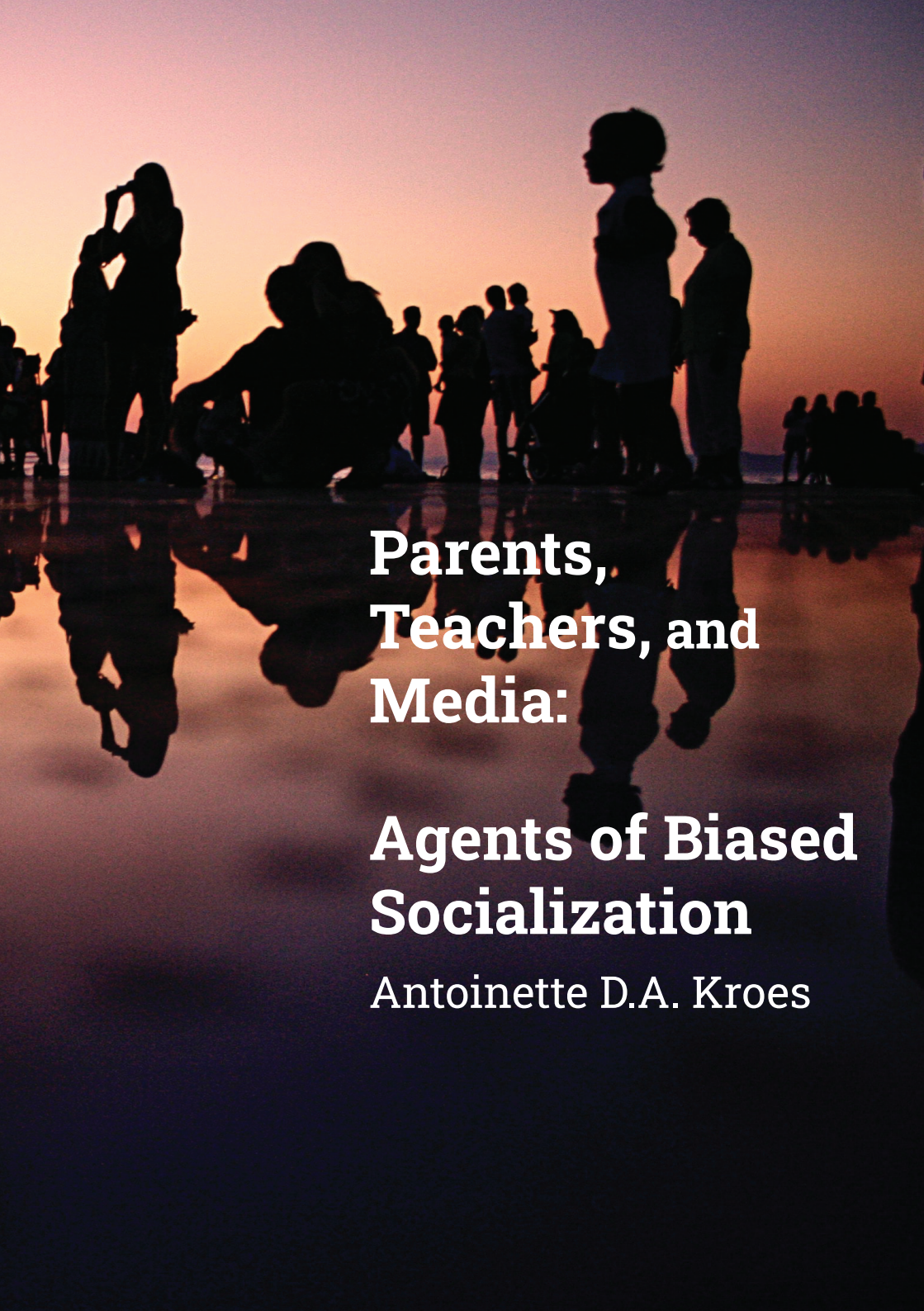
Parents, teachers, and media: agents of biased socialization

Kroes, A.D.A.



Parents, Teachers, and Media:

Agents of Biased Socialization

Antoinette D.A. Kroes

Parents, Teachers, and Media: Agents of Biased Socialization

Antoinette Dana Alice Kroes

Financial support for the printing of this thesis was kindly provided by
Fontys Opleidingscentrum Speciale Onderwijszorg
Leiden University

ISBN: 978-94-6469-591-5

All pictures via Unsplash.com, with many thanks to the photographers:
Mario Purisic (cover), Kelly Ann Tan (p. 8), Mariusz Slonski (p. 20), Patrick Hendry (p. 50), Adib Harith Fadzilah (p. 104), Jr Korpa (p. 144), Claudio Schwarz (p. 182), Harli Marten (p. 204), Bethany Beck (p. 226), Hugo B. Leohgh (p. 272), and Oudalay Sengkhounmany (p. 290).

Parents, Teachers, and Media: Agents of Biased Socialization

PROEFSCHRIFT

ter verkrijging van
de graad van doctor aan de Universiteit Leiden,
op gezag van rector magnificus prof.dr.ir. H. Bijl,
volgens besluit van het college voor promoties
te verdedigen op 22 november 2023
klokke 16:15 uur

door

Antoinette Dana Alice Kroes

Promotor:
Prof. dr. Judi Mesman

Copromotoren:
Dr. Marleen G. Groeneveld
Dr. Lotte D. van der Pol

Promotiecommissie:

Prof. dr. Sandra Groeneveld	Leiden University
Prof. dr. Eddie Denessen	Radboud University, Leiden University
Prof. dr. Greetje Timmerman	University of Groningen
Prof. dr. Christia Spears Brown	University of Kentucky
Dr. Linda van den Bergh	Fontys University of Applied Sciences

The research described in Chapters 2, 3, and 4 of this dissertation was supported by the European Research Council Consolidator Grant awarded to Judi Mesman (project number 726141). The study described in Chapter 5 was supported by the European Research Council Starting Grant awarded to Judi Mesman (project number 240885).

Voor Yilmaz, Bayram, Pieter, Aziz, Pawlik, Karel en Ben.

Voor altijd dankbaar voor hoe jullie me verwelkomden, lieten zien wat echt belangrijk is, en me hielpen om de eindstreep te behalen.
Jullie zijn onmisbaar.

Table of Contents

Preface

Reflexivity statement

> "Because of the big differences between boys and girls, sexuality for boys is a reality they begin to experience when puberty commences, while for girls sexuality in principle remains a closed book until the time comes that they are approached physically. The fairytale of Sleeping Beauty is based on this: she is awakened by the lover's kiss by the prince of her dreams. [...]
> The girls who experience [a short-term fling] for the first time are surprised by an *aha-erlebnis*, a sort of new world opens for them, they get butterflies in their stomachs. However, it results in nothing, it only gives a hangover. They are overcome by restlessness, a new desire for security and more, but it does not find a place, it is not fulfilled. Through these types of experiences, girls are very easily infected with some sort of 'boy craze', and because of that, girls are in danger of settling very easily for boys and men who present themselves because they have an urge.
> When a girl is awakened sexually, she more or less loses her objectivity that she naturally had. That doesn't just concern a physical desire, but also a broader and deeper desire, of her entire being (Baarsen, n.d., translated by the author)."

This is an excerpt from a pamphlet that I received in high school when I was 13 years old. I grew up in a relatively conservative Christian family, and attended churches and schools that were similarly conservative. Sexuality was a topic that was typically dealt with in a conservative matter, compared to contemporary Dutch standards. Both at school and in church, sexual education was focused on abstinence until marriage (Fine, 1988; Fine & McClelland, 2006), and there were clear double

standards about sexuality. The excerpt from the pamphlet above is an example of these types of double standards, in which girls are presented as asexual and passive beings in contrast to boys, who are sexual and active. Later in the pamphlet, the author goes into this further, explaining that men are meant to slowly guide and initiate women into sexuality. Women, according to the author, are naturally monogamous beings who can give away their virginity only once. When women are subsequently abandoned by men, they lose some of their capacity to attach themselves to someone and to form new relationships (Baarsen, n.d.). It is perhaps no surprise that texts like these had a big impact on me, being a teenager who was infatuated with someone new every few weeks. When I was introduced to feminist theory during a minor on gender and sexuality ten years later, I became very invested in examining these types of bias.

Some researchers may be attached to wearing a metaphorical cloak of objectivity, professing that their research practices are the result of purely rational considerations (Hamby, 2018). Nevertheless, when humans are involved, objectivity does not exist. Sharing a reflexivity statement, in which a researcher reflects on the position they adopt in relation to their research, is common in qualitative research, but often unexpected from quantitative researchers (Darwin Holmes, 2020; Rios-Aguilar, 2014). This was reflected in how objectivity and positionality were regarded during my bachelor studies, in which the focus was mainly on quantitative methods. Objectivity and positionality were not often discussed explicitly, and discussions mainly concerned how to make 'objective' observations. One of the methodological handbooks that was part of the mandatory readings states that in quantitative research "objectivity is sought and distance maintained between the researchers and participants" (Robson, 2011, p. 19). Conversely, for qualitative research, we can read on the same page "objectivity is not valued. It is seen as distancing the researchers from participants". Later on in the same book, novice researchers are warned that caring too much about helping research participants and answering research questions, puts the

researcher at risk of losing their objectivity and ability to evaluate data fairly. According to the author, researchers who only care about doing 'good research' (which apparently does not involve helping people and answering research questions) do not run that same risk (Robson, 2011). This is in stark contrast with the perspective of other academics, who see social justice and striving for equity as fundamental for social/behavioral science and education, especially when involving children (Brown et al., 2019; Grant & Agosto, 2008; Nieto, 2000). The position held by a (quantitative) researcher shapes much more than just the (evaluations of) observations they make. Their position is involved with the entire research process, including activities such as choosing the research topic, employing specific theories, designing the study, applying statistical techniques, and interpreting results. Therefore, by disclosing one's position, the validity and meaning of our research can be enhanced (Rios-Aguilar, 2014). The way I was raised, for instance by being socialized into the idea that sex should be different for girls than for boys, has been important in directing my position as a researcher. My position has further been influenced by many other experiences. For instance, I worked with kindergarten pupils at the time one of the biggest child sexual abuse cases in the Netherlands broke in the media, which made a big impression on me. Later, when I taught children in the highest grade of primary school, my Black male pupils would sometimes tell me how police officers would order them to show their ID while playing outside, even though my pupils were too young to have an ID. Most White people are never asked to provide identification, especially not White children. This ethnic profiling of children, who were very dear to me, fueled my desire for a fairer society. On the following pages, I will further reflect on my position, following the guidelines proposed by Sherry Hamby (2018).

I am a White Dutch cisgender[1] woman. Growing up female shaped my position immensely, and has been an essential aspect of my identity as a feminist, and my interest in studying gender. I was born in the Netherlands, as were my parents and grandparents. I grew up privileged, in a middleclass household in the Randstad, a conurbation in the central-western part of the Netherlands. Privilege stemming from my background is reflected for instance in how visiting libraries and being read to by my mother were part and parcel of my childhood. The attainment of an academic degree was a matter-of-course.

Before I started my PhD trajectory, I worked as a primary school teacher. I thoroughly enjoyed working with children, and was constantly made aware of how demographic characteristics like gender and ethnicity play a big role in their lives. For instance, in kindergarten, there was an omnipresent pink/blue gender division in the classroom, fueled by the popularity of animated movies and television shows like Frozen and Cars. Young children would also observe each other's appearances and wonder why some had darker skin, whether it was possible for a Black girl to be a princess, and why some children's eyes were shaped differently than others. In higher grades, children asked profound questions, such as: "Why do we call it football when boys do it, but *girls'* football when girls do it?" and "Why is it that when girls do boys' stuff they are cool, but when boys do girls' stuff they are losers?" I saw how some children were negatively impacted by biased expectations and prejudice. Through these experiences, my desire to act against and study various forms of bias grew. From 2017 to 2022, I worked within a larger interdisciplinary team, in an overarching longitudinal project, with colleagues who have a background in social psychology, developmental psychology, sociology, political science, and public administration. The overall goal of the longitudinal

[1] Cisgender indicates the congruence between gender identity and sex assigned at birth, e.g., I identify as female and was assigned female at birth.

project was studying gendered socialization of adolescents in the school and family context, and the gendered pathways of vocational interest development and educational choices of adolescents. During our longitudinal project, the world was hit with the COVID-19 pandemic. This had a big impact on our data collection and on the possibilities to recruit participants for our studies.

My position as a privileged White cis-gender woman, who is driven by a deep desire for social justice, can cause bias and oversights. In order to prevent bias and oversights as much as possible, I have collaborated with various people, both inside and outside of the interdisciplinary research team that I was part of. I have had very fruitful discussions with and received indispensable information and feedback from my team members, as well as others, including researchers, statisticians, methodologists, students, non-academic professionals, experts with first-hand experiences of intersecting marginalization, reviewers, editors, and the ethical committee of Leiden University. When working with an ethnically diverse sample (see Chapter 4), I specifically worked with experts on multi-ethnic communication, and with a multi-ethnic team of coders.

I have been diagnosed with ADHD and I believe that this has had an important influence on me as a researcher. Some of the symptoms of ADHD have been a blessing during my past years as a doctoral student. When my attention is grabbed I can dive into hyperfocus and spend hours upon hours investigating and reading about many topics. I am easily enthused for new projects. This is perhaps reflected in my somewhat eclectic dissertation. Still, my neurodiversity also leads to obstacles in life. It has been estimated that when a child with ADHD is 10 years old, they will have received thousands of corrective or negative statements like "Pay attention!" and "Get back on task!" in school alone (Jellinek, 2010), way more than their neurotypical classmates. Having been on the receiving end of many of these types of messages has peaked my interest in investigating the messages that teachers and parents communicate to

their children. Regularly, these corrective comments are gendered. I distinctly remember my 6th grade teacher holding up my somewhat messy attempt at bookbinding and exasperatedly exclaiming, "Antoinette, I thought you were a *girl*!?"

I am glad that now, over 20 years later, gendered messages like these have led me unto a path that resulted in this dissertation.

Terminology and binaries

In the following six chapters, I will give a general introduction to my dissertation and then present four studies for which I was the first author, and for which I collaborated with various co-authors. The last chapter consists of a general discussion. In this preface, the general introduction, and the general discussion I use singular first-person pronouns to describe my position, experiences, and ideas. In the following chapters of my dissertation I will use the plural first-person pronoun "we" when describing all research activities that were carried out with my co-authors.

In my dissertation, gender is the most important central concept. While the term sex commonly refers to biological sex, gender is constructed socio-culturally (R. W. Connell, 2005). Gender and sex can be used to label and interpret human behavior (American Psychological Association, 2020). Sex and gender are often presented as more or less inseparable, and as a binary: man and woman, or masculine and feminine (Hyde et al., 2019). In concordance with Hyde et al. (2019), I will refer to this binary as the gender/sex binary. The gender/sex binary forms meaningful social contexts in life. Within these contexts, people are implicitly sorted into certain binary gendered social roles, like the nurturing mother and the providing father, and the male construction worker and the female nurse. Without intervention, this process of social sorting often results in gender stereotypes, prejudice, and sexism (Hyde et al., 2019; Morgenroth & Ryan, 2018). This is to the detriment of people, albeit not to the same extent for all. Negative effects are greater for those

who do not conform to gendered expectations in society, and girls and women are affected more negatively than boys and men (Leaper & Brown, 2018). It is noteworthy that while all people are affected at least in some way by gendered inequality, for instance by gendered expectations and the gendered division of labor and care, many people do not desire any further emancipation (Kanne & van der Schelde, 2022).

Both on the biological and on the socio-cultural level, many variations beyond the sex and gender binaries exist (Fausto-Sterling, 2000). Additionally, the sex/gender binary is an important part of heteronormative frameworks, in which one is assumed to be born as either male or female, develop either a masculine or feminine gender identity, and be sexually attracted to someone of the opposite sex, with the opposite gender identity (Hyde et al., 2019; Martínez-San Miguel & Tobias, 2016). It is important to note that non-cisgender people and members of the LGBTIQ+[2] community experience greater marginalization through (cis)sexism than cisgender straight people (Fine & McClelland, 2006; Huijnk & van Beusekom, 2021; Leaper & Brown, 2018; Martínez-San Miguel & Tobias, 2016; Van Beusekom & Kuyper, 2018). This is one of the reasons that over the past decades, the number of studies that pay attention to the variations that go beyond the sex/gender binary has increased (Hyde et al., 2019; Leaper & Brown, 2018). Other reasons include that identities beyond the sex/gender binary are the reality, and excluding participants with these identities can contribute to their marginalization. Moreover, limiting research to the sex/gender binary impedes the discovery and development of new theories.

