This impossibility is further enhanced if we consider the inescapable possibility (which is true for all open systems, that is, developing phenomena) that the first application of X at time T_0 necessarily

participates in the very same developmental process that leads the subject to give us different results (to X) at the next time T_1. Virtual replicability is thus an idea that is antithetical to the notion of all research efforts being simultaneously "interventions" into the developmental processes that provide "input" for the further change of the subject.

Of course, most of contemporary psychology is not based on empirical studies of individual subjects over time (i.e., longitudinally), but on samples of subjects with the use of cross-sectional designs. Often the question of replicability becomes thus an issue of sampling comparable groups of subjects from the "general population," and thus creating the possibility of a "re-run" of the "effects" of a particular procedure. Here the problem described earlier is simply transposed to the level of a conglomerate of individuals (population) in which the particular subjects in the samples merely reflect some presumably stable characteristics of the "population." The "samples" become "comparable" because of their similar ways of "drawing them from the population" and hence replicability of procedure X (first used on Sample A) on a Sample B from the same population should allow for virtual replication. However, again it is the axiomatic assumptions of the investigator that may mislead our replication efforts—the "population" from which both samples (A and B) are "drawn" can be believed to be homogeneous by the investigator (in which case the samples A and B can be comparable), but not proven to be so. In fact, all statements about "general population" in psychological research are abstract statements that themselves cannot be proven (or disproven) other than deductively. In the deductive proof, the dangers of starting with an inadequate axiom (e.g., "the general population is homogeneous") may lead to acceptance of particular statements (e.g., "samples A and B, given their random sampling, are comparable") that can be undermined by alternative axioms (e.g., "all populations are heterogeneous" and "the particular form of heterogeneity cannot be determined"). If the assumption of heterogeneity of "general population" is accepted by an investigator, the comparison of samples A and B becomes inconsequential—even the same random sampling procedure does not guarantee their comparability, and hence the replication of the results of A in case of B again cannot indicate virtual replication.[1]

[1]Note that the problem of heterogeneity of a developing population cannot be solved also by way of parallel sampling of A and B from the "same" (at that moment) general population. This practice should overcome the problem of temporal changes in the assumed population, but will not be able to guarantee lack of uncontrollable (and unknowable) heterogeneity in it. The "randomly sampling" investigator is in no position to detect the structural organization of the population from which the random sampling takes place, and therefore the equality of the resulting samples remains in the head of the sampler, rather than in the groups of the sampled. The application of the same procedure in parallel to both samples may thus avoid any "sequential testing effects," but still is likely to fail to reach a state of exact replication of the results of the procedure obtained in sample A in the case of B.

Finally, from the developmental point of view it is reasonable to consider the possibility that the "general population" (however indeterminate our knowledge of it may be) is itself undergoing development. Hence if samples A and B are "drawn" from it at successive times, these samples reflect different "developmental cohorts" within the "general population". In this case, the problem becomes exactly parallel to the single case example described earlier. Namely, the application of X to sample A at time T_o can participate in the change that takes place in the "general population" between T_o and T_1, and when sample B is drawn at T_1 the application of X again cannot result in a virtual replication. The example given by Kloos (see Chapter 2, about Mead's and Holmes's restudy of adolescence in Samoa indicates part of the issue: The population (here taking the form of Samoan society) had undergone major changes between the two times the two investigators did their fieldwork, and hence it is not surprising that the anthropological accounts turned out to be different. Interestingly—a good example of the nondevelopmental orientation in cultural anthropology—Kloos considers the long time between the two studies of Samoa a "flaw" (rather than reflection of cultural development). In a similar vein, over historical time (reflecting cultural transformation) certain aspects of the experimental setting may change their shared meanings between members of the sample—for example, the replication of Istomina's experiment decades later lacks both the special meaning of collectivity of a kindergarten play (which is not even introduced in the replication experiments; see Chapter 10). The historical transformation of a society underlies the changes in the "general population" that happens to inhabit the given society. Therefore the samples of subjects (as well as individual cases) demonstrate changes due to their cohorts' life experiences, and hence most virtual replication efforts reflect these changes.

The role of the cultural and historical context in developmental phenomena is not always the same and, therefore, the possibility of replicating the outcome of developmental studies may vary greatly. The replication of Ainsworth's Baltimore study by Hubbard and Van IJzendoorn (see Chapter 12) is a case in point. Although the original study on maternal responsiveness and infant crying was carried out in the United States in the late 1960s, and the replication effort took place almost two decades later in the Netherlands, the development of infants' crying behavior in the first year of life proved to be describable by the same parameters. Even though different methods of data collection were used (paper and pencil versus electronic registration) the same crying curve could be drawn on the basis of the descriptive results of both studies. This remarkable converge may be explained in two ways: First, in the case of crying the time period necessary for essential changes in crying behavior to take place may be considerably larger than two decades. Dependent upon the specific developmental area, sociocultural changes may take a larger or shorter time to become effective. Second, crying behavior may be

somewhat more resistant to sociocultural changes as compared to other developmental phenomena. Being determined by both physiological and contextual processes it can be situated towards the lower end of the continuum of higher and lower psychological processes. As such it is probably more resistant to change than phenomena that are highly contextually determined. In some developmental areas replication efforts may indeed be considered particularly convincing if these yield similar results while separated from the original study in time and geographically (Rosenthal's maxim; see the General Introduction). In other developmental areas, however, a failure to replicate the outcome of the original study may demonstrate the role of the sociocultural context in developmental phenomena and stimulate a scientific discourse on the precise nature of this role. Relative stability of developmental phenomena should not be taken for granted—as seems to be the case in the traditional idea of replication as exact repetition of process and outcome of the original study—but has to be demonstrated through a series of carefully varied replications.