Throughout my dissertation, I use the sex/gender binary to label the participants as either male or female, and to measure differences between

[2] LGBTIQ stands for Lesbian, Gay, Bisexual, Transgender, Intersex, and Queer. The + indicates all other variations of (non hetero) sexual attraction and (non cis) gender expression, including people who identify as non-binary.

them. During recruiting and data collection, we asked the participants to identify with binary gender categories (e.g., "Are you a boy or a girl?"). We did not ask participants whether they were cisgendered. The omission of participants whose identity goes beyond the gender binary, is due to two reasons: firstly, for the samples in the studies presented in my dissertation I was dependent on two larger longitudinal research projects (Chapter 3, 4, and 5). The overarching aim of these projects involved the study of gendered socialization within families with a father, a mother, and (at least) two children. A substantial part of this sample had been recruited 7 years prior to the start of my PhD trajectory. Exclusion criteria included being raised outside of the Netherlands, and families with same-sex parents. To enable the possibility for an accelerated research design in the subsequent overarching longitudinal project that I was part of, the same exclusion criteria were used for the new participants. This limited the possibilities to include participants with identities that go beyond the sex/gender binary. Secondly, for the studies that employ statistical analyses, recruiting enough extra participants in various gender categories to reach sufficient power was not feasible. I want to acknowledge that this is a limitation in my dissertation. Studying the experiences and attitudes of people with non-binary and non-straight sex/gender identities is important, as they experience more and different forms of marginalization than those who fall within the heteronormative binary (Fine & McClelland, 2006; Huijnk & van Beusekom, 2021; Leaper & Brown, 2018; Martínez-San Miguel & Tobias, 2016; Van Beusekom & Kuyper, 2018). Additionally, people who do not conform to the sex/gender binary can be a rich source of information on how to dismantle gender stereotypes, prejudice, and sexism.

In Chapter 4, I have examined both gender and ethnic bias. In the general introduction of my dissertation, I therefore pay attention to this latter type of bias as well, albeit to a lesser extent than to gender bias. Similar to gender, I have employed ethnicity as a binary construct, as there was not enough power to differentiate further between the ethnic

backgrounds of the participants from ethnic minority groups. I am aware that working with these binary categories can obscure forms of intersectional bias (Crawford et al., 2019). However, simultaneously, the sex/gender binary does lie at the root of most gender stereotyping and gender discrimination (Hyde et al., 2019). In the case of ethnicity, any person from an ethnic minority background can be seen as 'the Other' by members of the dominant ethnic group, without distinctions regarding specific backgrounds (Jensen, 2011; Van Schie, 2018). Therefore, studying bias along these binaries is still a worthwhile endeavor.

Chapter 1

General Introduction

Many adolescents are confronted regularly with deeply misogynistic messages through social media platforms (Weimann & Masri, 2020), on which popular influencers state that men are better than women, and that women 'belong at home' and should 'make men a sandwich' (Fazackerley, 2023). In short format videos, especially aimed at boys, famous content creators say that women are a man's property, and (female) rape victims are responsible for the crimes committed against them. Recently, many schools have reported that these influencers are hugely well known among their pupils (Fazackerley, 2023). Teachers state that they are worried about how this affects the adolescents they educate. When social media apps are studied, reinforcement of stereotypical masculinity and femininity are found (Foster & Baker, 2022; Kennedy, 2020). For instance, in content related to self-harm, suicide, and eating disorders women and girls appear far more often than men and boys (Arendt, 2019; Minadeo & Pope, 2022). The selection of content presented to users is fueled by algorithms (Bueno-Fernandes & Campos-Pellanda, 2022; Das, 2022). Users are presented with gendered videos automatically, even without liking or searching for specific content. For instance, while both male and female users are confronted with heteronormative content, male users are more likely to see violent and sports related videos, and female users are more likely to see content related to skin care and sexuality (Bueno-Fernandes & Campos-Pellanda, 2022).

Gender-based prejudice and discrimination are well-known problems not just in social media but in many aspects of life (Leaper & Brown, 2018). Women structurally experience mistreatment based on their gender. They are, for instance, much more likely than men to experience sexual harassment, sexual violence (Laan et al., 2021; Leaper & Brown, 2018), domestic/partner violence (Dutch Central Bureau of Statistics, 2022a), earn less (Dutch Central Bureau of Statistics, 2022d), and be shamed and penalized for their sexual behavior (Endendijk et al., 2020). On the other hand, while cisgender men often reap benefits of gender inequality in society, there are many disadvantages for them too. Examples of these

disadvantages include that men receive far less paid parental leave than women, giving them less opportunities to build relationships with their children (Tamm, 2019). Men are more likely to experience physical assaults (Dutch Central Bureau of Statistics, 2022a), and are at much higher risk for suicide (Standish, 2021).

Gender inequality is pertinent in many aspects of children's lives too. For instance, in early childhood, they already perceive limits to what toys they can play with, according to their gender (Freeman, 2007). Children also experience gender inequality in schools. Gender-based grading bias has been found in many studies, with boys receiving lower grades than girls on non-standardized tests while there is no actual difference between their achievement (Voyer & Voyer, 2014). Teachers often have more positive attitudes towards girls than boys (Glock & Klapproth, 2017; Myhill & Jones, 2006), while boys receive more attention (Bassi et al., 2016), and higher quality feedback (Bašaragin & Savic, 2019). During sexual education, the penis is prioritized and the full anatomy of the clitoris is often omitted completely, even though it has been known for centuries (Gerritsen, 2022; Laan et al., 2021). For children from ethnic minority backgrounds, gender inequalities can be different than for the ethnic majority. Negative bias against ethnic minority children is found regularly (Geven et al., 2018; Wang et al., 2018), and ethnic minority pupils do not appear to reap the benefits from their gender like ethnic majority pupils do (Glock & Klapproth, 2017; Menegatti et al., 2017; Musto, 2019). When children experience negative bias, this can have various adverse effects on their achievements, behavior, and various socio-psychological aspects like self-esteem and motivation (Wang et al., 2018).

As gender and ethnicity based bias have many adverse effects on the lives of people, it is important to study the mechanisms that create and perpetuate bias. In this dissertation, I will do this by focusing on socialization agents. These are people and entities that model and instruct (potentially biased) behaviors and attitudes to individuals (Bandura, 1986). By investigating powerful socialization agents (parents, teachers,

and mass media; Zaikman & Marks, 2017) I aim to contribute to a deeper understanding into the mechanisms that underlie gender and ethnic bias. A better understanding of these mechanisms is beneficial in combatting bias and creating a fairer society.

Developing inequalities and bias

Because bias can lead to inequalities, it is important to understand how it works. Many theories explain the origin, development, and transmission of prejudiced and biased attitudes and behaviors (for overviews of several of these theories, see for instance: Bussey & Bandura, 1999; Levy & Hughes, 2009; Rutland et al., 2010; Zaikman & Marks, 2017). In this section, I will highlight and connect three of these theories that are relevant for my dissertation: constructivist-ecological perspective, social cognitive learning theory, and social role theory. Theoretical viewpoints are important for good research design, gaining deeper understanding into underlying mechanisms, and for theoretical integration across different areas of social and behavior science (Leaper, 2011; Zaikman & Marks, 2017). The constructivist-ecological perspective and social cognitive learning theory are common in developmental psychology, while social role theory is mostly used in sociology and social psychology.

Constructivist-ecological perspective

In the constructivist-ecological perspective, the development of biased attitudes is explained by drawing from constructivist theorists on the one hand, and from the ecological model of Urie Bronfenbrenner on the other (Liben, 2017).

Underlying constructivism is the premise that human development is cumulative. Knowledge is built upon previous knowledge, and prior knowledge affects how knowledge is modified or how new knowledge is constructed (Philips, 1995). Central to learning and development is active participation by the learner, as well as the social nature of learning (Leaper & Bigler, 2018; Liben, 2017). According to Lev Vygotsky, a

quintessential constructivist (Philips, 1995), the construction of knowledge always takes place on two planes: first on the social/interpersonal level, and then inside the individual, on the intrapersonal level (Duveen, 1996; Vygotsky, 1978). Missing in the works of Vygotsky is the notion that the construction of knowledge is not the same for each learner, as social groups and social identities play a role as well (Duveen, 1996). A similarly quintessential constructivist is Jean Piaget (Philips, 1995), who developed a comparable constructivist theory but with more attention to the individual qualities of the learner (Liben, 2017). Central to the work of Piaget are the concepts of assimilation and accommodation (Piaget, 1954). Assimilation refers to the process in which new information is made sense of and incorporated with existing knowledge and schemata of the learner. Accommodation refers to the way in which the learner modifies and adjusts their current understanding in response to new information. For example, on my first day as a teacher in kindergarten, one of my pupils enthusiastically shouted: "You are our teacher and you are also a mother!" She assimilated my gender and age with her idea of adult women who, in her experience, were always mothers. When I explained to her that I was indeed an adult woman, but that I did not have children, she accommodated her idea of adult women to include non-mothers. For the remainder of the year, she did not refer to me as a mother anymore. She had encountered this new knowledge first on the social level, and then internalized it, on the intrapsychic level. This exemplifies how constructivist theory can be applied to the construction of gender-related cognitions specifically (Duveen, 1996; Kohlberg, 1966; Liben, 2017). Children acquire their beliefs, behaviors, and cognitions from the social world around them, influenced by what they have acquired previously.

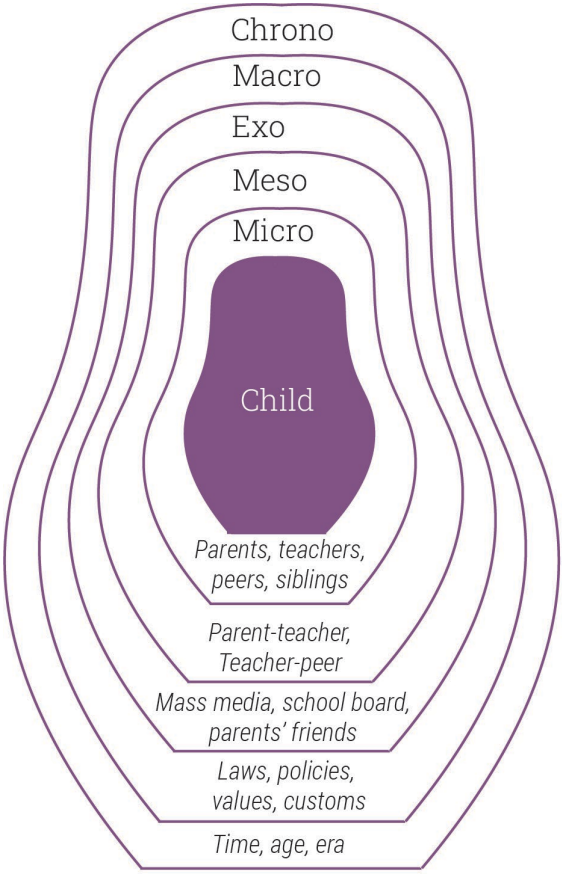
The importance of social contexts is further explained in the ecological model that was developed by Urie Bronfenbrenner (1979). Bronfenbrenner defined development as "a lasting change in the way in which a person perceives and deals with his environment (p. 3)." In the

ecological model, the developing person stands at the center. The model consists of five types of systems, each type is nested within the next, engulfing the developing child like Russian nesting dolls (see Figure 1.1; Bronfenbrenner, 1979). I will present the five types of systems below, illustrating for each system how they are involved with biased attitudes and/or behaviors with examples from literature. Working outwards from the center, the five types of systems are (Bronfenbrenner, 1979; Liben, 2017; Perry et al., 1993):

1) Microsystems: are the direct and immediate interactions in which the child is involved, like with parents and siblings in the home and with peers and teachers in the classroom. For example, when boys are penalized by their peers for not adhering to the standards of masculinity, they are likely to adjust their behavior to fit stereotypical ideas of masculinity and to avoid being stigmatized (Pascoe, 2007). Likewise, girls will adjust their behavior when their femininity is criticized (Allan, 2009).
2) Mesosystems: are the interconnections between the micro systems, like the connections between parents and schools, or between teachers and peers. For instance, when teachers show higher levels of ethnic prejudice, ethnic majority children are more likely to exclude and bully peers with a refugee background (Alan et al., 2021).
3) Exosystems: consist of one or more settings in which the child is not an active participant, and with which the child has no interactions, but which can still influence the child indirectly. These include mass media, the school board and its policies, parents' place of employment, and parents' friends. For example, mass media can reinforce homophobic attitudes, which can result in homophobic bullying (Hong & Garbarino, 2012), while positive portrayals of gay people can lead to more positive attitudes (Schiappa et al., 2006).

Figure 1.1

Bronfenbrenner's Ecological Systems Model



Note. The ecological systems model consists of five types of systems that engulf the developing child like nesting dolls. The names of the five types of systems are seen at the top of the model, examples of the systems are in italics at the bottom of the model.

4) Macrosystem: is the collection of broader systems like culturally based belief systems, economic system and the political systems in society. These systems influence the child's daily life through laws, values, policies, and customs. For example, in 2012 the Dutch government made an addition to the mandatory curriculum: all primary and secondary school pupils are to learn to respectfully interact with diversity within society, including sexual diversity (Bron et al., 2015).
5) Chronosystem: encompasses the role that time plays in the development of the child, both in the aging of the child as through the historical eras and zeitgeist. For instance, children's stereotypical attitudes typically become more flexible with age, as their cognitive skills develop (Liben & Bigler, 2002). On a historical scale, premarital and casual sex have become more acceptable over time for both men and women in countries with high levels of gender equality, while women are judged more harshly for this behavior in other countries (Endendijk et al., 2020; Wouters, 2013).

As reflected by the examples given for each of the systems, there is support for the constructivist-ecological perspective on the development of multiple kinds of intergroup bias and stereotypes (Bigler & Patterson, 2017; Liben, 2017).

Social cognitive learning theory

Social cognitive learning theory offers a second explanation on how biased attitudes are transmitted and how subsequent behaviors and socio-psychological characteristics are influenced. Social cognitive learning theory was first introduced by Albert Bandura (Bandura, 1977, 1986; Bandura & Walters, 1963; Bussey & Bandura, 1999). In social cognitive learning theory, three socio-cognitive modes of influence are distinguished that bolster the development of learning in general, and of biased attitudes specifically (Bussey & Bandura, 1999). Firstly, through

direct tuition, in which (biased) ideas and meaning are explicitly taught. Secondly, through modeling, in which the individual models their behavior after a (role) model. Like the constructivist-ecological perspective, social cognitive learning theory places emphasis on the idea that learning is informed by social interaction. According to this theory, individuals learn by observing and imitating the behavior of others (Bandura, 1977, 1986; Bandura & Walters, 1963; Bussey & Bandura, 1999). The third socio-cognitive mode of influence on the development of biased attitudes is enactive experience. This entails observing the social reinforcement and social sanctions that follow from specific behavior. When certain behavior is met with disapproval, this discourages the individual to repeat said behavior. Not just actual reinforcement and punishment promote and discourage certain behaviors, expected reinforcement and punishment are also encouraging and discouraging respectively (Zaikman & Marks, 2017). Positive reinforcement can also be given simply by the presence of certain things, and the absence of others. For example, infants show gendered preferences for toys, according to the presence of the types of toys in their home (Boe & Woods, 2018). In classrooms, social sanctions have been found to be gender-differentiated, with boys receiving reinforcement for academic success and sanctions for misbehavior, while girls receive praise for tidiness and compliance, and sanctions for academic failure (Bussey & Bandura, 1999). This leads to differences in self-efficacy, sense of agency, and competencies, as well constraints on career aspirations. These negative effects have been found to be stronger for ethnic minority girls.

Social role theory

In his work on the ecological model, Bronfenbrenner discussed social roles (1979). He defined social roles as the "activities and relations expected of a person occupying a particular position in society (p. 85)." For instance, the activity expected of a primary school teacher is to teach, and the expected relation is with primary school pupils. According to

Bronfenbrenner, social roles are differentiated not just by occupation, but also by various other characteristics, such as sex/gender, age, social status, ethnicity, and religion. By interacting with people in various social roles, human development is facilitated. In the following decade, social role theory was developed with the aim to understand sex/gender differences and similarities in social behavior (Eagly et al., 2000). Later, social role theory was also used to examine other social roles (Koenig & Eagly, 2014), making the connection with Bronfenbrenner's definition stronger. According to social role theory, prejudice and biased attitudes originate in the uneven representation of specific social groups in specific social roles (Eagly et al., 2000; Eagly & Wood, 1999; Koenig & Eagly, 2014; Morgenroth & Ryan, 2018). People enact certain behaviors within their social roles, and when specific social groups are overrepresented in certain roles, this affects the traits that are assumed to be characteristic of these groups (Koenig & Eagly, 2014). For instance, there is a widespread division of labor, with women doing more unpaid labor in the home, while men do more paid work outside of the home (Dutch Central Bureau of Statistics, 2022a; Petersen & Hyde, 2010). Because of this, women and girls are more involved in home-making roles, in which they learn skills like cooking and taking care of children (Eagly et al., 2000). Simultaneously, men and boys acquire more skills that are rewarded in paid occupations. The division of labor is an important facet of gender-based stereotypes, which positions men as breadwinners and women as homemakers. Other important facets of gender based bias include the idea that men are expected to be dominant, assertive, independent, and leaders, whereas women should be submissive, caring, kind, considerate, helpful, and nurturing (Eagly et al., 2000; Endendijk et al., 2020). Social roles related to gender can also concern specific traits that are important in school. For instance, girls are often positioned as well behaved and diligent, and boys as more disruptive (Glock & Klapproth, 2017; Myhill & Jones, 2006; Timmermans et al., 2016, 2018). Teachers are more tolerant to disruptive boys than they are to girls (Musto, 2019). Children are aware

of these assumed traits, and perceive that their teachers expect trait related behaviors from them (Myhill & Jones, 2006).

Social role theory also alludes to the idea that all social behavior is in a sense scripted and performed, explained further in social script theory (Wiederman, 2005, 2015; Zaikman & Marks, 2017). Shared meanings, beliefs, and values exist within social groups. Together, they construct an interpretation of reality, metaphorical social scripts, which are internalized and enacted by the individual members of the social group (Simon & Gagnon, 1984, 1986), not unlike actors who perform scripts on stage (Wiederman, 2005). Social scripts operate on three different levels. Firstly, there are cultural scripts. These are collective, broad, and general social norms, which are often conveyed by cultural institutions like mass media, politics, governments, laws, policies, education, and religion (Wiederman, 2015), comparable to the exo- and macrosystems in the ecological model by Bronfenbrenner. For instance, religious messages in the Bible, which are further conveyed by Christian political parties, prescribe that men and women have distinct gendered roles: a man should be the breadwinner and head of the household, while a woman should be a caring and nurturing mother and homemaker. The performance of virtually every social role has to adhere at least to some extent to the appropriate cultural scripts (Simon & Gagnon, 1984). Cultural scripts are often rather abstract, and it is quite possible that a cultural script is incongruent with a concrete situation (Simon & Gagnon, 1984). To resolve this incongruence, and to allow social actors to participate in social interactions, interpersonal scripts are used. This second level of scripts concerns the adaptation of cultural scripts within social interaction (Wiederman, 2015). The idea of adaptation is comparable to the constructivist idea of assimilation (Piaget, 1954). For example, a parent may be aware of certain cultural scripts, like "women are well suited to be home-makers", and adapt this script in child rearing by giving their daughter dolls and miniature household appliances to play with. In this context, the interpersonal script is how the child is provided with tools

that symbolize women in the role of homemaker. This can be further reinforced by social interactions with their parents, who adapt and adhere to the cultural scripts of gendered social roles. Lastly, the third level of social scripts are intrapsychic social scripts. These are social scripts at the most personal level, which include fantasies, mental rehearsals, desires, and memories (Wiederman, 2015). Intrapsychic social scripts exist on the most personal level; they connect social meanings to individual desires and ideas (Simon & Gagnon, 1984), similar to the constructivist idea of accommodation (Piaget, 1954). These intrapsychic scripts could include the make believe play, fantasies, and drawings of a girl, who imagines herself being a mother and a homemaker, her desire to fulfil those roles when she grows up, and her memories of her own mother in those roles (see Figure 1.2). Thus, the three levels of social scripts are dynamically interrelated and constantly co-act in behavior and social interaction.

People incorporate their social roles into their identity (Zaikman & Marks, 2017). Because of this, social roles have a considerable influence on human behavior (Eagly & Wood, 1999). Adherence to prescribed social scripts and roles is important, as people are judged on how well they conform to these scripts and roles (Zaikman & Marks, 2017).

Conformity leads to praise and positive evaluations, while violation leads to penalization and negative evaluations. Through these positive and negative consequences of the adherence or rejection of prescribed social roles, gender bias and stereotypes are reinforced and perpetuated. People who belong to a high-status group are inclined to be more rigid in guarding the boundaries of their group (Leaper, 2015). In many (patriarchal) societies, men are valued higher than women, and consequently gender roles are more rigid for men and boys than they are for women and girls (Leaper, 2015; Leaper & Brown, 2018). For girls, adhering to prescribed gender roles also means devaluing their own gender (Duveen, 1996). For example, stereotypical gender roles prescribe

Figure 1.2

Intrapsychic Script: Drawing of a Young Girl's Wish to Become a Mother

Note. This drawing represents an intrapsychic script: the wish and fantasy of a young girl to become a mother when she grows up. I made this drawing in 1994 (age 6 years) in response to the question "What do you want to be when you grow up?" The teacher wrote down my answer to the question at the bottom: "I want to become a mother and nothing else!" At the top of the image, I drew a bassinet, to represent a baby. I drew a man on the left, my future husband, wearing a tie as he works in an office. I drew myself with substantial breasts, something I saw as a necessity for being a mother.

that girls should prioritize being sexy and attractive to boys, which is perceived to be incompatible with being intelligent and performing well in school (A. A. Nelson & Brown, 2019). Therefore, gender roles are more flexible and non-conforming behavior is more acceptable for girls than for boys (Leaper & Brown, 2018).

While social role theory has mainly been used to investigate gender-based bias and prejudice (Zaikman & Marks, 2017), research has shown that the theory also applies in other contexts, like ethnicity, body type, and sexual orientation (Koenig & Eagly, 2014).

Socialization Agents

A concept binding social role theory, the constructivist-ecological perspective, and social cognitive learning theory are socialization agents (also: socializing agents, agents of socialization). In society, there are various socialization agents, people and entities that instruct, model, and through praise and penalties show which behaviors are appropriate and which are not (Bandura, 1986). Important socialization agents include the media, cultural and governmental institutions, schools, peers, and parents (Bussey & Bandura, 1999; C. Connell & Elliott, 2009; Endendijk et al., 2022; Zaikman & Marks, 2017). These agents are situated in the cultural and interpersonal social scripts of social role theory (Wiederman, 2015), and exist within the various systems in the ecological model of Bronfenbrenner (see Figure 1.1). In my dissertation, I will focus on three different socializing agents: parents, teachers, and media in Chapters 3, 4, and 5 respectively.

Parents

From the constructivist-ecological perspective, and both within social cognitive learning theory and social role theory, parents are seen as largely influential on their offspring. When parents raise their children, they teach them many types of essential cognitions and skills. Especially from infancy to early adolescence (0-12 years), parents are very important

agents of socialization. They play a large part in children's development through direct modeling, instruction, praise, and discouragement (Astle & Anders, 2022; Bussey & Bandura, 1999). During this process, parents have been found to shape the gender attitudes of their children (Kågesten et al., 2016). This often happens through implicit processes. From birth (and possibly prenatally as well) parents show differences in their behaviors towards boys and girls. Even when there are no physical differences between boys and girls in their size and their activity, parents have gendered perceptions of newborn girls being more fragile, softer, and finer featured than newborn boys (Bussey & Bandura, 1999). Parents vary and adjust their body movements, the way they touch, vocalize, and make facial expressions according to the gender of their infant (Fausto-Sterling, 2012). More explicitly, gender differentiation and bias can be transmitted through the physical appearance and environment of babies, which are adjusted according to their gender. Parents dress boys and girls differently, style their hair differently, decorate their rooms differently, and buy different toys for sons than for daughters (Brown & Stone, 2018; Bussey & Bandura, 1999). Parents have different rules for sons than for daughters (Axinn et al., 2011; Endendijk et al., 2022), and may directly formulate differential gendered expectations, or show their bias through emotional responses (Astle & Anders, 2022). In similar processes, parents have been shown to transfer other types of attitudes and bias to their children, like intergroup/ethnic bias (Bigler & Liben, 2006; Bigler & Patterson, 2017; Degner & Dalege, 2013; Levy & Hughes, 2009), and political attitudes (Torney-Purta, 2006). Because of their crucial role in the socialization of children, parents are central in developmental psychological research.

Teachers

Like parents, teachers socialize through (differential) instruction, modeling, praise, and punishment (Bandura, 1977; Bussey & Bandura, 1999). There are various socializing practices that are quite unique to the

position of the teacher. Teachers of course use instruction on a daily basis as a socio-cognitive mode of influence on the development of children. Other socializing practices that are mostly reserved to teachers include giving turns in a classroom setting, grading, writing evaluations, giving (extra) homework, and dividing children into groups. Teacher bias has been found in many of these practices (see for instance, Bašaragin & Savic, 2019; Bassi et al., 2016; Denessen et al., 2020; Frawley, 2005; Kågesten et al., 2016; Mason et al., 2014; Voyer & Voyer, 2014; Wang et al., 2018). Teachers shape biased attitudes of pupils. For instance, teachers reinforce stereotypical gender norms by pardoning the bullying of girls by boys, and regarding it as a form of heterosexual attraction (Kågesten et al., 2016). Teachers have also been found to actively exclude girls from sports that promote male-typed qualities like toughness, aggression, and competitiveness, while boys who do not conform to these qualities are stigmatized. Teachers' interethnic prejudice has been found to put refugee children at higher risk of peer violence, and to be negatively linked to inter-ethnic social contact (Alan et al., 2021).

The differential treatment of pupils fueled by teacher bias also has impact on various other pupil outcomes like achievement, self-efficacy, and motivation (Denessen et al., 2020; Gentrup et al., 2020; Jussim, 2009; Wang et al., 2018). The importance of investigating biased socialization by teachers therefore is important even when their pupils do not adopt and internalize the same bias as their teachers.

Media

In media portrayals, bias and stereotypes can often be found. For example, in newspaper articles, White men in politics are most often described by their political ideology, while for other politicians aspects of their identity are often emphasized, like their gender identity and ethnic background (Runderkamp et al., 2022). By stressing these identities, the media implicitly conveys the message that women and people belonging to ethnic minority groups defy the norm and are out of place in politics.

Other examples of bias in media include magazines aimed at adolescents that perpetuate gender stereotypes (Joshi et al., 2011), heteronormative television shows (Aubrey et al., 2020), sexist music videos (Endendijk et al., 2022), children's TV shows and advertisements that portray strongly gender stereotyped behaviors (Brown & Stone, 2018; Coyne et al., 2014, 2016; Walsh & Leaper, 2020), and the way in which Black people are more likely to be depicted as criminals, unemployed, or working in lower status jobs than White people (Signorielli, 2009). Through these types of bias in media, social roles are conveyed and modeled, which could affect the media consumer's own attitudes and behavior (Bussey & Bandura, 1999). For instance, in news media, crime is overrepresented, and specific groups like ethnic minorities, men, and juveniles are overrepresented as (potential) criminals, which can lead to negative attitudes towards these groups (Arendt & Northup, 2015; Grosholz & Kubrin, 2007; Popović, 2018a; Windzio & Kleimann, 2009). Specific forms of crimes are notably overrepresented, like child sexual abuse (Grosholz & Kubrin, 2007; Hove et al., 2013). Especially cases that are particularly abhorrent, for instance because there is a large number of victims or because the victims are exceptionally young, result in spikes of media attention, with emphasis on the perpetrator, who is almost always male (Cromer & Goldsmith, 2010a; Hove et al., 2013; Popović, 2018b; Weatherred, 2015, 2017).

A specific case of how the media possibly negatively affects gendered attitudes is the case of male early childcare professionals. Men in early childcare are quite rare. For instance, according to the most recent data, in the Netherlands only 12% of early childcare professionals are male (Dutch Central Bureau of Statistics, 2022b). Early childcare professions are seen as traditionally feminine roles, particularly due to the physical contact that is involved in working with very young children (Eidevald et al., 2018). This is reflected in the media portrayal of girls and women as nurturing and caring, while men and boys are far less likely to be portrayed as such (Steyer, 2014). These gender roles and the status of child care professions as feminine are in itself reason for many men not to work

in early childcare (Rolfe, 2006). Additionally, men working in early childcare can be exposed to the risk of being seen as a (potential) pedophile or abuser, as them working with and touching children can be seen as unnatural (Eidevald et al., 2018). Men working as early childcare professionals can lead to parental concern and suspicion, exactly because they are somewhat of a novelty (Rentzou, 2011). Some studies have additionally theorized that media attention to child sexual abuse may lead to negative attitudes towards male (early) childcare professionals (Eidevald et al., 2018; Fahmy & Ibrahim, 2022; Munk et al., 2013; Rentzou, 2011). In Chapter 5, I will present a study on this topic.

While many studies have focused on the biased content of mass media, less attention has been paid to the possible impact of this content (Mutz & Goldman, 2010). There is consensus that media affect attitudes, but views differ to what extent and in what way (Easteal et al., 2015). Some studies have shown longitudinal effects of media consumption on attitudes and social-emotional behaviors (Arendt & Northup, 2015; Coyne et al., 2014, 2016; Schiappa et al., 2006; Slater, 2007). For instance, boys who watch superhero programs that contain strong masculine gender stereotypes, show more male gender stereotyped play (Coyne et al., 2014). Likewise, children who watch princess programs that contain strong feminine gender stereotypes, show more female gender-stereotypical behavior a year later (Coyne et al., 2016). However, a few years later, this effect is no longer present (Coyne et al., 2021). Similarly, other studies show no proof for longitudinal effects (Breuer et al., 2015; Shi et al., 2018), or longitudinal effects for negative bias towards some social groups but not towards others (Tukachinsky et al., 2017). Because of these conflicting results, studying the (longitudinal) effects of media as a socializing agent of bias remains important. The effect of media attention to child sexual abuse on attitudes towards male childcare professionals has not been investigated yet. As the media attention on child sexual abuse is capricious, with stark peaks (Weatherred, 2017), this may be especially interesting.

Subtle bias: does it really matter?

When my paternal grandfather was born in 1918, women did not yet have the right to vote, nor did the inhabitants of the Dutch overseas territories. When my paternal grandparents married in 1942, my grandmother lost her legal capacity (*juridische handelingsbekwaamheid*), meaning that she could not sign a contract, open a bank account, or make decisions independently on how she wanted to rear her children. Both my paternal and maternal grandparents belonged to the working class, which meant that they only received a primary school education. Any further education was beyond their means. It should be mentioned that unlike my maternal grandmother, my maternal granduncles did receive further education. My greatgrandparents decided that their small funds were better spent on the education of their sons than on the education of their daughter. Over the past century, many of these types of social inequalities have shrunk considerably (Liben, 2016). However, biases rooted in mechanisms such as sexism and classism persist and are still expressed, albeit in different and often more subtle ways (Dovidio et al., 2016; Jones et al., 2016; Leaper & Brown, 2018). For example, bias can be expressed by differential treatment through eye contact, smiling, tone of voice (İnan-Kaya & Rubie-Davies, 2021), exclusionary discipline, and name mispronunciations (Childs & Wooten, 2023). These types of subtle forms of discrimination are also known as microaggressions. Microaggressions are subtle, low-intensity incidents, that express discrimination or hostility towards marginalized or underrepresented groups, like ethnic minority groups and members of the LGBTIQ+ community (Byers et al., 2020; Ogunyemi et al., 2020). The various types of subtle expressions of bias are often enacted unconsciously or unintentionally (Jones et al., 2016). Expressions in spoken and written language, including linguistic abstraction and word use, constitute one field in which subtle forms of bias are studied (Beukeboom, 2014; Biernat et al., 2012; Menegatti & Rubini, 2017; Rojek et al., 2019).

Attention to biased and discriminatory language regularly meets mocking or even hostile reactions (Murray, 2018). In contemporary media and online discourse many people are of the opinion that tackling issues of potentially biased or offending language is unimportant or even unworthy of attention. People who do bring such issues to attention are seen as too sensitive and too easily offended (Murray, 2018). Another common response, for instance voiced by politicians, is that there are bigger, more important problems to focus on (Lowrie, 2018; Van der Aa, 2023). Likewise, acknowledging bias in language, opting for more inclusive language, and investigating more subtle forms of biased expressions has been met with intense discussions as well as resistance in academia (see for instance Funnekotter, 2021; Hofhuis & Pas, 2021; Natri et al., 2023; Özdil, 2014; Singer et al., 2023; Veldhuis & Weeda, 2022).

This begs the question: do these types of subtle bias matter? Most people oppose overt forms of discrimination, which can be defined as differential and unfair treatment of members of specific social groups, with observable negative outcomes (Van Laer & Janssens, 2011). Examples of overt discrimination include denying a person a promotion, position, education or housing because of their social characteristics, like their gender identity or ethnicity. Overt forms of discrimination have become quite rare, are largely considered unacceptable, and are unlawful in many cases (Dovidio et al., 2016; Jones et al., 2016; Van Laer & Janssens, 2011). Explicit attitudes towards groups that have traditionally been marginalized have in general become much more positive (Dovidio et al., 2016). In polling research, many people indicate that emancipation has come far enough (Kanne & van der Schelde, 2022). So is studying these subtle types of bias and discrimination a worthwhile endeavor, or should we be satisfied with how far we have come and focus on other topics?