REPLICATIONS AND REORIENTATIONS IN SCIENCE: KNOWLEDGE CONSTRUCTION

Similar developments of the phenomena can be observed at the level of science itself—the perspective from which investigators of human development approach issues of replication has vastly changed over the present century. From the time of the controversies about replication that were on the minds of scientists at the turn of the century (i.e., the Wundt vs. "Wurzburg School" and Baldwin vs. Titchener oppositions - see Chapter 1) psychology has changed towards a very different emphasis on sample-based research and its meta-analyses (as was discussed in Chapter 3, Chapter 6, Chapter 7). That change of perspective leads the issue of replicability itself being set up for investigators in terms of conceptual development. What that concept itself meant at different periods of the history of psychology guarantees that the meaning of replicability has become extended. The wide range of approaches presently subsumed under the focus of this volume on replicability (refer to Figure 3.1) can be seen as differentiations of the themes used in the beginning of this Century, rather than their substitutions.

We now reach a major conclusion about the issue of replicability: The developmental nature of the phenomena under study (and that is the case with all aspects of human or animal development) can be replicable only in the conceptual sense, but never virtually. Thus, the prevailing "diagnosis" of the problematic nature of virtual replications that echoes from all the contributions to this volume is an experiential reflection of a basic general principle

(that of development) which deductively assumes the presence of constant and never-ending generation of novelty on behalf of the developing systems.

Having put to rest the vain hopes for achieving virtual replicability in the case of developmental sciences (i.e., any discipline—psychology, anthropology, sociology—that may take a developmental perspective), we need to reach some reasonable understanding of the notion of "conceptual replication." If the developmental nature of the phenomena rules out virtual replication possibilities, then the issue of conceptual replicability is complicated by the developmental nature of the scientific understanding itself. The strong emphasis on the "conceptual" notion here looks rather appealing, but has an inherent complication in it. Namely, if our scientific knowledge is a constructed entity that emerges from the scientific discourse (see Chapter 3), then the developing mind of the investigator that targets a developing phenomenon as an object of investigation can come up with a variety of conceptual models of the phenomena and may persevere in the ways in which a particular conceptual model is linked with empirical work (=is replicated). For example, one can create one's own "rules of thumb" of replicability on the basis of the variety of perspectives. One investigator may accept only the results of a secondary analysis of an existing data set or specimen of phenomena as convincing replications; another may concentrate on the accumulation of data from various studies by way of meta-analyses (see Chapters 6 and 7, respectively). Likewise, investigators may decide that replication of the original study using novel inference methodologies may be of relevance (see Chapter 12). Likewise, the whole "mindscape" of replicability procedures provides a wide range of opportunities for rejection of some. One can reject the use of meta-analysis, for example, because of its inductive inferential nature and lack of explicit linkages with theoretical underpinnings of the different studies that "produce" the data. Likewise, the use of secondary analysis can be rejected on the grounds of theoretical incompatability (e.g., a "cognitivist" who has collected samples of 3–5-year-old children's behavior in "theories of mind" type experiments may consider any re-analysis that a "behaviorist" is likely to provide for the same specimens, as inconsequential for replication of one's original results).

Taken all together, the ways in which scientists define their norms of replicability of particular results retain a flavor of all of the subjectivity inherently involved in our knowledge construction. In Chapter 7, Siegfried illustrated the subjective insistence of so-called "hard scientists" upon the rules by which to keep up their formed understanding of the constructed knowledge. The rules by which concepts and empirical study practices are related are often consensually made and kept up in the community of scientists. In science, subjectivity is an inevitable facet of researchers' slow and complicated move toward greater objectivity in our constructed understanding of the world. In other terms, no method in science guarantees

objectivity in any automatic way, but objectivity can be constructed with the help of the human mind that critically investigates both the phenomena and the previous investigations of those phenomena. Thus, Lytton's exposure of the sequence of decisions that the user of a meta-analysis technique has to make (see Chapter 6) demonstrates convincingly that no meta-analytic procedure in and by itself can give us an objective summary of the data, but instead the investigator's thoughtful sequence of decisions while performing meta-analysis and equally thoughtful interpretation of its results may lead us towards greater objectivity. That focus is resonated also by Siegfried (Chapter 7) who emphasized the relevance of theorizing in conjunction with any meta-analytic adventure.

CAN VIRTUAL NONREPLICATION BE A SPECIAL CASE OF CONCEPTUAL REPLICATION?

At first glance, this question seems to be highly paradoxical. Indeed, our focus so far has been on replicability as a highly desired state of affairs. This, however, need not automatically mean that explicit failure to replicate a particular study in its exact form is necessarily tantamount to proving the conceptual level of that study "wrong." Since the conceptual and empirical levels of research can be linked in a multiplicity of ways (due to the developmental nature of the phenomena, or due to selective focus of an empirical study on some part of the phenomenon, rather than its entirety), it is possible to look for examples of how virtual nonreplications of a study might constitute conceptual replications.

Once More on Istomina: What Actually Was Replicated?

The case of replications of the "Istomina-experiment" that is covered in great detail in the present volume can provide a good example. As the different versions of that experiment provided varied results (see details in Chapter 9 by Schneider and Hasselhorn) in the sample-level summaries of the "lesson" versus "activity" conditions, one can be sure to recognize in these results all the difficulties of virtual (exact) replication that most of the contributors discussed in this volume. Furthermore, the exact procedures of Istomina were proven to be not repeatable (see also Chapter 10). Thus, Istomina's experiment at the level of exact replications turns out to be either inconsistently replicated or in principle nonreplicable—dependent upon the empiricist or theoretical analytic focus one or another investigator uses to look at the Istomina case.

However, the limited success in virtual replication efforts need not indicate that Istomina's study was not replicated at its conceptual level. Indeed, the

main conceptual focus of the study was the relevance of activity context in its motivating of the mnemonic task solutions. When Istomina built that activity "pull" into her task by way of unreplicated (and nonreplicable) "Kindergarten play" that relied heavily on the children's socioeconomic and symbolic worlds of the time, the same conceptual focus can be encoded in the empirical procedures differently in other places and at other time. Thus, for instance, if the "lesson" setting becomes highly motivating in some (schooled) children, it may take over the role of the most motivating activity setting in another context. Hence, if investigators compare "lesson" and "play" settings in an effort to replicate Istomina at her conceptual level, then the higher results in mnemonic tasks in "lesson" contexts (in comparison with "play") can be seen exactly as conceptual replication of Istomina, while in terms of exact replication these results amount to a nonreplication. Furthermore, any experimental setting can be structured in ways that either enhance or suppress children's fragile motivation (see discussion of that point in Chapter 9), hence, the mere labeling of a memory-task setting into "lesson" or "game" (or by any other label) does not reveal the whole texture of the setting that actually constitutes the children's environment in working on the given tasks.

Finally, the Istomina replications that were brought together in this volume indicate also the historical shift in what has become the target of replication. In Istomina's own study, the data on individual subjects' handling of the setting are given in exact narrative detail. These data are the main data that allow for direct test of Istomina's theoretical propositions. However, all the replication efforts have concentrated at the sample-level summary data of comparing the different conditions. If we were to try to replicate Istomina's work at its direct conceptual level, we would have to study single cases of children who are moved through a series of memorization-demanding tasks with varied activities, and look for the process data (on how the child proceeds to memorize) in each of those settings. For example, one child in the "Kindergarten" play setting might show high rehearsal strategies and succeed in memorizing many more items than in the "lesson" setting. If another child demonstrates a vastly different pattern of handling these two settings—is little involved in Kindergarten play and does not rehearse the items, plus ends up not remembering them—although that child may be actively following the teacher's demands in the "lesson" setting by rehearsing the word list (and showing superior performance), then for that child the two activity contexts are functionally reversed: The "lesson" context is that of activity-embedded learning context, whereas the "Kindergarten play" happens to be a context of little personal involvement. For the two children the data show opposite outcomes in their memory performance, but similar processes (only in different contexts) in involvement in the different activity contexts. Hence, although the second child can be seen as nonreplicating Istomina's study in its exact/virtual sense, conceptually that nonreplication equals replication. In a

more general sense—as long as children are active co-constructors of their life-worlds, the structure of their life-worlds may be varied from case to case, and seemingly standardized external situation ("lesson" versus "play") is interpreted in accordance with one's own personal experiences by the given child him- or herself.

REPLICATION EFFORTS AND THEIR THEORETICAL BACKGROUNDS: ONCE MORE ABOUT "NATURE" AND "NURTURE"

Another issue in the realm of replicability is the diversity of investigators' perspectives from which they attempt to replicate other studies. If an investigator takes a perspective that is divergent with that of an established alternative "classic" view, then one possibility to discredit the opposing view is to undermine its empirical basis. This introduces a strategic quality to scientists' discourse about replicability: If by demonstrating nonreplicability of one's opponent's experiment it is possible to promote one's own perspective, then the function of replicability ceases to be that of "check" of the adequacy of the scientific knowledge. Instead, it becomes a rhetoric device in the duel of opposing general ideas or ideologies. Very often, these duels are fought out from one party taking one (partially adequate) side of the issue, and the other the other side (equally partially adequate). However, the resulting "fight" may be quite counterproductive for our knowledge construction, since in their opposition the two parties fail to recognize that they overlook the essential unity of their opposing ideas. Still, claims of nonreplicability of the other side's studies may be heard loud in the scientific discourse.

A good example of the ways in which different "replication efforts" that start from different conceptual perspectives lead to nonreplicability is the supposed nature of adolescence on Samoa (see Chapter 2). The contrasting of "nature" (i.e., the belief that adolescence is a stage determined by biological maturation) and "nurture" (i.e., that culture "molds" or "shapes" the ways in which adolescents proceed through their becoming-adults phase) is an old ideological dualism that prevails in psychology and anthropology. By setting up one's research question here in either–or terms, the investigators guarantee that different researchers will find each other's "data" nonreplicable. Thus, the role of both biological and cultural facets in adolescents is necessarily plausible. However, a given investigator may want to explain adolescence conceptually by making a choice between the two equally plausible explanations (let A = "maturation" and B = "social expectations"). Based on his or her metatheory, one investigator is eager to view most of the phenomenon as a reflection of A, and selectively picks up empirical evidence that corroborates that perspective. Another investigator may try to "replicate"

the first one's study, but starting from a different conceptual basis (B). Not surprisingly, this leads to the failure of replication, as this other investigator selects exactly those aspects of the phenomenon to which the previous investigator did not pay attention. The result is seemingly a nonreplication, but its roots are in the conceptually opposite perspectives the two investigators have taken.

Interestingly, both of the investigators overlook the possibility that the nature of the phenomenon can be generated by the systemic linkedness of A and B. This of course implies another metatheoretical starting point—that of a systems theory and its characteristic focus on linkages between parts of the system—in contrast to verifying the relevance of one of its parts over the others. From that perspective, the failure to replicate the studies of the opposite "camp" indicates that both opposing perspectives are in a state of theoretical impasse from which there is no empirical exit possible. Hence the whole issue of nonreplication and its problems (as those look from the side of empirical investigation) may turn out to be misleading—both investigators have turned a partial explanation into an absolute one, and have failed to put together a more appropriate explanation of another (systemic) kind (e.g., one that takes the form "A linked to B" and specifies the particulars of the assumed links). Indications that the controversy about replicability of Margaret Mead's description of Samoan adolescence is of that kind are given by Kloos (see Chapter 2). When the claims that adolescents' psychological state "is **related** to culture" and "is **not the** outcome of . . . psychobiological changes" are viewed as irreconcilable opposites, the above-mentioned controversy would remain in its fighting force. However, the meaning of "related to culture" can be seen in systemic terms, and the "psychobiological changes" in adolescence can be linked with the functioning of culture (rather than its "effects" upon the adolescents). This way the issue of nonreplicability at the level of mutually opposed partial pictures of the phenomena can become integrated at another conceptual (systemic) level. If that happens, examples of nonreplicability have participated in the reconstruction of scientific knowledge.