There are several arguments to be made in favor of the continued study of bias. Firstly, research shows that, perhaps counterintuitively, subtle forms of discrimination have comparable adverse effects to overt discrimination (Jones et al., 2016). These adverse effects include negative

psychological health, physical health, and work-related outcomes. Thus researching subtle forms of discrimination is important for the sake of its victims. Secondly, due to its subtle and often implicit and unintentional nature, subtle bias is harder to recognize and to change (Dovidio et al., 2016). By investigating and documenting subtle forms of bias, we can contribute to awareness of contemporary biases, especially when they evolve into phenomena that are more elusive. Investigating contemporary biases therefore demands a critical lens when choosing research instruments, and perhaps demands improved or new methodologies.

Practical significance

In research into the effects of gender and ethnicity on (subtle) socialization practices and outcomes, it is not uncommon to find mainly or only small effects (see for instance Gentrup et al., 2020; Menegatti et al., 2017; Mertens et al., 2019; Miller, 2012). Like the question whether subtle forms of bias are important, this may lead to the question whether small effects are relevant. To answer this question, it is first necessary to understand what constitutes a small effect.

In quantitative research, statistical analysis is employed to investigate whether certain characteristics, conditions, or treatments have an effect on a dependent variable. For a long time, there has been a heavy emphasis on significance testing in quantitative research in the social and behavioral sciences (American Psychological Association, 2020; Cohen, 1994; Hyde, 2001; Keppel, 1991). This entails the calculation of the probability that any effect in a study was found by chance or coincidence. This probability is represented by the *p*-value, and usually it is agreed that when the *p*-value is lower than .05, it can be ruled out that the effect found was based on chance (Kirk, 1996). The *p*-value represents the statistical significance of a test and is affected by sample size. When we investigate a hypothesized effect among a larger sample, the chance of finding the effect is higher. However, the *p*-value only tells us what the probability is

of an effect being present, it does not inform us about the strength of an effect (Fritz et al., 2012). To measure the strength of an effect, several effect sizes exist. Over the past two decades, scientific attention to the importance of reporting effect sizes had increased (Alhija & Levy, 2009; Fritz et al., 2012; Peng et al., 2013; Sun et al., 2006; Zhou & Skidmore, 2017). Contrary to statistical significance, effect sizes are not dependent on sample size (Levine & Hullett, 2002). Effect sizes help us understand the impact of the treatment, condition, or characteristic that we investigate. Because of this, effect sizes have been called measures of *practical* significance (Ellis, 2010; Kirk, 1996). It is possible to find a difference between two groups, for instance men and women, which is statistically but not practically significant. For example, imagine we examine one million people who make a standardized language test, and men make significantly more mistakes than women do. However, the difference in mistakes is so small that this does not result in a difference between the grades men and women receive. While the difference is significant statistically, it is not a meaningful difference. Vice versa, it is possible that we find a quite large effect size when investigating a small number of people, but the effect is not statistically significant. This indicates that a bigger sample size is needed to reach statistical significance (Fritz et al., 2012).

This may lead to thinking that the larger the effect size, the better, but this is not quite true. Small effects are actually rather common in social and behavioral research, as the investigated issues are often subtle (Cohen, 1988). Large effects are often already well-known, so further verification and investigation is often not scientifically interesting and could be a waste of resources (Keppel & Wickens, 2004). Conversely, small effects can be much more interesting than large ones for several reasons. Firstly, a small effect can still have great real-life consequences (Maxwell et al., 2018), especially when many people are affected by it. For instance, a small negative effect on ethnic bias among police officers in a city the size of New York, could result in a decrease of thousands of Black people

being ethnically profiled (Greenwald et al., 2015). Secondly, small effects may accumulate, leading to larger effects (Greenwald et al., 2015; Jones et al., 2016; Prentice & Miller, 1992). For instance, research into the effects of (biased) teacher expectations on the learning opportunities and achievements of, and interactions with pupils reveals only small effects, but the accumulative effects have widened the achievement gaps between specific groups of pupils (Rubie-Davies, 2015). Thirdly, small effects may be important in providing a decision between competing theoretical explanations (Keppel, 1991), like whether gender differentiation can be explained by evolutionary or biosocial theory (Endendijk et al., 2020; Zaikman & Marks, 2017). Lastly, when an effect is tested under inopportune or tough circumstances and a small effect is found, it shows that an effect is pervasive (Prentice & Miller, 1992). However, it should be noted that it is important that the right effect size is chosen and that the calculation of effect sizes is done correctly. Research has shown that this often is not the case (Peng et al., 2013; Zhou & Skidmore, 2017).

Dutch context

It is important to discuss the larger cultural/national context in which the studies in my dissertation took place. As explained by Bronfenbrenner's ecological model (1979; see Figure 1.1), socialization is embedded in various social structures. The way in which bias plays a role differs between national societies and cultural groups. Below, I give a brief description of how gender and ethnicity play a role in Dutch society.

Gender

On various international indices for gender equality, the Netherlands rank high (Dutch Central Bureau of Statistics, 2022a). Compared to most other countries, women often do paid work. Dutch men and women attain a comparable high level of education, and have relatively equal political power (although women still form less than 40% of parliament; The Economist, 2022). In other aspects, the Dutch do less well. While political

power is shared relatively equally, the Netherlands has never had a female prime minister. The gender pay gap in the Netherlands is still relatively large and men take up way more leadership positions than women (World Economic Forum, 2021). Contributing to the fact that the Netherlands has the highest rate of women working part-time worldwide, is that, compared to other OECD countries, maternity leave is short in the Netherlands, and child-care costs are high (The Economist, 2022). Additionally, are women more often in occupations in which working fulltime is difficult due to irregular shifts and scheduling issues, like home care and childcare (Daub, 2022; UWV, 2020). This division of labor is striking. While being overrepresented in occupations involving (health) care and children aged 0-12 years, Dutch women are greatly underrepresented in occupations related to science, technology, engineering, and mathematics (STEM; Dutch Central Bureau of Statistics, 2022a). This division is reflected in school choices of Dutch students.

Although Dutch men and women share household tasks and care responsibilities more equally compared to people in other countries (Dutch Central Bureau of Statistics, 2022a), women spend more hours each week on these tasks than men (Kanne & van der Schelde, 2022). Even when men indicate that they find it important to share care for their children equally, they generally spend over 10 hours per week less on care for their children than women. While Dutch men and women indicate that they find gender equality important and see themselves as emancipated, they simultaneously endorse traditional gender roles, indicate that they do not want to emancipate any further, and especially men have negative attitudes towards feminism (Kanne & van der Schelde, 2022).

Ethnicity

In the Netherlands, there are high levels of ethnic and racial segregation, both geographically and educationally (Boterman et al., 2019; Hondius, 2009). Racist and Islamophobic practices can been found in many facets of Dutch society, like classrooms (Weiner, 2015, 2016), newspapers

(Runderkamp et al., 2022), television broadcasts, and in political campaigns (Welten & Abbas, 2022). There are various social movements that (for example) make efforts to bring attention to the colonial history and present of the Netherlands, address inequalities in the educational system and on the labor market, as well as protests against ethnically biased police brutality and the use of Black caricatures in Dutch festivities (Rose, 2022). There are also various examples of recent institutional racism in the Netherlands. After controlling for several demographic characteristics, parents from ethnic minority groups are at much higher risk to have their children placed in out-of-home care by court order than White Dutch parents (Dutch Central Bureau of Statistics, 2022c). In the so-called *Toeslagenaffaire*, parents were unlawfully targeted based on their ethnicity by tax authorities with false allegations of fraud with childcare benefits (Van der Bunt & Verheul, 2022). In a report published in late 2022, it was revealed that there is pervasive institutional racism at the Dutch ministry of Foreign Affairs (NOS, 2022). For example, employees of the ministry reported that the n-word is used to address Black people, as well as words like 'apes'. African countries have been described as 'monkey-countries' by employees of departments aimed at foreign aid. Specific cultural backgrounds were connected to being lazy and criminal. Employees indicated that supervisors rarely intervened in these types of situations.

While there are many examples of how ethnic bias plays a serious role in Dutch society, and Dutch people indicate that they find ethnic equality important (Kanne & van der Schelde, 2022), there is a large tendency to refuse to acknowledge race and racial discrimination (Wekker, 2016). The Dutch often position themselves not just as very tolerant, but also as 'colorblind', indicating that race and skin color are so unimportant that they do not even see it (Hondius, 2009; Özdil, 2014; Rose, 2022; Weiner, 2016). This is for instance reflected in the belief that the Dutch school system is meritocratic and that ethnicity is unimportant in the attainment of success (Stevens et al., 2019; Weiner, 2015). This attitude of

White Dutch people towards ethnic bias appears to follow the same pattern as the attitude towards gender equality: they find it important, but many feel like the final destination has been reached and improvement is no longer required. Research into the remaining structures of gender- and ethnicity-based inequalities therefore remains very important. Academia can play a pivotal role into uncovering bias that society refuses to see.

Aim and outline of the dissertation

The general aim of my dissertation is to provide insight into subtle forms of bias in socialization by various socializing agents, and exploring novel ways to document bias. In **Chapter 2,** we present a tutorial on understanding and calculating effect sizes in the context of the analysis of variance. As my dissertation focuses on a research area that is characterized by small and subtle effects, calculating correct and unbiased effect sizes is important. The tutorial is subsequently used to calculate effect sizes in Chapters 4 and 5. In **Chapter 3,** parents of 15- and 17-year old adolescents are the socializing agents of interest. We examine the responses of parents to adolescents who engage in multiple romantic relationships and kissing a stranger. Men are often praised while women are penalized when engaging in the same sexual behavior. This differential way in which men and women are treated for sexual behavior is called the sexual double standard. However, what gendered messages parents convey when talking about sexual behavior has rarely been studied. We use content analysis to investigate whether the sexual double standard is present in parents' responses. In **Chapter 4,** we study teachers as socializing agents. We analyze what teachers write on the school reports of primary school pupils in grade 6-8. We focus on three aspects: 1) the proportion of positive and negative remarks, 2) whether the remarks evaluate school work or other pupil aspects like behavior and character traits, and 3) the linguistic categories of the remarks. We examine whether differences exist between boys and girls, and between

pupils from the ethnic majority and pupils from ethnic minority groups. In **Chapter 5,** we investigate the media and their role as socializing agents. We examine the attention to a big child sexual abuse story with a male perpetrator in the media, and the influence of this attention on parents' attitude towards hiring a male babysitter over the course of four years. We investigate whether parents of 1- to 7-year old children become more negative towards hiring a male babysitter after prolonged attention to a male perpetrator of child sexual abuse, and whether this differs between fathers and mothers. Finally, in **Chapter 6,** a general discussion of this dissertation is presented, describing the main findings, limitations, implications, and future directions of the research.

Chapter 2

Demystifying Omega Squared

Practical Guidance for Effect Size in Common ANOVA Designs

Antoinette D. A. Kroes, Jason R. Finley

Advanced online publication in *Psychological Methods*, 2023, https://doi.org/10.1037/met0000581

Abstract

Omega squared ($\widehat{\omega}^2$) is a measure of effect size for ANOVA designs. It is less biased than eta squared, but reported less often. This is in part due to lack of clear guidance on how to calculate it. In this paper, we discuss the logic behind effect size measures, the problem with eta squared, the history of omega squared, and why it has been underused. We then provide a user-friendly guide to omega squared and partial omega squared for ANOVA designs with fixed factors, including one-way, two-way, and three-way designs, using within-subjects factors and/or between-subjects factors. We show how to calculate omega squared using output from SPSS. We provide information on the calculation of confidence intervals. We examine the problems of non-additivity, and intrinsic versus extrinsic factors. We argue that statistical package developers could play an important role in making the calculation of omega squared easier. Finally, we recommend that researchers report the formulas used in calculating effect sizes, include confidence intervals if possible, and include ANOVA tables in the supplemental materials of their work.

Keywords: Omega squared; partial omega squared; effect size; ANOVA

Comparison of means is a common analysis in psychological science. For example, researchers often want to compare performance (dependent variable) across different groups of people or different conditions of a treatment (independent variable). ANOVA (analysis of variance) is the inferential statistical test for comparing means across three or more groups/conditions, and/or for comparing means across two or more independent variables. The *F* test statistic and *p*-value from an ANOVA indicate the statistical significance of the test—that is, the probability of obtaining these differences in means just due to chance. However, it is crucially important for researchers to also report a measure of *effect size*, which tells us not just whether there was likely a difference in means at all, but how *large* that difference was, or how *strongly* an independent variable effected the dependent variable (American Psychological Association, 2020; Keppel & Wickens, 2004; Thompson, 1999a; L. Wilkinson, 1999).

Effect sizes are also important for meta-analysis, in which treatment effects are compared across studies (Thompson, 1999a), and for calculating sample size required to obtain certain levels of power (Keppel & Wickens, 2004).

A commonly used measure of effect size for ANOVA is eta squared (η^2) or partial eta squared (η_p^2). Eta squared estimates the amount of variance in the dependent variable that is accounted for by one or more independent variables. However, eta squared is problematic because it is biased: it tends to overestimate the true effect size in a population. Many authors have pointed out this flaw and have advised alternatives (e.g., Albers & Lakens, 2018; Field, 2017; Lakens, 2015; Okada, 2013; Olejnik & Algina, 2000; Tabachnick & Fidell, 2007a; Yigit & Mendes, 2018). The foremost alternative measures of effect size are omega squared ($\widehat{\omega}^2$) and partial omega squared ($\widehat{\omega}_p^2$), which were first proposed over 50 years ago (Hays, 1963). Omega squared is much less biased than eta squared, and thus is a superior measure. However, it is still rarely used (Alhija & Levy,

2009; Zhou & Skidmore, 2017). The goal of this paper is to help remedy that by providing user-friendly explanations and instructions.[1]

In this paper, we will explain the logic of eta and omega squared, and the shortcomings of eta squared. We will give a brief history of omega squared, and we will discuss the reasons we believe it is still underused. Most helpfully, we will provide formulas to calculate omega squared for the most commonly used ANOVA designs in the behavioral and social sciences, up to three-way ANOVAs. We will explain how to use these formulas with output from SPSS. We will conclude with recommendations on how to calculate and report omega squared.

Clarifying Our Scope and Terminology

Let us first clarify some terminology, and define the scope of what we will and will not cover.

Variables (aka factors): are characteristics that vary across entities. In experiments, an *independent variable* is one that is manipulated by the researcher (i.e., they determine the possible values and assign participants to those values), and a *dependent variable* is one that is simply measured by the researcher. Common inferential statistical tests such as *t*-test and ANOVA are used to see if an independent variable(s) has an effect on the dependent variable. ANOVA may also be used for non-experimental data, in which case the variables may be referred to as *predictor variable* and *outcome variable*, rather than independent variable and dependent variable, respectively. For convenience, we will simply use the terms independent variable (IV) and dependent variable (DV). We will

[1] Seasoned researchers may note how painstakingly we describe and explain basic concepts in this paper. We have deliberately chosen to do this, as we found that most of the previous articles on omega squared assume a good deal of specific prior knowledge, and are difficult to decipher for non-statisticians. That may be one of the obstacles to more widespread use of omega squared. Another obstacle is that a number of the sources we have scoured are out-of-print textbooks. We wish to remove such obstacles.

also use the term *treatment* as a synonym for IV. As for labeling, the outcome or dependent variable is typically labeled as Y. The predictor or independent variable may often be labeled as X, though we will use the label A instead of X in order to be consistent with many of our sources. The words variable and factor are synonyms, and the latter is often used to refer to IVs in ANOVA; we will use the words variable and factor interchangeably in this paper.

ANOVA (analysis of variance): A statistical test (aka model) of the relationship between one or more categorical independent variables and one continuous dependent variable. A categorical variable has values that are treated as named categories without inherent order or numeric value, and can have two or more such categories; a continuous variable has numeric values and may be of interval or ratio scale of measurement.[2] The purpose of ANOVA is to compare means across three or more groups/conditions. The variances of the groups/conditions are simply a tool toward that end, and are not actually themselves the subject of analysis. An ANOVA produces an ANOVA summary table, which contains the outcome(s) of the test(s) for statistical significance, as well as the components needed for calculating effect size. The terms one-way, two-way, and three-way refer to the number of independent variables in the design.

Statistical significance*:* When conducting statistical analyses, we choose a significance level, represented by α (alpha). This value represents the probability of rejecting the null hypothesis while it is in fact true. Traditionally, this value is set at .05 or .01. When carrying out statistical analyses we calculate a *p*-value. This value is the probability of getting the obtained result (or a more extreme value) while the null hypothesis is true. When the *p*-value is equal to or lower than the

[2] It is also possible to use ANOVA with a Likert-scale DV (Norman, 2010) or a dichotomous DV (Lunney, 1970).

significance level α, we reject the null hypothesis. When the *p*-value exceeds α, it cannot be ruled out that the found effect is due to chance (Kirk, 1996). Reporting the *p*-value is a longstanding and important tradition in the social and behavioral sciences (American Psychological Association, 2020).