If seen from this angle, blatant failures of conceptual replication may be of good use for science. These failures are not failures of one of the previously accepted conceptual models in the sense of its replacement by an opposite model, but are rather diagnostic of the need for reconstructing the whole search for better conceptual models that could fit the phenomena. A similar function of replication difficulties may be viewed to be present in case of inconclusive results of meta-analyses, or of secondary analyses. In those cases, the discrepancy between replication efforts can be interpreted as co-presence of different interpretations within the dynamically changing developing system. When meta-analyses or secondary analyses give contradictory, or inconclusive summaries of the empirical data, it is useful to

contemplate whether the theoretical questions underlying the empirical data accumulation efforts are phrased in an adequate manner.

To summarize, the developing nature of the understanding by science (=by scientists) is itself constructive in ways that involve both selection of different aspects of the phenomena and their transformation into data by way of some metatheoretic schema. The emergence of replication difficulties may indicate a fruitful, but incomplete construction effort of the conceptual perspectives involved, rather than cast doubt on the "rightness" of the previously substantiated conceptual model.

REPLICABILITY AND MULTILEVEL CONSTRUCTION OF SCIENTIFIC KNOWLEDGE

We hope that this book has provided a new look at the old problem of replicability. Indeed, from the merely empirical question of "can one replicate study X" and a general authoritative rule "all empirical results must in principle be replicable," we have moved toward a general investigation of how scientific construction of knowledge operates. Because this volume primarily focused on the issues of the developmental perspective in science, it brought into focus a number of issues that other methodologists of replication (e.g., Robert Rosenthal's "maxim or replicability" described earlier) have overlooked, due to their lack of emphasis on development.

We reach the understanding that the issue of replicability is but a tip of an iceberg, of the whole process by which science constructs its knowledge. Only part of that process is empirical, and much of it takes place at the methodological, theoretical, and metatheoretical levels. It can be argued that the whole issue of replicability also belongs to the methodological level and cannot be analyzed if we only look at the empirical level of scientific activities. Instead, the question of replicability is closely linked with the issue of methodological innovation (at the same, methodological, level), and closely guided by the next higher level (that of theory). Because of that latter guidance it becomes possible to find "camps" of opposing theories pointing to the difficulties to replicate their opponents' studies. That finger pointing is, of course, in itself a futile rhetoric device used in scientific discourse. However, it can be also fruitful as it can lead to the necessity to rethink whether the opposing theoretical perspectives are so much in opposition as they claim to be, and whether in their "fight" they have overlooked some more relevant focus on the phenomena. If the question of replicability leads to the need of general reformulation of our reconstructions about our world then its relevance transcends the limited function of "checking the empirical data" for their consensually defined "goodness" or "flaws." Rather, the issue of replicability leads us to a better picture of the functioning of the scientific process at all levels with all of its contradictions, inconsistencies, misunderstandings,

jealousies, and journalistically advertised "breakthroughs." In the middle of that polyphony of the drama of science is the birthplace of new knowledge, and the personal lives of the scientists who are determined to live through the difficult road to the rare moments of discovery. The general issue of replicability is thus an important, but episodic theme in our endless striving for better knowledge.

Author Index

A

Abrami, P.C., 145, *147*
Achenbach, T.M., 98, *113*
Adams, A.K., *236*, 237, 241, *242*, 243, 244, 245, 246, 247, 248, 249, 250, 251, 253, *253*
Adams, K., 46, *54*
Ainsworth, L.L., 58, *69*
Ainsworth, M.D.S., 60, 62, 67, *69*, 121, *147*, 172, *172*, 255, 256, 257, 258, 261, 262, 263, 264, 265, 266, 267, 268, *269*
Aivano, S., 225, *230*
Alansky, J.A., 67, *69*
Allerbeck, K.R., 155, *168*
Allinsmith, W., 133, *147*
Amir, Y., 4, 7, *8*
Angell, J.R., 22, *34*
Arbinger, R., 189, *204*
Asher, S.R., 98, *114*
Aspelin, P.L., 47,*54*

B

Bahr, H.M., 59, 65, *69*, 151, 155, 160, 161, 164, *168*
Baker, J., 187, *204*
Baker-Ward, L., 202, *204*
Baldwin, A.L., 119, *147*
Baldwin, J.M., 21, *34*
Bangert-Drowns, R.J., 129, *147*
Bank, L., 99, 111, 112, *113*
Barnard, K.E., 122, *147*
Barnes, T., 162, *168*
Basov, M.I., 176, *181*
Baumrind, D., 119, 122, *147*
Beach, F., 97, *113*
Beals, A.R., 41, *55*
Beck, A.T., 162, *170*
Bell, A.H., 162, *168*
Bell, S.M., 121, *147*, 172, *172*, 255, 256, 257, 258, 261, 262, 263, 264, 265, 266, 267, *269*
Belsky, J., 257, *269*
Benbow, C.P., 142, *147*
Benedict, R., 45, *55*
Bennett, J., 49, 54, *55*
Bentley, A.F., 86, *91*
Bernauer, F., 161, *168*, *169*
Best, D.L., 218, *230*
Bjorklund, D.F., 218, 226, *230*
Bjorn, K.M., 236, 237, 241, 242, 243, 244, 245, 246, 247, 248, 249, 250, 251, 253, *253*
Blehar, M.C., 60, 62, 67, *69*, 255, *269*
Block, J.H., 126, *147*
Blumenthal, A.L., 3, *9*, 58, *70*
Böhme, G., 22, *34*
Boomsma, A., 263, *269*
Boring, E.G., 30, *35*
Borkowski, J.G., 227, *230*

Bornstein, R.F., 6, *8*, *9*
Bowlby, J., 255, 268, *269*
Box, J.F., 31, 32, *35*
Boyd, E.F., 256, 257, 258, 264, 266, 268, *269*
Bozarth, J.D., 6, *9*
Bradley, R.H., 122, *147*
Brainard, P.P., 234, *253*
Brand, D., 265n, *270*
Breese, F.H., 119, *147*
Broad, W., 37, *55*
Bronfenbrenner, U., 187, *204*
Brown, P.J., 42, *55*
Brownstein, L., *91*
Brun, H., 187, 191, 192, 193, 194, 200, 201, 202, *205*, 209, *230*
Bruner, J., 233, *253*
Bryant, F.B., 66, *69*
Bühler, K., 20, *35*, 234, *253*
Bulldock, R.J., 151, *168*
Burton, R.V., 118, 119, 120, 133, *147*, *149*
Burtt, C., 162, *170*
Bush, R.R., 123, *149*
Byrne, D., 108, *113*

C

Cahen, L., 143, *148*
Caldera, Y.M., 141, *147*
Caldwell, B.M., 122, *147*
Campbell, D.R., 151, *168*
Campbell, D.T., 104, 107, *113*, 122, *147*, 264, *269*
Campbell, J.D., 118, 119, 120, *149*
Capaldi, D.M., 111, *113*
Caplow, T., 59, *65*, *69*, 151, 155, 160, 161, 164, *168*
Carli, L.L., 121, *148*
Caton, G., 162, *168*
Cattell, J.M., 25, *35*
Ceci, S.J., 187, *204*
Chadwick, B.A., 59, 65, *69*, 151, 155, 160, 161, 164, *168*
Charnov, E.L., 267, *270*
Christophersen, E.R., 162, *168*
Clarke-Stewart, K.A., 120, 121, *147*
Cochran, W.G., 28, 31, *35*
Cohen, D., 161, *170*
Cohen, P.A., 145, *147*
Cole, M., 187, *204*
Collins, H.M., 18, *35*, 59, 61, 63, 64, *69*, 74n, 81, *91*
Cook, T. D., 264, *269*
Cordray, D.S., 139, *149*
Cornwell, D., 3, *9*
Corsale, K., 218, *230*
Cowan, R.S., 24, *35*
Coy, P., 50, *55*
Craig, J.R., 65, *69*
Crandall, R., 6, *9*
Crockenberg, S., 257, *269*
Cronbach, L.J., 66, *69*, 100, *113*
Crowley, M., 136, *148*

D

d'Apollonia, S., 145, *147*
Danziger, K., *35*
Davis, S.F., 3, *9*
Dawes, R.B., 123, *148*
Degerando, J.M., 39, *55*
DeGroot, A.D., 73, 74, *91*
Denzin, N.K., 151, *168*
DeVinney, L.E., 151, *170*
Dewey, J., 86, 87, 88, 90, *91*, 234, *253*
Dicker, G., *91*
Dishion, T., 99, 111, 112, *113*
Doane, J.A., 162, *170*
Donati, R., 161, *168*, *169*
Draaisma, D., 73, 81, 82, *91*
Duncan, O.D., 151, *168*
Duxbury, R.A., 162, *168*

E

Eagly, A.H., 121, 136, 146, *148*
Edelberg, R., 162, *169*
Edwards, L.A., 121, *148*
Eisinger, R.A., 58–59, 66, *70*
Ellen, R.F., 51, *55*
Elwood, R.W., 162, *168*
Epstein, S., 120, *148*
Ericcson, K.A., 110, *114*
Evans-Pritchard, E.E., *55*

F

Fancher, R.E., 24, 27, *35*
Faust, D., 123, *148*
Fechner, G.T., 25, *35*
Feinberg, B.M., 58–59, 66, *70*
Feinberg, R., *55*
Feyerabend, P., 1, *9*

Filby, N., 143, *148*
Finifter, B.M., 151, 157, 158, 159, *168, 204*
Finney, J.W., 162, *168*
Firth, R., 46, *55*
Fisher, R.A., 32, *35*
Fiske, D.W., 97, 99, 101, 102, 103, 104, 107, 108, 109, 112, *113, 114,* 130, *148*
Fitzgerald, S.G., 162, *169*
Foa, E.B., 162, *169*
Fogelin, R.J., 62, *69*
Fogg, L., 112, *114*
Folds, T.H., 218, 226, *230*
Fortune, R., 47, *55*
Fraccon, I.G., 162, *168*
Freeman, D., *55,* 58, 60, *69*
Fuqua, R.W., 162, *169*
Furnham, A., 155, *168*

G

Gale, R.M., 1, *9*
Gal'perin, P.I., 210, *230*
Gardner, W., 267, *270*
Gasparikova-Krasnec, M., 151, *168*
Gellner, E., 51, *55*
Gergen, K.J., 156, *168*
Gewirtz, J.L., 256, 257, 258, 264, 266, 268, *269, 270*
Gigerenzer, G., 32, 34, *35*
Gilbert, G.N., 71, 73, 74, 75, *92*
Gillin, J., 46, *55*
Ging, N., 151, *168*
Glaser, B.N., 53, *55*
Glass, G.V., 67, *69,* 107, *114,* 123, 124, 132, 133, 143, 144, *148*
Glastra van Loon, J.F., *91*
Goldsmith, H.H., 67, *69*
Goldstein, G., 162, *170*
Goldstein, M.J., 162, *170*
Goldwyn, R., 62, *70*
Gottfried, A.W., 122, *147*
Gottman, J.M., 98, 101, *114*
Gould, S.J., 3, *9*
Grawe, K., 161, *168, 169*
Gray, C., 122, *147*
Grayson, J.B., 162, *169*
Greenwald, A.G., 6, *9,* 142, *148*
Grossmann, K.E., 61, *70,* 257, *269*
Grossmann, K., 61, *70,* 257, *269*
Grusec, J.E., 120, 142, *148*
Guillaume, P., 234, *253*