Effect size: is a measure that estimates the strength of the investigated effects of the IV(s). Whereas statistical significance only indicates whether an effect is present, effect sizes describe the quantitative size of the effect (Fritz et al., 2012). Effect sizes help us understand the expected impact of a treatment or condition. Thus, an effect size gives an indication whether an effect is meaningful in the real world, and is therefore called *practical significance* (Ellis, 2010; Kirk, 1996). While statistical significance is dependent on sample size, effect sizes are not; they should be comparable across studies, regardless of sample size (Levine & Hullett, 2002). A large effect that is not significant indicates that greater power may be needed, while a very small significant effect cautions against overvaluing the effect (Fritz et al., 2012).

Between-subjects, within-subjects, and split-plot designs: These distinctions have to do with how an independent variable(s) is manipulated. For an IV that is manipulated *between-subjects* (aka independent samples), each participant experiences only one value of the IV, and thus the comparison of DV means is made between different groups of participants. For an IV that is manipulated *within-subjects* (aka repeated measures), each participant experiences all values of the IV, and thus the comparison of DV means is made within that one group of participants such that each participant's score in one condition is compared to their scores in the other conditions. By convention, the values of a between-subjects IV are often called *groups*, and the values of a within-subjects IV are often called *conditions*. The term *levels* may also be used in either design. In this paper, we will use the term groups/conditions to refer generally to values of an IV. Note also that the terms *participant* and *subject* are synonyms.

A one-way ANOVA has only one IV and will be either between-subjects or within-subjects. For two-way ANOVAs and higher, a *split-plot* design is one in which there is at least one between-subjects IV and at least one within-subjects IV. The name originates from agricultural research, where experiments were conducted on different plots and subplots of land (Goos, 2010). Some authors have referred to this model as a mixed design, as it is a mix of within-subjects and between-subjects factors (Gaebelein & Soderquist, 1978; Keppel & Wickens, 2004). However, this can be quite confusing, as the term "mixed" has different meanings in other contexts, such as a "mixed effects model" (aka mixed model, linear mixed model, multilevel model, ANOVA Model III) which is one that includes both fixed factors and random factors, or "mixed methods research" (aka mixed research) which combines qualitative and quantitative methods. Thus, in this paper we use the term split-plot to avoid confusion. Note that the distinction of between-subjects versus within-subjects is unrelated to the distinction of fixed effects versus random effects, which we will discuss next.[3]

Fixed effects versus random effects (aka fixed factors vs. random factors): This distinction has to do with how the conditions of an independent variable are chosen. A fixed effect is when the researcher chooses a fixed set of conditions for an IV. A random effect is when the researcher randomly samples conditions from a range of possible values, so that the conditions used may vary across experiments.[4]

The distinction is important for several reasons. First, the conclusions from an experiment using fixed effects should be limited to just those conditions that were used, whereas the conclusions based on random effects can be broader. Second, the expected mean squares, and thus the

[3] The encyclopedia entry Mixed Model Design (Kraska, 2010) confuses these two uses of the word "mixed."

[4] More explanation of fixed versus random effects can be found in the following textbooks: Keppel and Wickens (2004, pp. 533-549), and Myers, Well, and Lorch (2010, p. 335).

appropriate error terms, differ for fixed versus random effects. This is why there are three different overall ANOVA models. ANOVA Model I is for designs including only fixed effects, ANOVA Model II is for designs including only random effects, and ANOVA Model III is for designs including both fixed and random effects. The way that omega squared is calculated differs across these three models. The most common scenario in psychology research is an experiment using only fixed effects (i.e., ANOVA Model I). Thus, we will limit the scope of this paper to only fixed effect IVs (aka fixed factors). If you want to know about calculating omega squared for designs that include random effects consult Dodd and Schultz (1973), Olejnik and Algina (2000), and Vaughan and Corballis (1969). Alternatively, the intraclass correlation coefficient (ICC, or $\hat{\rho}^2$) has been recommended for designs that include random effects (Kirk, 2012; Maxwell et al., 1981).

Summary of scope: In this paper, we will address between-subjects, within-subjects, and split-plot designs, up to three-way ANOVAs, for fixed factors only.

Logic of ANOVA Effect Size Measures

In order to understand the logic of effect size measures for ANOVA, including omega squared, we must first consider the larger context. When we measure a group of people's performance on some task, their scores will vary from each other. The job of psychological science is to understand *why* that variance happens, to "account for" the variance.[5]

Let us use Y to represent a variable we have measured, for example performance on a memory test for a list of words. In the context of an experiment, we call this the dependent variable. If we have no other information about the Y scores—that is, we know nothing about any other

[5] For a further discussion of what "accounting for variance" means, see Sechrest and Yeaton (1982).

variables—then all we can do is *describe* the variance of those scores. We cannot explain any of it. Every possible conceivable other variable in the world could be influencing the spread of those Y scores.

But suppose that we do have more information about those Y scores. Suppose we know the value of another variable, let us call it A, that goes along with each Y score. In fact, the reason we know the A values in this example is because we randomly assigned people to three different conditions of A (e.g., short, medium, and long amount of time to study the list) before we measured Y. That is, we ran a between-subjects experiment and A was the independent variable.

An inferential statistical test, in this case a one-way between-subjects ANOVA, would tell us the extent to which we might want to believe that there is truly any effect at all of variable A on variable Y. If the ANOVA tells us that $p < .05$, the effect of A on Y is statistically significant. But the statistical significance does not tell us how *big* the effect of A on Y is (i.e., the *effect size*). For that we must see *how much of the variance* in Y may be attributable to A. That is, how much of a role did study time (A) play in the variance of peoples' memory test performance (Y)?

To do so, we must *partition the total variance into two different components*. The first component is the variance in Y attributable to A, which might be called between-groups variance, or treatment effect. The second component is all the remaining leftover variance in Y, which could be due to any other variables in the world (aka extraneous variables). This component is often called error variance. The word error does not mean mistake. It means the unexplained deviation of a Y score from what we would have predicted it to be, based on the overall mean of Y and the mean of the relevant A condition. Error variance is simply variance in Y that is due to variables that are not included in our statistical model.

We quantify variance components by calculating several different *sums of squared deviations*. SS_{total} is the sum of squared deviations of all Y scores from the grand mean (i.e., the overall mean of all Y scores,

disregarding any other variable). This represents the total variability of all the Y scores:

$$SS_{total} = \sum_j \sum_i \left(Y_{ij} - M\right)^2 \quad (1)$$

In the above formula, Σ means summation, *j* indicates a particular group/condition of the independent variable (A), *i* indicates a particular participant, Y_{ij} is the dependent variable score of participant *i* in group/condition *j*, and *M* is the grand mean.

SS_{total} can be partitioned into two components: $SS_{between}$ and SS_{within}. $SS_{between}$ (also called SS_{effect} or $SS_{treatment}$) is the between-groups variance component, which tells us how much of the variance in Y is due to A. It is quantified by calculating how much the group means (e.g., mean test performance for the short, medium, and long study time groups) vary around the grand mean:

$$SS_{between} = \sum_j n_j \left(M_j - M\right)^2 \quad (2)$$

In the above formula, n_j is the number of participants in group/condition *j*, and *Mj* is the mean of group/condition *j*.

SS_{within} (also called SS_{error}) is the leftover variability, which is our best estimate of the influence of all other conceivable variables on Y. It is quantified by the sum of squared deviations of individual Y scores from their respective group means:

$$SS_{within} = \sum_j \sum_i \left(Y_{ij} - M_j\right)^2 \quad (3)$$

This formula is very similar to the SS_{total} formula, except that we are comparing each participant's score to their respective group/condition mean, instead of the grand mean. Adding $SS_{between}$ and SS_{within} gives us SS_{total}.

That is, SS_{total} consists of two components: SS_{between} and SS_{within}. Keep in mind, this is for a one-way between-subjects ANOVA.

Now that we have partitioned the variance, we can see *how much of the total variance* is due to a particular component. That is, effect size. Let us start by using the eta squared measure of effect size. Eta squared gives an intuitive use of the partitioned variance:

$$\eta^2 = \frac{SS_{\text{between}}}{SS_{\text{total}}} \tag{4}$$

Eta squared expresses the between-groups variance as a proportion of the total variance, telling us how much of the variance in dependent variable Y can be attributed to independent variable A. For example, how big of a role did study time play in performance on our memory test? Eta squared can be thought of conceptually as either the *proportion of variance accounted for*, or as a *proportional reduction in error/uncertainty*. As a proportion, its possible values range from 0 to 1. So, for example, a value of .50 would mean that independent variable A accounts for 50% of the variance in dependent variable Y. This makes sense, and eta squared works well as a descriptive statistic of our sample data. However, there are problems with eta squared.

What is Wrong With Eta Squared?

In reporting on ANOVA, eta squared and partial eta squared are the most popular effect sizes (Peng et al., 2013; Zhou & Skidmore, 2017). It has become common practice in statistics to use Greek letters to indicate a population parameter (American Psychological Association, 2020; Keppel & Wickens, 2004). This might give the impression that eta squared (η^2) is an estimator for the effect size in the population. This however, is not the case. Eta squared is simply a descriptive statistic of the sample data

(Lakens, 2013; Maxwell et al., 2018; Tabachnick & Fidell, 2007a).[6] If we want to make inferences to the broader population of people (i.e., everyone not in our sample), eta squared is flawed. This is why eta squared is also known as R^2 or the correlation ratio in the context of regression (Keppel & Wickens, 2004).

To understand the problem with eta squared, let us remind ourselves about the difference between population and sample. In psychology research, the population is the hypothetical set of all possible participants of interest (e.g., all humans, past, present, and future), whereas the sample is a single finite set of participants drawn from that population. Descriptive statistics merely describe our sample data. Inferential statistics draw conclusions about the population from the sample.

Some sample statistics, such as the mean, are unbiased estimators of their corresponding population parameters, meaning that the expected valued of the sample mean in the long run is equal to the population mean. That is, if we were to endlessly draw new samples from the population, calculate the sample mean each time, and build a sampling distribution out of those means, the mean of that sampling distribution would be equal to the population mean.

However, some sample statistics are biased estimators, such as the variance. The expected value of the sample variance is smaller than the actual population variance. An adjustment is necessary to make an unbiased estimate of the population variance. That adjustment is Bessel's correction, which uses n-1 in the denominator of the variance formula, instead of n. This adjusted version of the sample variance is often denoted as s^2.

Eta squared is a ratio of sums of squares calculated from the sample data (SS_{between} / SS_{total}). Just as the unadjusted sample variance is a biased

[6] The persistent misleading use of the Greek letter eta is probably a result of the notation used by SPSS.

estimator of the population variance, an unadjusted measure based on sample sums of squares will also be a biased estimator of the population value. In this case, eta squared tends to *overestimate* the true effect size, especially when it is calculated from small samples. The reason for this flaw is that the numerator, $SS_{between}$, consists of variance due to the factor, *as well as* some random variance in the means of the groups of that factor. This is due to sampling error. Even when the population group means do not differ, the sample group means will always differ somewhat. This is due to the random chance involved in sampling participants from the population (Ellis, 2010; Maxwell et al., 2018; Myers & Well, 2003). In eta squared, these coincidental differences between sample groups are treated as systematic (Keppel & Wickens, 2004). Treating coincidental differences as systematic results in positive bias, meaning it will overestimate the effect, especially with small sample sizes (Albers & Lakens, 2018; Keppel & Wickens, 2004; Lakens, 2015) and even more so for partial eta squared (Levine & Hullett, 2002). Simultaneously, as the explained variance due to the factor is overestimated (numerator), the unexplained variance is underestimated (denominator, Maxwell et al., 2018). Multiple studies using Monte Carlo simulations have shown the extent of the positive bias that occurs in eta squared and partial eta squared (Keselman, 1975; Okada, 2013; Yigit & Mendes, 2018). Recently, Liu (2022) proposed a bootstrapping method to correct for bias in eta squared. This method requires some proficiency with the program R. Importantly, this method is only suitable for one-way between-subjects ANOVAs and not for other designs. The positive bias in eta squared and partial eta squared is problematic as it can lead to overvaluing effects, as well as underpowering subsequent studies. In a 2015 blogpost, statistician Daniël Lakens wrote: "If η^2 was a flight from New York to Amsterdam, you would end up in Berlin." Although that may be a bit of an overstatement (Albers, 2015), the bias in eta squared should motivate researchers to choose a better measure of effect size, namely omega squared.

Development/History of Omega Squared

The origins of the effect size omega squared are somewhat murky. We can broadly think of most effect size measures as belonging to one of two families: standardized differences (such as Cohen's *d*), and associative strength (such as Pearson's *r*; Ellis, 2010, pp. 6-15). The ANOVA effect sizes belong to the latter. Using Google Books Ngram Viewer and the PsychINFO database, we conclude that the standard version of omega squared was introduced by Hays (1963) in the first edition of his well-regarded graduate-level statistics textbook (pp. 323-332, and especially pp. 381-384). Several other sources also point to Hays (1963) as the origin (Dwyer, 1974; Keren & Lewis, 1979; Sechrest & Yeaton, 1982).[7]

However, there were several precursors to omega squared that have been mentioned in histories by other authors (Dwyer, 1974; Glass & Hakstian, 1969; Huberty, 2002; Keren & Lewis, 1979). Examining these precursors can help readers in understanding effect sizes in general and omega squared specifically. It is unclear whether Hays knew of these precursors, as he did not cite anything when introducing omega squared, and simply justified his use of the omega symbol by referring to it as a "relatively neutral symbol," presumably meaning one that had not already been used very much for other measures.

The general idea of quantifying the strength of association dates back to at least Pearson (1905)[8] who defined and labeled eta as the correlation ratio. Eta squared was introduced for regression by Pearson (1911), and for ANOVA by Fisher (1925, 1928). As stated before, the squared correlation ratio, η^2 ($SS_{\text{between}}/SS_{\text{total}}$), is synonymous with R^2 (aka $r^2 = 1 - \frac{SS_{\text{residual}}}{SS_{\text{total}}}$), which is often called the coefficient of determination. The term

[7] There are other statistics that are referred to as omega: Cramer-von-Mises omega squared, an alternative to the Kolmogorov-Smirnov test; Cohen's Omega, an effect size for Chi Square tests; and McDonald's coefficient omega, an alternative to Cronbach's alpha measure of reliability. These are unrelated to the omega squared measure of effect size for ANOVA.

[8] For a more thorough history of effect sizes, see Huberty (2002).

R^2 is used in the context of regression, whereas η^2 is used in the context of ANOVA. To the best of our knowledge, omega squared is not used in regression analysis.

Several measures descended from eta squared. Bolles and Messick (1958) proposed a utility index or coefficient of utility, *U*, which appears to be equivalent to eta squared (Gaito & Firth, 1973). *U* has not apparently had a lasting impact, but was notable for its early emphasis on the importance of an effect (i.e., its utility). Another relative of eta squared was the intraclass correlation coefficient (rho-i, ρ_i), introduced by Fisher (1925) for use with random effects; it is analogous to omega squared which is used more with fixed effects. For fixed effects, Kelley (1935) recognized the bias in eta squared, which tends to overestimate the population effect size, and developed epsilon squared (ε^2) as an improvement. Epsilon squared is equivalent to adjusted R^2 in regression contexts (Vogt & Johnson, 2015, p. 142), and is very similar to omega squared, differing only in the denominator. Finally, Hays (1963) developed omega squared, also intended mainly for fixed effects ANOVA. In the years since then, there have been refinements of and debates about omega squared, including epsilon versus omega (Glass & Hakstian, 1969), slightly unequal sample sizes (Vaughan & Corballis, 1969), formulas for designs that include random effects (Dodd & Schultz, 1973), formulas for MANOVA and ANCOVA (Olejnik & Algina, 2000), the introduction of partial omega squared (Keren & Lewis, 1979), and the introduction of generalized omega squared (Olejnik & Algina, 2003).

It is worth considering the conceptual formulas that would be used to calculate omega squared if we somehow magically knew population values. A first helpful stepping stone in understanding the formulas for the effect sizes is provided by Keppel & Wickens (2004; p. 162), who express the idea of effect sizes in words:

$$\text{effect size} = \frac{\text{variability explained}}{\text{total variability}} = \frac{\text{total variability} - \text{unexplained variability}}{\text{total variability}} \quad (5)$$

Effect size, in this context, represents the proportion of variability that is accounted for by an effect (aka variable, factor, treatment). Variance in the population can be represented by sigma squared (σ^2). This results in the population formulas for omega squared shown in Table 2.1, which are all equivalent to each other.

Table 2.1

Population formulas for omega squared.