H

Habermas, J., 85, *91*
Haccoun, D.M., 155, *169*
Hacking, I., 17, *35*
Hahn, W.K., 105, *114*
Hall, C.L., 162, *168*
Hammond, M.A., 122, *147*
Hans, S.L., 162, *169*
Harlow, H.F., *70*
Harré, R., 152, *169*
Harris, B., 3, *9*
Harris, F.N., 156, *169*
Harris, M., 41, *55*
Harrison, J., 162, *169*
Hartke, D.D., 138, *149*
Hartmann, E., 162, *169*
Hasselhorn, M., 194, 201, 202, *204*
Hauri, P.J., 162, *169*
Haynes, S.N., 162, *169*
Hearnshaw, L.S., 3 , *9, 35*
Hedges, L.V., 133, 138, *148*
Heider, K., 46, *55*
Heisenberg, W., 52, *55*
Hellekson, C., 162, *169*
Hendrick, C., 4, 5, 8, *9*
Hilgard, E.R., 3, *9*
Hoag, W.J., 155, *168*
Hobbs, S., 3, *9*
Hoffman, M.L., 120, *148*
Hogarth, R., *114*
Holcomb, J., 108, *113*
Holden, D.J., 202, *204*
Holden, G.W., 121, *148*
Holmberg, A.R., 46, *55*
Holmes, L.D., 43, 44, *55, 70*
Howell, C.T., 98, *113*
Hubbard, F.O.A., 121, *148,* 257, 259, 260, *263, 270*
Humphrey, G., 20, *35*
Hunt, K., 71, 74n, 75n, *91,* 158, 163, 164, *169*
Hunter, J.E., 125, 132, 137, 138, 144, *148*
Huston, A.C., 141, *147*
Hyman, H.H., 164, *169*
Hymel, S., 98, *114*

I

Istomina, Z.M., 174, 175, 176, 177, 178–179, 180–181, *181,* 186, *204,* 210, 211, 212, 213, 215, *230*

J

Jacklin, C.N., 125, 127, 131, 140, 141, 142, *149*
Jackson, G.B., 125, 132, 137, 138, 144, *148*
Jacobson, N.S., 162, *168*
Jegerski, J.A., 7, *9*
Jenson, W.R., 156, *169*
Johnson, D.L., 122, *147*

K

Kagan, J., 98, 99, 112, *114*, 156, *169*
Kalhorn, J., 119, *147*
Kamin, L.J., 3, *9*, 58, 65, *70*, 125, *148*
Kenny, D.A., 264, *270*
Kerwin, K., 227, *230*
Kiesler, D.J., 165, *169*
Killian, G.A., 120, *147*
Kimball, M.M., 142, *148*
Klein, D.F., 153, *169*
Kloos, P., 47, 52, *55*, 60, *70*
Knorr, K.D., 81n, *91*
Kohlberg, L., *9*
Köhler, W., 233, *253*
Koren, K., 61, *70*
Kotok, A.B., 58–59, 66, *70*
Krantz, D.L., 22, *35*
Krauss, E., 108, *115*
Kroonenberg, P.M., 67, *70*
Kubica, G., 51, *55*
Kubsda, J., 189, *204*
Kuhn, T.S., 63, 66, *70*, 72, 78, *91*, 117, *148*
Kuncel, R.R., 108, *114*
Kunneman, H., 72, 76, 77, 78, 79n, *91*
Kuper, A.J., 39, *55*
Kurtz, B.E., 227, *230*

L

Lakatos, I., 79n, *91*
Lamb, M.E., 267, 268, *270*
Landau, R., 268, *270*
Larsen, K.S., 2, *9*
La Sorte, M.A., 65, *70*, 151, *169*
Latour, B., *80*, 81n, *83*, 89, *91*
Lavigueur, H., 155, *169*
Lazurski, A.F., 176, *181*
Leach, E.R., 53, *55*
Leff, J.P., 162, *170*
Leontjev, A.N. (Leontiev, Leont'ev), 185, *204*, 210, 211, *230*, 233, 234, 235, 237, 238, 239, 240, 245, 246, 247, 248, 252, 253, *253*
Lester, B.M., 268, *270*
Leventhal, D., 162, *170*
Levin, H., 120, *149*
Levinson, D.F., 162, *169*
Lévi-Strauss, C., 47, 55
Lewis, C.C., 122, *148*
Lewis, O., 44, 45, 49, 50, *56*
Lewontin, R.C., 3, *9*
Li An-Che, 45, *56*
Light, R.J., 122, *148*
Lilienfeld, A.M., 29, *35*
Lind, J., 261, *270*
Lord, F.M., 265n, *270*
Louis, C., 161, *168*
Lovie, A.D., 32, *35*
Luria, A.R., 234, 241, *254*
Lykken, D.T., 65, *70*, 71, *92*
Lytton, H., 120, 127, 141, 142, *148*