Formula	Source
$\omega^2 = \frac{\sigma^2_Y - \sigma^2_{Y\|X}}{\sigma^2_Y}$	Hays (1963, pp. 325, 381-382)
$\omega^2 = \frac{\sigma^2_Y - \sigma^2_e}{\sigma^2_Y}$	Maxwell et al. (1981, pp. 526-527)
$\omega^2 = \frac{\sigma^2_A}{\sigma^2_A+\sigma^2_e}$ $\omega^2 = \frac{\sigma^2_\alpha}{\sigma^2_\alpha+\sigma^2_\epsilon}$ $\omega^2 = \frac{\sigma^2_A}{\sigma^2_A+\sigma^2_{S/A}}$	Myers & Well (2003, p. 208), Vaughan & Corballis (1969, p. 206), Keppel (1991, p. 64)
$\omega^2 = \frac{\sigma^2_\alpha}{\sigma^2_Y}$	Cardinal and Aitken (2005)

The overall logic is the same across all these forms of the population formula. Starting with the first row (formula 10.19.2 in Hays, 1963), σ^2_Y is the total population variance of variable Y, and $\sigma^2_{Y|X}$ is the population variance of Y within a particular group/condition of variable X. That is, $\sigma^2_{Y|X}$ is the remaining variance left in Y (DV) given that you know X (IV). Assuming equal variance of Y across all groups/conditions of X (i.e., homogeneity of variance, aka homoscedasticity, which is an assumption underlying ANOVA), $\sigma^2_{Y|X}$ is simply the error variance, σ^2_e, which is the leftover variance due to unkown variables. Thus, $\sigma^2_Y - \sigma^2_{Y|X}$ gives us the variance in Y attributable to variable X. Dividing that by σ^2_Y gives us the proportion of Y's variance that is attributable to X, or the proportional

reduction in uncertainty. This last interpretation is because we can consider variance to be a kind of uncertainty (aka error). Say we want to predict one person's score on a memory test. Our best guess is the mean score across all people; but the person's actual score might be some distance from that mean. Exactly how far? We are uncertain. But the smaller the variance around that mean, the closer we can get to making a good guess, and thus the less uncertain we are. The next formula, from Maxwell et al., 1981 (pp. 526-527) simply replaces $\sigma^2_{Y|X}$ with σ^2_e.

The three formulas in the third row switch to using A or α as the label for the IV, instead of X. These formulas use the single term σ^2_A or σ^2_α to represent the variance in Y that is attributable to X, then divide by a sum that yields the total variance of Y. Finally, the simplest population formula is provided by Cardinal and Aitken (2005): the variance in Y due to X, divided by the total variance in Y.[9]

In reality, we cannot truly know the population variances, because for the purpose of inferential statistics in psychology, we typically conceive of the population as an infinitely large hypothetical distribution of all possible individuals of interest. Thus, the best we can do is to estimate those variances from sample data. This returns us to the bias in eta squared, and how omega squared improves upon that. Omega squared corrects for bias by both shrinking the numerator and enlarging the denominator.[10]

Hays (1963) made these corrections by using a combination of ANOVA expected mean squares (E(*MS*)) and some "nasty" algebra (Keppel & Wickens, 2004, p. 163). The E(*MS*) are based on the idea of repeatedly

---

[9] Some authors have argued that the very idea of total population variance of Y is nonsensical. See Maxwell et al. (1981) for discussion.

[10] Another very similar measure of effect size, epsilon squared, just shrinks the numerator from eta squared, and does not change the denominator. For detailed comparison of epsilon squared versus omega squared, see Glass and Hakstian (1969), and Carroll and Nordholm (1975).

drawing samples from the population and calculating $MS_{between}$ and MS_{within} for each sample, resulting in two sampling distributions of these two *MS* values (Myers & Well, 2003, p. 203). The means of these sampling distributions are the E(*MS*) and they play an important role in ANOVA. Myers and Well (2003) present the following formula for ω^2 (p. 208; see also our Table 2.1):

$$\omega^2 = \frac{\sigma_A^2}{\sigma_A^2+\sigma_{error}^2} \tag{6}$$

The numerator (σ_A^2) represents the variance of the treatment effect, while the denominator ($\sigma_A^2 + \sigma_{error}^2$) represents total population variance. Each of these components can be replaced with formulas for the E(*MS*). By using the E(*MS*) formulas, the formula for ω^2 can be rewritten as the estimate of the population parameter. In the general formula for omega squared we replace "between" with "effect", and add a hat (circumflex) on top of the omega, to indicate that it is an estimator. Myers and Well (2003, p. 209) rewrite the formula as follows:

$$\widehat{\omega}^2 = \frac{[(a-1)/a](1/n)(MS_{effect}-MS_{within})}{[(a-1)/a](1/n)(MS_{effect}-MS_{within})+MS_{within}} \tag{7}$$

where *a* is the number of groups/conditions in Factor A. Using algebra, this formula can be rewritten more simply, using components from the ANOVA table:

$$\widehat{\omega}^2 = \frac{SS_{between}-(df_{effect}\times MS_{within})}{SS_{total}+MS_{within}} \tag{8}$$

Let us first consider the denominator, which contains the total variance (SS_{total}), as well as the mean square of the error variance (MS_{within}). By adding MS_{within}, the denominator is enlarged compared to the formula for η^2 (Equation 4). It may be counterintuitive that a measure for error variance is added, while SS_{total} already contains SS_{within}. This

can be explained by the formulas for E(*MS*), as they follow the argument that the *F*-ratio ($MS_{\text{between}}/MS_{\text{within}}$) equals 1 when the null hypothesis is true (Keppel & Wickens, 2004, p. 36). The idea is that, when the null hypothesis is true, MS_{between} and MS_{within} are both estimates of σ^2_{error}, the error variance (Myers & Well, 2003, p. 203). Therefore, it is argued that we should "choose an error term such that its E(*MS*) and the E(*MS*) of the term to be tested are identical when the null hypothesis is true" (p. 204). The E(*MS*) formula for σ^2_{A} in the context of a one-way between-subjects ANOVA is:

$$\sigma^2_{\text{A}} = n\theta^2_{\text{A}} + \sigma^2_{\text{error}} \tag{9}$$

This means that σ^2_{error} is part of the E(*MS*) formulas for σ^2_{A} as well as added (again) to the total population variance (Glass & Hakstian, 1969, p. 406). MS_{within} is considered to be a good estimator for σ^2_{error}.[11]

Now let us consider the numerator. Since the numerator of eta squared (SS_{between}) includes variance due to the independent variable A as well as variance due to error, we need to subtract out the error variance amount. The exact amount to be subtracted is worked out using the E(*MS*) and degrees of freedom of MS_{between} and MS_{within}. MS_{within} represents the amount that each group mean in the sample is expected to vary from its respective mean in the population. The number of times we subtract MS_{within} is the degrees of freedom of our IV: the number of independent observations (i. e., number of conditions in the IV) minus one for our estimation of the overall population mean from sample data.

The formula for standard omega squared for one-way between-subjects designs can also be rewritten as a function of the values of *F*,

[11] For more in-depth information on E(*MS*) and their use in ANOVA and $\hat{\omega}^2$, we recommend: Carroll & Nordholm, 1975; Dodd & Schultz, 1973; Glass & Hakstian, 1969; Hays, 1963; Maxwell et al., 1981; Myers & Well, 2003; Vaughan & Corballis, 1969.

df_{effect}, and N (Carroll & Nordholm, 1975; Keppel & Wickens, 2004; Maxwell et al., 2018):

$$\widehat{\omega}^2 = \frac{df_{\text{effect}}(F_{\text{effect}}-1)}{df_{\text{effect}}(F_{\text{effect}}-1)+N} \tag{10}$$

Conceptually, this formula may be the most clear in how $\widehat{\omega}^2$ corrects for the sampling error which is present in η^2. As stated before, when there is no treatment effect and the null hypothesis holds true, the F-value will approximate 1, as MS_{between} and MS_{within} in that case are both estimates of σ^2_{error} (Keppel & Wickens, 2004, p.164). By subtracting 1 from F, the sampling error is corrected. Because of this correction, $\widehat{\omega}^2$ is always smaller than η^2.

To sum up, effect size for ANOVA consists of the proportion of total variance in a DV that is attributable to an IV. Eta squared calculates this very simply from sample data. Omega squared corrects for the bias in eta squared by adjusting both the numerator and the denominator, providing a better estimate of the effect size in the broader population. Both eta squared and omega squared use values that are readily available in an ANOVA summary table.

Underuse of Omega Squared

Bias in eta squared was described over 80 years ago (T. L. Kelley, 1935) and the less-biased alternative omega squared was proposed 60 years ago (Hays, 1963). Still, omega squared is rarely used. There are several explanations for this. Traditionally, there has been a heavy emphasis on significance testing, while behavioral scientists were educated far less on effect sizes and power analysis (APA, 2020; Cohen, 1994; Hyde, 2001; Keppel, 1991). The emphasis on the importance of effect size is relatively new. An explicit recommendation to utilize effect sizes was added to the fourth publication manual of the American Psychological Association (APA) in 1994, but this did not result in increased use of effect sizes in the

following years (Hyde, 2001; Thompson, 1999a, 1999b; L. Wilkinson, 1999). This was perhaps due to the large number of editions of the publication manual. The following editions have paid increasingly more attention to effect sizes. The fifth edition of the publication manual mentioned that "it is almost always necessary to include some index of effect size or strength of relationship in your Results section" (APA, 2001, p. 26). In the sixth and seventh editions, effect sizes have been added as a requirement for all publications, and guidance was added on what types of effect size should be reported (APA, 2010, 2020). The reporting of effect sizes has increased considerably since the late nineties; research over the past two decades shows that approximately half of the reported ANOVA tests are accompanied by any measure of effect size (Alhija & Levy, 2009; Fritz et al. , 2012; Peng et al., 2013; Sun et al., 2006; Zhou & Skidmore, 2017).

The shortcomings of (partial) eta squared have not been discussed in the publication manual (APA, 2020). Despite the overall increase in reporting of effect sizes, eta squared and partial eta squared have continued to dominate, as reviewed by Peng et al., 2013 (see also: Alhija & Levy, 2009; Barry et al., 2016; Fritz et al., 2012; Kirk, 1996). For example, Fritz et al. (2012) examined articles published in the *Journal of Experimental Psychology: General* in 2009 and 2010, and out of all the articles that reported ANOVA results, they found only one use of omega squared, compared to 32 uses of eta squared (either standard or partial). Although not as precise as the manual counting done for review articles like those cited here, we conducted a search of PsycINFO for publications in 2019 or 2020, and we found 454 results that contained any of the following search terms: η^2, $\eta\hat{}2$, η_p^2, or $\eta p\hat{}2$. By contrast only 18 results contained any of the following search terms: ω^2, $\omega\hat{}2$, ω_p^2, or $\omega p\hat{}2$. Thus, the underuse of omega squared continues.

The positive bias in eta squared and partial eta squared may actually be one of the reasons why they are preferred over other effect size measures (Fritz et al., 2012). Other reasons for the high prevalence of eta squared and partial eta squared are likely familiarity and convenience.

Partial eta squared can be automatically produced by the most used statistical packages, like SPSS. This easy accessibility promotes its use, even when it is not appropriate (Kirk, 1996; Zhou & Skidmore, 2017). Levine and Hullett (2002) further documented problems with effect sizes reported from SPSS. Omega squared was not included at all in SPSS until the 27th version in 2020 (Mathew, 2020), and then only for the one-way between-subjects ANOVA design.

The formulas for omega squared and partial omega squared differ across designs, which makes them more cumbersome to calculate than eta squared and partial eta squared. Finding the right formulas, and guidance on how to make calculations, is difficult. On the one hand, much information is published only in statistical textbooks. This content is often not indexed in commonly used internet search engines. On the other hand, and perhaps most importantly, there is no definite consensus on what formulas to use to calculate omega squared (Maxwell et al., 2018). This holds especially true for multifactor ANOVAs with one or more within-subjects factors. While some authors argue that it is impossible to calculate omega squared in these models (Keppel & Wickens, 2004), others recommend it (Gaebelein & Soderquist, 1978). Moreover, most publications only supply formulas for one- and two-way ANOVA designs, due to the large number of possible formulas (for an exception, see Dodd & Schultz, 1973; who also present formulas for three-way designs). Lastly, different authors often use different symbols and subscripts to describe the same components; while in other instances one term is used to describe opposing constructs. This can be very confusing, especially for inexperienced researchers. To aid in demystifying omega squared, we have included a disambiguation table where we clarify the different notations used in a selection of papers (see Appendix A), and we have done the legwork of gathering and verifying the formulas to use for the most common ANOVA designs.

Formulas for Omega Squared

In this section we will present formulas for omega squared and partial omega squared for ANOVA designs with fixed factors. For designs with random factors, we recommend readers to consult Dodd and Schultz (1973), Olejnik and Algina (2000), and Vaughan and Corballis (1969). A quick overview of the formulas can be found in Appendix B. For each design, we will show an example of output from SPSS (version 28), highlighting where each component of the relevant formula can be found. We recommend calculating the formulas using widely available software such as Microsoft Excel, and we provide examples of this in the supplemental materials (see Appendix D).

Between-Subjects Designs

One-Way Between-Subjects Designs

Conceptually, the effect size in the population is estimated with the following formula (see also Table 2.1):

$$\omega^2 = \frac{\sigma^2_{\text{A}}}{\sigma^2_{\text{A}}+\sigma^2_{\text{error}}} \qquad \text{(6, repeated)}$$

For any size between-subjects design with fixed factors, the formula for standard omega squared can be expressed in terms found in the ANOVA table (Dodd & Schultz, 1973):

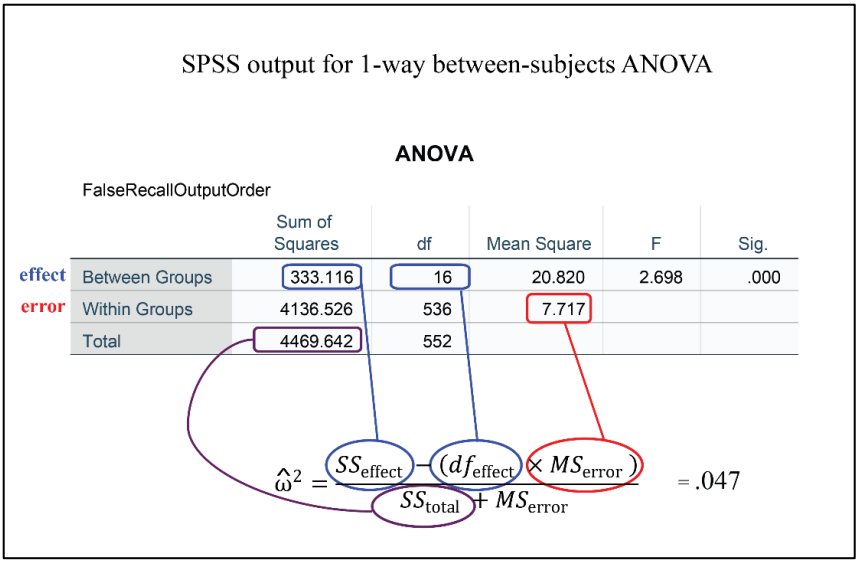
$$\widehat{\omega}^2 = \frac{SS_{\text{effect}}-(df_{\text{effect}}\times MS_{\text{error}})}{SS_{\text{total}}+MS_{\text{error}}} \qquad (11)$$

In the one-way between-subjects ANOVAs, the subscripts *between* and *within* are often used (see Formula 8). In other designs, these subscripts are replaced with *effect* and *error* respectively. For consistency, we will use *effect* and *error* in the formulas for all designs. See Figure 2.1 for an example of SPSS output from a one-way between-subjects ANOVA, showing the components used for omega squared. In this example,

$\widehat{\omega}^2$ = .047. By comparison, η^2 = .075. Notice that $\widehat{\omega}^2 < \eta^2$, as expected. Note also that we report omega squared to three decimal places, and never include a leading zero.

Figure 2.1

The Components for Standard Omega Squared in SPSS Output For a One-Way Between-Subjects Design.



Note. Data are from Finley et al. (2017), available in the supplemental materials. SPSS version 28.

Multi-Factor Between-Subjects Designs and Partial Omega Squared

For multi-factor between-subjects designs (aka higher order designs; two-way, three-way, etc.) it can be argued that it is inappropriate to use the standard omega squared formula (Keppel & Wickens, 2004). This is because the estimated total variance ($\widehat{\sigma}^2_{total}$, see Table 2.1) varies across designs. So, in a one-way design, the formula for the effect size of Factor A would be:

$$\widehat{\omega}^2 = \frac{\widehat{\sigma}_A^2}{\widehat{\sigma}_A^2 + \widehat{\sigma}_{error}^2} \qquad \text{(6, estimated version)}$$

And in a two-way design, where Factor B is added, the formula would be:

$$\widehat{\omega}^2 = \frac{\widehat{\sigma}_A^2}{\widehat{\sigma}_A^2 + \widehat{\sigma}_B^2 + \widehat{\sigma}_{AB}^2 + \widehat{\sigma}_{error}^2} \qquad (12)$$

This means that even when the estimated variance components for Factor A and the random error are identical in both designs, this will result in a different value for the estimated effect size. Therefore, *partial omega squared* is proposed, which always consists of the same components as the one-way design, regardless of the total number of factors in the design (Keppel, 1991). The effects of other factors are *partialled out* (Keren & Lewis, 1979) and can be written as:

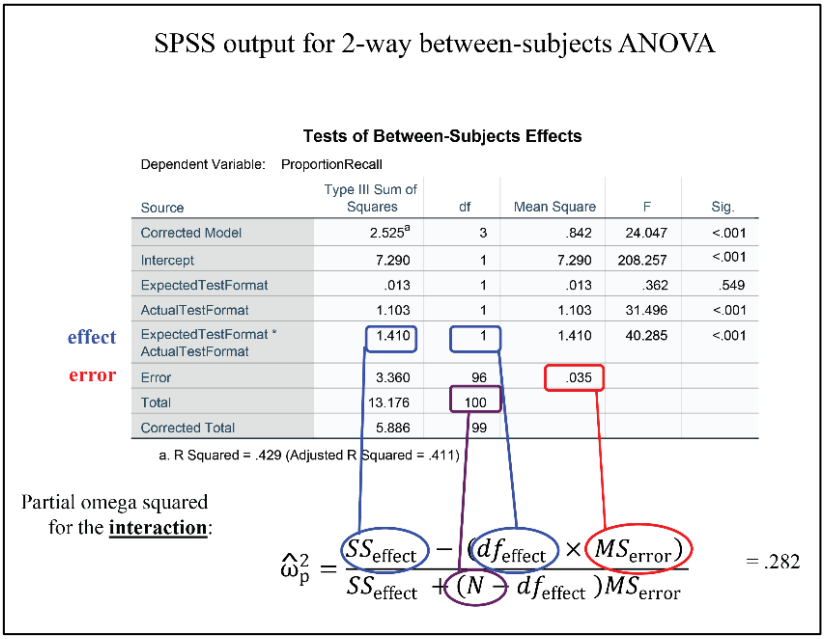
$$\widehat{\omega}_p^2 = \frac{SS_{effect} - (df_{effect} \times MS_{error})}{SS_{effect} + (N - df_{effect})MS_{error}} \qquad (13)$$

Standard omega squared represents the proportion of the total variance explained by the effect of one factor. Partial omega squared represents the proportion of variance explained *exclusively* by one factor that is *not* explained by other factors in the model. As the denominator for partial omega squared is smaller, this will always result in a higher value. The standard effect size is useful to compare the effect of the various factors within the design. This is impossible with the partial effect size, as they do not share the same denominator (Sechrest & Yeaton, 1982). The partial effect size is suitable for power analysis (Keppel & Wickens, 2004) and usually more suited to make comparisons across designs (Fritz et al., 2012; Keppel & Wickens, 2004; Levine & Hullett, 2002). We will go into the matter of partialling out factors and comparability further in the section on obstacles in using omega squared.

See Figure 2.2 for an example of SPSS output from a two-way between-subjects ANOVA, showing the components used for partial omega squared.

Figure 2.2

The Components for Partial Omega Squared in SPSS Output for a Two-Way Between-Subjects Design.



Note. Data are from Finley and Benjamin (2012), available in the supplemental materials. SPSS version 28.

The formulas for omega squared and partial omega squared in between-subject designs can be rewritten to be computed from the *F*-test statistic, *df*, and *N* (Maxwell et al., 2018). This is especially straightforward for partial omega squared (which is the same as the formula for standard omega squared in one-way designs, see Formula 10).

$$\widehat{\omega}^2 = \frac{df_{\text{effect}}(F_{\text{effect}}-1)}{\sum_{\text{all effects}}(df_{\text{effect}}F_{\text{effect}})+df_{\text{error}}+1} \quad (14)$$

$$\widehat{\omega}_{\text{p}}^2 = \frac{df_{\text{effect}}(F_{\text{effect}}-1)}{df_{\text{effect}}(F_{\text{effect}}-1)+N} \quad (15)$$

As the sample size and levels of the factors should be reported in any research paper, this means that omega squared and partial omega squared in between-subjects designs can often be calculated even in absence of an ANOVA table. For standard omega squared this requires the *F*-values and *df* for all factors in the design.

Within-Subjects Designs

About Within-Subjects Factors

When all subjects receive all levels of a treatment, the treatment is a within-subjects factor (Keppel & Wickens, 2004, p. 347). As the subjects are measured more than once, they are a potential source of variance (Tabachnick & Fidell, 2007a, p. 249). A one-way within-subjects design can therefore be thought of as a two-factor design, often indicated as an $\overline{\text{A}}\times\text{S}$ design. Factor $\overline{\text{A}}$ is the fixed within-subjects factor (a line is added above the letter to signify a within-subjects factor) and the subjects form the second, random between-subjects Factor S (Dodd & Schultz, 1973; Keppel & Wickens, 2004; Olejnik & Algina, 2000). The presence of Factor S complicates the conceptualization of standard and partial omega squared in ANOVAs that include within-subjects designs. This has led some authors to claim that it is impossible (Keppel, 1991; Keppel & Wickens, 2004) or at least problematic (Gaebelein & Soderquist, 1978; Vaughan & Corballis, 1969) to calculate omega squared for within-subjects designs. We will describe these complications and subsequently argue that omega squared can indeed be calculated.

When we insert a one-way within-subjects design into a table, this table holds one observation in each cell (Keppel & Wickens, 2004; Olejnik & Algina, 2000; Tabachnick & Fidell, 2007a), as shown in Table 2.2.

Table 2.2

One-way Within-Subjects ($\bar{A}\times$ S) ANOVA Design With a = 3 Conditions and N = 3 Subjects.

	Conditions of Factor A		
Subjects	a1	a2	a3
s1	Y_{11}	Y_{12}	Y_{13}
s2	Y_{21}	Y_{22}	Y_{23}
s3	Y_{31}	Y_{32}	Y_{33}

In a one-way within-subjects design, there are two identifiable sources of variance: variance due to treatment $\overline{A}$ ($\hat{\sigma}^2_{\text{effect}}$) and variance due to the systematic differences between subjects ($\hat{\sigma}^2_{\text{subject}}$). The remaining variance in the model is indicated with $\hat{\sigma}^2_{\text{effect}\times\text{subject}}$. Ideally, $\hat{\sigma}^2_{\text{effect}\times\text{subject}}$ is made up solely of random variance (error). This is the strength of within-subjects designs compared to between-subjects designs (Keppel & Wickens, 2004; Lakens, 2013; Loftus & Masson, 1994). In a between-subjects design, no distinction can be made between random error and systematic differences between subjects, and the total variance is therefore simply defined as:

$$\hat{\sigma}^2_{\text{total}} = \hat{\sigma}^2_{\text{effect}} + \hat{\sigma}^2_{\text{error}} \tag{16}$$

While in a within-subjects design, $\hat{\sigma}^2_{\text{error}}$ is replaced with $\hat{\sigma}^2_{\text{subjects}} + \hat{\sigma}^2_{\text{effect}\times\text{subject}}$. There are two ways to define the total variance for within-subjects designs (Keppel, 1991):

$$\hat{\sigma}^2_{\text{total}} = \hat{\sigma}^2_{\text{effect}} + \hat{\sigma}^2_{\text{subjects}} + \hat{\sigma}^2_{\text{effect}\times\text{subject}} \tag{17}$$

And

$$\hat{\sigma}^2_{\text{total}} = \hat{\sigma}^2_{\text{effect}} + \hat{\sigma}^2_{\text{effect} \times \text{subject}} \quad (18)$$

Excluding the variance due to subjects from the total variance leads to higher effect sizes and power, which are important reasons for authors to adopt a within-subjects design (Keppel, 1991; Keppel & Wickens, 2004; Lakens, 2013). This can be seen as partialling out the random Factor S, similar to how one would calculate partial effect sizes in a multi-factor between-subjects design. However, there are three important arguments *not* to partial out the systematic differences between subjects. Firstly, because of the viewpoint that effect sizes are meant to provide a standard metric that can be used across designs (Lakens, 2013; Maxwell et al., 2018). As the systematic differences between subjects are not partialled out in between-subjects designs, they should not be partialled out in within-subjects designs either, for comparability purposes. The higher effect sizes found in within-subjects designs when the systematic differences between subjects are partialled out, are therefore often seen as overestimations of the actual effect size (Lakens, 2013; Maxwell et al., 2018; Olejnik & Algina, 2003). Secondly, systematic differences between individuals are seen as part of the total population variance (Maxwell et al., 2018). It appears illogical to disregard them as such in within-subjects designs. The third argument relates to the assumption of additivity, which we will discuss below. We will then present formulas in which the variance due to subjects is not partialled out. In the section on obstacles for using omega squared we will discuss alternatives in which the subject variance can be partialled out.

Non-Additivity

The third argument to not partial out the variance due to subjects from the total variance relates to (non-)additivity. ANOVAs for designs that

include one or more within-subjects factors have to meet the assumption of additivity, which is a component of the assumption of sphericity (Tabachnik & Fidell, 2007). The assumption of sphericity holds true when the variances across conditions, as well as the covariances between pairs of conditions, are equal (Field, 2017, p. 654). The F test is not robust against violation of the assumption of sphericity, and violations increase the chance of Type I error (Maxwell et al., 2018). Departure from sphericity can be measured. It is denoted with a lower case epsilon (ε) and when the assumption of sphericity is met perfectly, it has a value of 1 (Maxwell et al. 2018). There are three well-known estimators for epsilon: Greenhouse-Geisser, Huynh-Feldt, and the lower-bound estimate of sphericity. In SPSS output for an ANOVA that includes within-subjects factors, those estimators are listed in a table titled "Mauchly's Test of Sphericity". When the assumption of sphericity is false, the estimators will be smaller than 1 and a correction should be made to the degrees of freedom. Mauchly's test itself is problematic, especially for small samples (Field, 2017, p. 655; Maxwell et al., 2018, p. 629), but the general idea is that if the test is statistically significant, then corrected *df* should be used.[12] The corrected *df* are listed as separate rows in the "Tests of Within-Subjects Effects" table of the SPSS output. While using the corrected degrees of freedom increases the *p*-value, it does not impact omega squared. This is because adjustments are made to all components used to calculate omega squared, so while the absolute values are changed, the ratio between the numerator and the denominator remains the same. The bottom line is that for calculating omega squared, it will not matter whether or not you use *df* corrected for violation of sphericity.

Additivity and sphericity are regularly discussed as if they are interchangeable, but in fact additivity is a component of sphericity. For a

[12] For guidelines on how to assess sphericity and which estimator to choose, see Maxwell et al. (2018, pp. 627-634) and Tabachnick and Fidell (2007, pp. 284-288).

more thorough explanation of the relation between sphericity and additivity see Tabachnick and Fidell (2007, pp. 247-248, pp. 284-288). As a general term, additivity signifies that there is no interaction between factors. Take for instance the effect of fertilizer and picking the bugs off our fictional rosebushes. We have tried these measures separately and know our roses grow 3cm a week because of fertilizer and 2cm because we remove the bugs. Now we decide to treat our bushes to both of the treatments at the same time. If the model is additive, the roses grow 5cm in total. But if they suddenly grow 7cm the model is non-additive, the bug-picking and fertilizer *interact* and contribute to an additional 2cm of growth.

In within-subjects designs, a distinction can be made between additive and non-additive models (Dodd & Schultz, 1973; Keppel, 1991; Tabachnick & Fidell, 2007a; Vaughan & Corballis, 1969; Winer, 1962). This does not refer to the possible interaction of the factors we are investigating, but to the possible interaction of the random subjects factor S with the IVs. Recall that in the within-subjects design we have one variance component that simultaneously indicates random variance (error) and the interaction between the treatment and the subjects: $\widehat{\sigma}^2_{\text{effect} \times \text{subject}}$ (Tabachnick & Fidell, 2007a). Ideally, we would have an additive design, where there is no interaction between the treatment and the subjects. When there is no interaction, $\widehat{\sigma}^2_{\text{effect} \times \text{subject}}$ represents only random error and this gives us the possibility to partial out the systematic differences between subjects ($\widehat{\sigma}^2_{\text{subject}}$). However, in any design that includes at least one within-subjects factor with more than two levels it is unreasonable to assume that the model is additive (Keppel, 1991; Tabachnick & Fidell, 2007a). In the social and behavioral sciences it is practically impossible to imagine a repeated measure that will not interact with the individual (Tabachnick & Fidell, 2007a, p. 248). It is often impossible to differentiate which part of $\widehat{\sigma}^2_{\text{effect} \times \text{subject}}$ is due to random error and which is due to the interaction between the subjects and the IV (Dodd & Schultz, 1973; Gaebelein & Soderquist, 1978; Keppel & Wickens, 2004; Vaughan &

Corballis, 1969). A test for non-additivity was devised by Tukey (Tukey, 1949; Winer, 1962) and is included in statistical packages like SPSS (Myers & Well, 2003) and SAS (Zambarano, 1992). However, this test is only suitable to assess one specific type of non-additivity, while other kinds exist (Dodd & Schultz, 1973; Zambarano, 1992). Zambarano (1992) provides some guidance on other methods to assess additivity, but this requires advanced statistical knowledge and still does not lead to a conclusive answer whether or not a model is additive. In practice, additivity is not often tested for; tests for sphericity and homogeneity of variance are seen as sufficient (Tabachnick & Fidell, 2007a). When a model is completely additive, this automatically means that the assumption of sphericity is met. However, a model can be non-additive while still meeting the assumption of sphericity through compound symmetry, and without inflating the *F* test. Thus, meeting the assumption of sphericity does not ensure additivity. A model where non-additivity may cause problems can be improved by adding a between-subjects "blocking" factor (Tabachnick & Fidell, 2007a, p. 285). For instance, if it is expected that native speakers recall more words in a memory test than non-native speakers, mother tongue can be added as a between-subjects factor. The interaction of mother tongue with the factor is then removed from the error term. Still, other sources of non-additivity could remain. It is also possible to transform the data to an additive scale (Myers & Well, 2003), although it can be challenging to find guidance on how to exactly transform the data. Some authors conclude that it is impossible to calculate omega squared for within-subjects designs, as it is impossible to estimate each source of error variance independently (Gaebelein & Soderquist, 1978; Keppel & Wickens, 2004).

In summary, non-additivity is a complicated matter. In practice, tests for the assumption homogeneity of variance and sphericity are seen as sufficient when carrying out ANOVAs (Tabachnick & Fidell, 2007a), but this does not solve the conundrum of choosing which formula to use to calculate the effect size. We argue that using the formulas presented in

this paper circumvents the problem by not partialling out variance caused by systematic differences between subjects. At the end of this section on designs that include within-subjects factors, we will discuss alternatives proposed by other authors. As there are many other options, we recommend to always report the formula used when reporting omega squared, especially when within-subjects factors are involved, and to include the full ANOVA table.

One-Way Within-Subjects Designs

For within-subjects designs the formula for standard omega squared is similar to the between-subjects design. The differences lie in the use of the two variance components (Dodd & Schultz, 1973; Keppel & Wickens, 2004; Maxwell et al., 2018):

$$\widehat{\omega}^2 = \frac{SS_{\text{effect}} - (df_{\text{effect}} \times MS_{\text{effect}\times\text{subject}})}{SS_{\text{total}} + MS_{\text{subject}}} \tag{19}$$

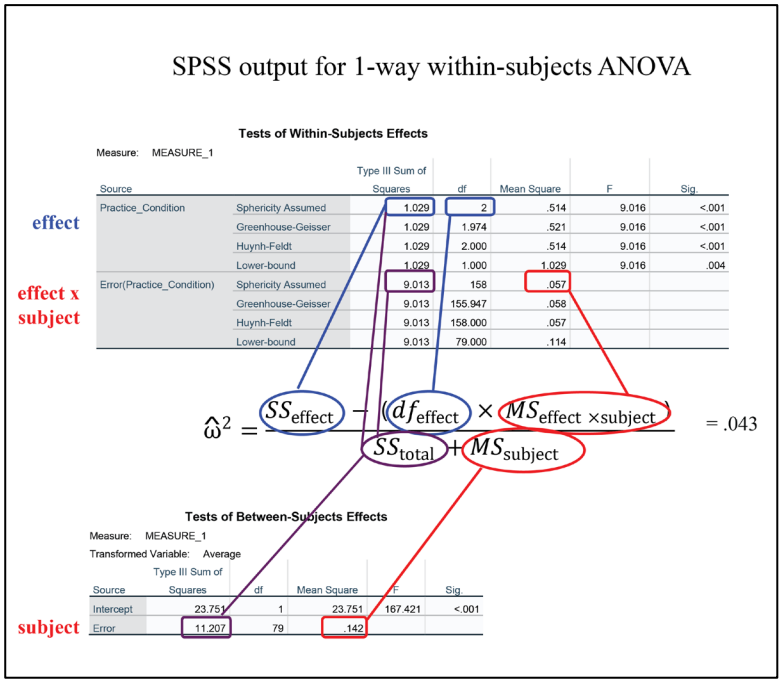
See Figure 2.3 for an example of SPSS output, showing the components used for omega squared. SPSS does not output the total sum of squares for within-subjects designs, and thus it is necessary to calculate SS_{total} from three values, as shown in Figure 2.3. It is necessary to use values from the "Tests of Between-Subjects Effects" table, even though this is a within-subjects design, because that is where the subject factor values can be found (labeled "error" by SPSS). The "intercept" factor in that table should be ignored for our purposes.