M

Maccoby, E.E., 120, 125, 127, 131, 133, 140, 141, *147*, *148*, *149*
Mackenzie, D.A., 24, 31, *35*
Mahoney, M.J., *9*
Main, M., *62*, *70*
Malinowski, B., 50, *56*
Marcus, J., 162, *169*
Maricq, H.R., 162, *169*
Marquis, D.G., 3, *9*
Marsh, H.W., 104, *114*
Martin, J.A., 120, *148*
Matt, G., 132, 133, 144, *148*
McCartney, K., 67, *70*, 141, *149*
McCluskey, K., 257, *269*
McConaughy, S.H., 98, *113*
McElroy, M.G., 162, *170*
McGaw, B., 67, *69*
McKinley, T., 162, *169*
Meacham, J.A., 187, *204*
Mead, M., 37, 41, 43–44, 47, *56*, 60, *70*
Mednick, S.A., 162, *169*
Meehl, P.E., 123, *148*
Mehner, M., 22, *35*
Messick, S., 108, *114*
Meyerson, I., 234, *253*
Michelse, N., 162, *169*
Miedema, S., 65, *70*
Milgram, S., 156, *169*
Miller, T.I., 107, *114*, 132, 133, 144, *149*
Miltenberger, R.G., 162, *169*
Mitchell, S., 122, *147*

Mitroff, I.I., 154, *169*
Moore, A.W., 22, *34*
Mosteller, F., 123, *149*
Mucha, J., 51, *55*
Mulkay, M., 61, 64, *70*, 71, 73, 74, 75, *92*
Mullen, B., 135, 138, *149*
Munsinger, H., 125, *149*
Murray, A.D., 268, *270*
Murray, B.J., 32, *35*
Murray, H.A., 101, *114*
Musgrave, A., 59–60, *70*

N

Nagel, E., *92*
Naus, M.J., 225, *230*
Neisser, U., 187, *204*
Neuliep, J.W., 6, *9*
Norton, B., 26, *35*
Norton, R.S., 59, *70*
Nurius, P.S., 144, 146, *149*

O

O'Brien, M., 141, *147*
O'Connor, E.F., 265n, *270*
Olkin, I., 133, 138, *148*
O'Meary, M., 162, *169*
Ornstein, P.A., 202, *204*, 218, 225, 226, *230*
Orwin, R.G., 133, 139, *149*
Oster, G.D., 3, *9*

P

Paris, S.G., 187, 188, 189, 193, 198, 200, 202, *205*, 209, 214, *231*
Parker, J.G., 98, *114*
Partanen, T., 261, *270*
Patterson, G.R., 99, 111, 112, *113*
Paul, D.B., 3, *9*, 58, *70*
Payer, L., *8*, *9*
Peacock, J.L., 52, *56*
Peirce, C.S., 74, *92*
Percy, L., 162, *169*
Perlmutter, M., 187, 202, *205*
Pesner, R., 151, 156, 158, 160, *170*
Peters, L.H., 138, *149*
Petrovich, S.B., 268, *270*
Phillips, D., 67, *70*
Piaget, J., 157, *169*
Pohlmann, J.T., 138, *149*
Polyani, M., 63, *70*
Popper, K.R., 17, *35*, 57, 58, 60, 63, 64, *70*, 73, 77, 78, *92*
Porter, T.M., 24, *35*
Powdermaker, H., 43, 53, *56*
Pressley, M., 187, 188, *205*
Prinz, R.J., 98, *115*
Prytula, R.E., 3, *9*

Q

Quetelet, L.A., 24, *35*

R

Radder, H., 84, 85, *92*
Radin, P., *56*
Ramey, C.T., 122, *147*
Rapoff, M.A., 162, *168*
Rasinski, K.A., 108, 110, *114*
Rayner, R., 2, *9*
Redfield, R., 49, 50, *56*
Reese, S.C., 65, *69*
Reid, L.N., 6, *9*
Reiss, A.F., Jr., 108, *114*
Roberts, R.R., 6, *9*
Rock, S.L., 122, *147*
Rodrigues, A., 151, *169*
Rogosa, D., 265n, *270*
Romney, D.M., 127, 141, *148*
Rose, S., 3, *9*
Rosenthal, R., 4, 5, 7, *9*, 128, 133, 135, 138, *149*
Ross, D.C., 153, *169*
Rothbart, M.K., 131, *149*
Rovine, M., 257, *269*
Rubin, K.H., 98, *114*
Rucci, A.J., 32, *35*
Rusk, R.R., 30, *35*
Russ, D., 162, *169*

S

Sagi, A., 61, *70*
Salkovskis, R.M., 162, *169*
Samelson, F., 3, *9*, 58, *70*
Sandahl, C., 162, *169*
Sattath, S., 109–110, *115*
Scabini, E., 162, *170*
Scarr, S., 141, *149*
Schlenker, B.R., 164, *170*
Schmidt, F.L., 125, 132, 137, 138, 144, *148*
Schmitt, N., 104, *114*

Schneider, W., 187, 188, 191, 192, 193, 194, 200, 201, 202, *204*, *205*, 227, *230*
Schulsin, F., 162, *169*
Schultz, D.P., 22, *35*
Sciortino-Brudzynski, A.P., 236, 237, 241, *242*, *243*, *244*, *245*, *246*, 247, 248, 249, 250, 251, 253, *253*
Sears, R.R., 120, *149*
Sharon, I., 4, 7, *8*
Shaver, J.P., 59, *70*
Shelley, D., 161, *170*
Shelly, C., 162, *170*
Shute, G.E., *169*
Sidman, M., 29, *36*
Siegel, L., 122, *147*
Siegfried, J., 152, 161, *168*, *169*, *170*
Simon, H.A., 110, *114*
Sjöberg, L., 156, *170*
Skinner, M., 99, 111, 112, *113*
Slavin, R.E., 143, *149*
Sleeper, R.W., 86, *92*
Sleight, W.G., 30, *36*
Slotkin, J.S., 39, *56*
Slovic, P., 109–110, *115*
Smedslund, J., 156, *170*
Smith, G.P., 151, *170*
Smith, M.L., 67, *69*, 107, *114*, 132, 143, 144, *148*, *149*
Smith, N.C., 60, *70*, 151, *170*
Smith, P., 257, *269*
Smith, P.V., 122, *148*
Snow, M.E., 141, *149*
Sodian, B., 202, *205*
Soley, L.C., 6, *9*
Solomon, R.L., 30, *36*
Spangler, G., 257, *269*
Spearman, C., 26, *36*
Sroufe, L.A., *270*
Srull, T.K., 108, *115*
Stanley, J.C., 151, *168*
Star, S.A., 151, *170*
Stearman, A.M., 46, *56*
Steer, R.A., 162, *170*
Steketee, G.S., 162, *169*
Stevenson, R.L., 58–59, 66, *70*
Stocking, G.W., 39, *56*
Stouffer, S.A., *170*
Strachan, A.M., 162, *170*
Strauss, A.L., 53, *55*
Stults, D.M., 104, *114*
Suchman, E.A., 151, *170*
Suess, G., 257, *269*
Syvantek, D.J., 151, *168*

T

Tanaka, J.S., 263, *270*
Tavecchio, L.W.C., 268, *170*
Taylor, D.G., 257, *269*
Tharp, R.G., 236, 237, 241, 242, 243, 244, 245, 246, 247, 248, 249, 250, 251, 253, *253*
Thompson, R.A., 267, 268, *270*
Thurstone, L.L., 28, *36*
Tibbets, P., 79, *92*
Titchener, E.B., 21, *36*
Toulmin, S., 78, *92*
Tourangeau, R., 108, 110, *114*
Travis, G.D.L., 61, 62, 63, *70*
Turner, C.F., 107, 108, *115*
Tversky, A., 109–110, *115*
Tweney, R.D., 32, *35*

U

Unzner, L., 257, *269*
Urry, J., *56*

V

Valanné, E., 261, *270*
Valsiner, J., 2, 4, *9*, 14, *15*, 156, *170*, 235, *253*
VanderStoep, L.P., 120, *147*
van der Veer, R., 14, *15*, 235, *253*
van IJzendoorn, M.H., 61, 67, *70*, 121, *148*, 259, 263, 268, *270*
Vaughn, C.E., 162, *170*
Visser, R., 265n, *270*
Vuorenkoski, V., 261, *270*
Vygotsky, L.S., 233, 234, 241, *254*

W

Waal Malefijt, A. de, 38, *56*
Wade, N., 37, *55*
Wagenaar, W.A., 2, *9*
Walker, H., 31, *36*
Wall, S., 60, 62, 67, *69*, 255, *269*
Walsh, D., 162, *168*
Waters, E., 60, 62, 67, *69*, *70*, 255, *269*
Wasz-Höckert, O., 261, *270*
Watson, J.B., 2, *9*
Webb, N., 66, *69*

Weiner, A.B., 51, *56*
Weissberg, J.A., 187, 188, 189, 193, 198, 200, 202, *205*, 209, 214, *231*
Wells, D.S., 162, *170*
Werner, H., 210, *231*
Wertsch, J.V., 187, *205*, 210, *231*, 233, *254*
West, M.O., 98, *115*
White, D.R., 151, 156, 158, 160, *170*
Williams, R.H., 265n, *270*
Williams, R.M., 151, *170*
Wimmer, R.D., 6, *9*
Winch, W.H., 30, *36*
Wippich, W., 188, 189, *205*
Woerner, M.G., 153, *169*
Wolff, P.H., 263, 268, *270*
Wolins, L., 65, *70*
Wood, W., 106, *115*
Woolgar, S., 80, 81n, 83, 89, *91*
Wortman, P.M., 66, *69*
Wothke, W., 104, *115*
Wozniak, R.H., 157, *170*
Wright, C.R., 164, *169*
Wundt, W., 19, *36*
Wyer, R.S., Jr., 108, *115*

Y

Yarrow, M.R., 118, 119, 120, *149*
Yeaton, W.H., 144, 146, *149*

Z

Zaporozhets, A.V., 210, *231*
Zeman, B.R., 226, *230*
Zimmerman, D.W., 265, *270*
Zimowski, M., 265, *270*
Zinchenko, P.I., 210, 211, *231*
Zitrin, C.M., 153, *169*

Subject Index

C

Collective object, 25
Constructive empiricism, 84
Correlation, 25–28
Cross-sectional design, 177, 252, 274

D

Differential responsiveness, 268

E

Effect size, 67, 107, 132–140, 143–146
Enculturation (indoctrination) model, 62, 112
Epistemic subject, 157
Experimenter regression, 63

F

Fieldwork as *rite de passage*, 41
File drawer problem, 7
Fraud, 2, 3

L

Longitudinal design, 177–178, 267, 274

M

Meta-analysis, 65–69, 94, 106, 117, 122–147, 151, 223, 271, 276–278, 281
Method effects, 94, 99–104, 107–112
Moderator variables, 129
Multitrait–multimethod model, 4, 93–94, 99, 104–105, 107
My tribe syndrome, 42

N

Nature–nurture, 280
Normal law of error, 32

P

Pseudoreplications, 159–160

R

Referential realism, 84–86

Replication
conceptual (constructive, imprecise, inexact, varied), 4–6, 8, 62, 65, 68–69, 71, 90, 158, 161, 163–165, 167, 201, 271–272, 277–279, 281
of conclusions, 12, 75–76, 79
and ontology, 52, 72, 80
of results, 12
of setup, 12
strict (algorithmic, carbon copy, literal, precise, virtually exact), 4–6, 8, 14, 54, 57–62, 64–65, 69, 71, 74–75, 81, 118, 157, 161, 163–164, 167, 174, 188, 201, 269, 271–272, 276–279
Replication studies (frequency of), 6, 59, 93, 160–163
Researcher as human instrument, 43–44, 53
Restudy, 38, 44–47

S

Secondary analysis, 65–66, 122–123, 138, 269, 277, 281

T

Transactional realism, 84, 86–90