The exclusion of SS_{total} from SPSS output may relate to the discussion on whether SS_{subject} should be included in SS_{total}. For formulas that do partial out the systematic differences between subjects we recommend Olejnik & Algina (2000) and Keppel & Wickens (2004). It should be noted that these authors present different formulas that yield quite different results. This is because Olejnik and Algina (2000) add the product of the

error term and the number of *subjects* to the denominator, while Keppel and Wickens (2004) add the number of *observations*.

Figure 2.3

The Components for Standard Omega Squared in SPSS Output for a One-Way Within-Subjects Design.



Note. Data are from Finley et al. (2011), available in the supplemental materials. SPSS version 28.

Multi-Factor Within-Subjects Designs

For multi-factor within-subjects designs (two-way, three-way, etc.) partial omega squared can be calculated. Maxwell et al. (2018, p. 674) state: "We generally believe that the value of omega squared we calculate for a main effect in a factorial design should be identical to the value we would have obtained for that effect in a single-factor design." This reasoning follows the same logic as the partial omega squared formula for higher-order between-subjects designs. For partial omega squared the total sum of squares in the formula of standard omega squared is replaced with the components that form the total sum of squares in the one-way design:

$$\widehat{\omega}_{\mathrm{p}}^{2} = \frac{SS_{\text{effect}} - (df_{\text{effect}} \times MS_{\text{effect}\times\text{subject}})}{SS_{\text{effect}}+SS_{\text{effect}\times\text{subject}}+SS_{\text{subject}}+MS_{\text{subject}}} \tag{20}$$

See Figure 2.4 for an example of SPSS output from a two-way within-subjects ANOVA, showing the components used for partial omega squared. Again, the "Tests of Between-Subjects Effects" table is necessary to find the subject values (labeled "error" by SPSS), and the "intercept" factor should be ignored.

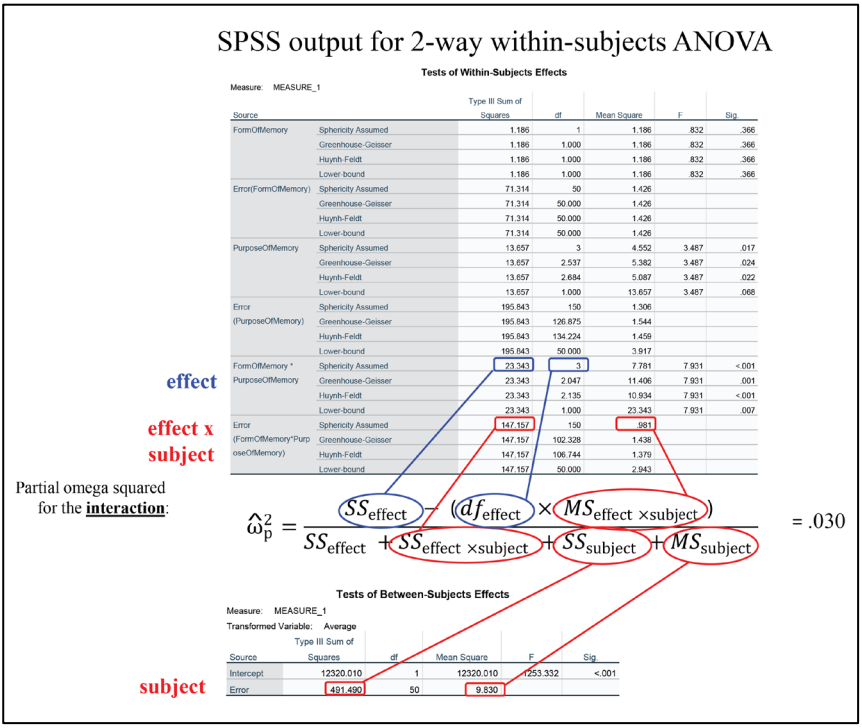
As with the one-way within-subjects design, the systematic differences between subjects are not partialled out, but all other factors are. Because partial eta squared does partial out the variance between subjects, Maxwell et al. (2018) argue that referring to this statistic as partial omega squared may lead to confusion and they refrain from doing so (p. 725). However, we argue that *not* making a distinction between this formula and the formula for standard omega squared, is still confusing. As we follow the idea that formulas for omega squared and partial omega squared should be comparable across designs, it is justifiable to define this formula as partial omega squared.

For between-subjects designs the formula for omega squared can be rewritten as a function of the *F* test statistic, *df*, and *N* (Formulas 14 and 15). This is possible for between-subjects designs because omega squared

is calculated with variance components from the effect and from subject error, both of which are also part of the calculation of *F*.

Figure 2.4

The Components for Partial Omega Squared in SPSS Output for a Two-Way Within-Subjects Design.



Note. Data are from Finley and Naaz (2023), available in the supplemental materials. SPSS version 28.

But rewriting the formula in such a way is impossible for designs that include within-subjects factors. For designs including any within-subjects factors (including split-plot designs), *F* is calculated without the variance due to subjects. It only includes the variance due to the interaction between the effect and the subjects ($\bar{A}$×S). As our formulas for omega

squared include the variance due to subjects, omega squared cannot be deduced from *F* for designs that include within-subjects factors. Some authors do suggest formulas to calculate omega squared from *F*. For the one-way within-subjects design, Keppel and Wickens (2004, p. 362) suggests using the same formula as for between-subjects designs (Formula 14 in this paper). As this excludes the variance due to subjects, this yields a higher value for omega squared than the formulas we have presented. For multifactor-designs, Keppel and Wickens (2004) present two formulas based on *F* to calculate a range in which omega squared falls. This often results in a rather large range, which is "too broad to be useful" (p. 427). It should be noted that the formula that they propose for the lower end of the range still yields a higher value than the formula we have presented, while the higher end of the range yields a lower value than partial eta squared. The bottom line is that there is no certain way to calculate omega squared from *F* for any within-subjects design.

Split-Plot Designs

A split-plot design is one in which there is at least one between-subjects factor and at least one within-subjects factor. These designs do include estimates of random error due to the within-subjects factor ($\widehat{\sigma}^2_{\text{effect} \times \text{subject}}$), as well as an error term that includes systematic differences due to the between-subjects factor ($\widehat{\sigma}^2_{\text{error}}$). For standard omega squared, the formula is as follows (Dodd & Schultz, 1973):

$$\widehat{\omega}^2 = \frac{SS_{\text{effect}} - (df_{\text{effect}} \times MS_{\text{appropriate term}})}{SS_{\text{total}} + MS_{\text{subject/A}}} \tag{21}$$

The appropriate error term for the numerator ($MS_{\text{appropriate term}}$) depends on the factor of interest. To clarify, we will use an example with between-subjects Factor A and within-subjects Factor $\overline{\text{B}}$. For the main effect of Factor A the appropriate term is $MS_{\text{subject/A}}$; for the main effect of Factor $\overline{\text{B}}$ and for the interaction A$\overline{\text{B}}$ the appropriate error term is

$MS_{\text{B}\times\text{subject/A}}$ (Maxwell et al., 2018). For calculating standard omega squared in split-plot designs with more factors (and thus with more possible interactions), the appropriate error term for main effects and interaction effects of between-subjects factors will always be $MS_{\text{subject/X}}$, where X stands for all the between-subjects factors in the design. When the effect of interest includes one or more within-subjects factors, the error term will be based on that factor(s). For example in an $AB\bar{C}S$ model, when investigating the $A\bar{\text{C}}$ interaction, the appropriate error term is $MS_{\text{C}\times\text{subject/AB}}$. In an $\text{A}\bar{\text{B}}\bar{\text{C}}\text{S}$ model, when investigating the $\text{A}\bar{\text{B}}\bar{\text{C}}$ interaction, the appropriate error term is $MS_{\text{BC}\times\text{subject/A}}$.

When calculating partial omega squared for a split-plot design the formula also depends on what type of factor is used. If the effect of interest includes one or more within-subjects factors, we use the following formula (see Figure 2.5a):

$$\widehat{\omega}_{\text{p}}^2 = \frac{SS_{\text{effect}} - (df_{\text{effect}} \times MS_{\text{effect}\times\text{subject/A}})}{SS_{\text{effect}} + SS_{\text{effect}\times\text{subject/A}} + SS_{\text{subject/A}} + MS_{\text{subject/A}}} \tag{22}$$

Note that even an interaction between a within-subjects factor and a between-subjects factor should use the within-subjects formula.

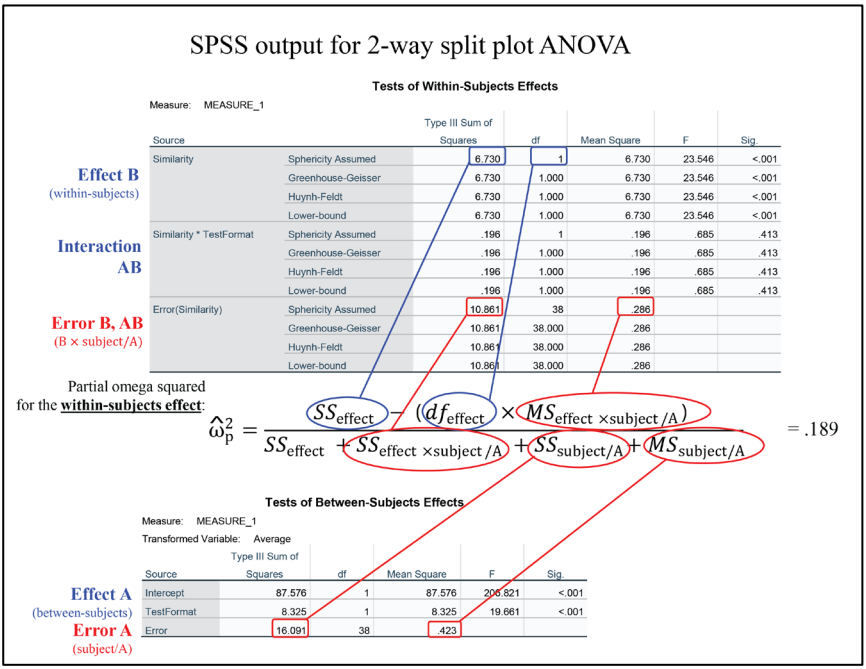
If the effect of interest concerns only between-subjects factors, we use the following formula (see Figure 2.5b):

$$\widehat{\omega}_{\text{p}}^2 = \frac{SS_{\text{effect}} - (df_{\text{effect}} \times MS_{\text{subject/A}})}{SS_{\text{effect}} + SS_{\text{subject/A}} + MS_{\text{subject/A}}} \tag{23}$$

See Figures 2.5a and 2.5b for an example of SPSS output from a two-way split-plot ANOVA, showing the components used for partial omega squared for the within-subjects factor (2.5a), and for the between-subjects factor (2.5b).

Figure 2.5a

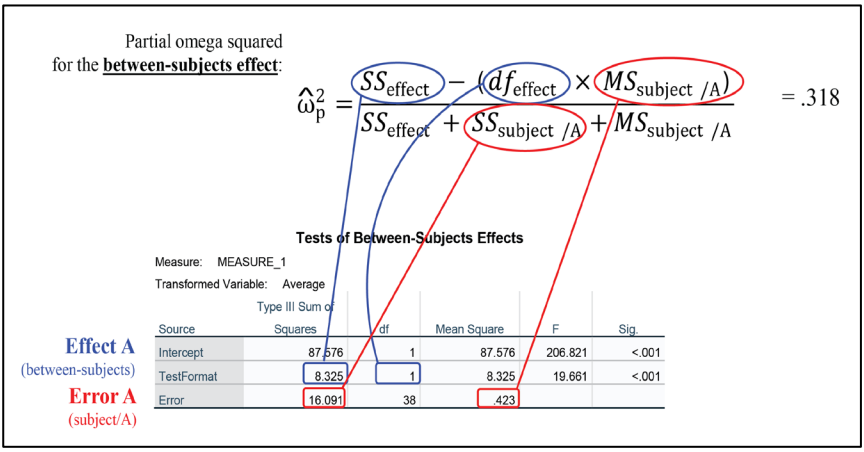
The components for partial omega squared in SPSS output for the effect of the within-subjects factor in a two-way split-plot design.



Note. Data are from Finley et al. (2015), available in the supplemental materials. SPSS version 28.

Figure 2.5b

The components for partial omega squared in SPSS output for the effect of the between-subjects factor in a two-way split-plot design.



Note. Data are from Finley et al. (2015), available in the supplemental materials. SPSS version 28.

In theory, Formulas 21, 22 and 23 can be used for split-plot designs with any number of factors, as long as the right error terms are selected. Olejnik and Algina (2000) present other formulas for split-plot designs (Table 18, pp. 278-279). For the effects that include within-subjects factors, the systematic differences are partialled out. However, readers should use those formulas with caution, as they apparently contain mistakes for the within-subjects factors; the examples provided by Olejnik and Algina in Table 18 of their paper are calculated differently from what the formulas prescribe.

Alternatives for Dealing with Non-Additivity

We have presented formulas in which the variability due to subjects is not partialled out. Alternatives to this method exist. Some authors have proposed ways to calculate a possible range for omega squared (Myers & Well, 2003; Tabachnick & Fidell, 2007a). Other authors propose specific

formulas in which the denominator is increased by adding an error term for each non-additive effect. Interested readers can find these formulas for one, two and three-way designs in Dodd and Schultz (1973); and Vaughan and Corballis (1969). It should be noted that Vaughan and Corballis exclude SS_{subjects} from SS_{total}, while Dodd and Schultz do not specify whether they include SS_{subjects} in SS_{total}. We argue that by not partialling out the variance caused by systematic differences between subjects, the problem of non-additivity is avoided altogether. The papers that provide formulas that correct for non-additivity do not provide formulas for partial omega squared (Dodd & Schultz, 1973; Vaughan & Corballis, 1969). Following the logic that the partial effect size should be equivalent to the effect size that would be obtained in a one-way design (Keppel & Wickens, 2004; Maxwell et al., 2018), one can use the correction provided by Dodd & Schultz (1973) for the one way design by adding $(N \times MS_{\text{effect}\times\text{subject}})$ to the denominator:

$$\widehat{\omega}_{\text{p}}^2 = \frac{df_{\text{effect}}(MS_{\text{effect}} - MS_{\text{error}})}{SS_{\text{effect}} + SS_{\text{effect}\times\text{subject}} + SS_{\text{subject}} + MS_{\text{subject}} + (N \times MS_{\text{effect}\times\text{subject}})} \quad (24)$$

Another solution to the problem of non-additivity is to investigate the contrasts for each level of the within-subjects factor of interest (Boik, 1981). When investigating the 1 *df* contrasts, the assumptions of additivity and sphericity are no longer in effect. Lastly, it is important to note that using (partial) eta squared does not circumvent the problems caused by non-additivity. Eta squared "treats the ambiguity in the error variances in a way that gives it the largest possible value" (Keppel & Wickens, p.428), giving a result that is positively biased.

Remaining Obstacles in Using Omega Squared

Partialling Out Factors

Partial eta squared is the most popular effect size (Alhija & Levy, 2009; Fritz et al., 2012; Peng et al., 2013; Sun et al., 2006; Zhou & Skidmore, 2017)

and some authors recommend always reporting a partial measure of effect (Keppel, 1991; Keppel & Wickens, 2004). However, partial effect sizes can be misleading and harder to interpret than standard effect sizes (Cohen, 1973; Levine & Hullett, 2002; Olejnik & Algina, 2003). The more independent causes there are for an effect, the smaller the effect size is for any of the individual causes. This is disregarded when using a partial effect size (Levine & Hullett, 2002). Identifying causes in a model can reduce the unexplained variance by adding them as a factor (Tabachnick & Fidell, 2007a). However, it is likely that any factor that is added to an ANOVA design can simultaneously reduce error while instilling additional variance (Levine & Hullett, 2002). This makes it hard to decide whether a factor can be partialled out. To use partial omega squared it is necessary that the other factor(s) in the design can be seen as extrinsic (Maxwell et al., 2018). It has to be reasonable to partial out Factor B when investigating the effect of Factor A. Cohen (1973) advised that manipulated variables and variables that are held constant can usually be safely partialled out. When it is likely that an additional factor has an influence on Factor A, the factor is called intrinsic and it should not be excluded when calculating an effect size (Maxwell et al., 2018). When the effect size is inflated because intrinsic factors are partialled out, it is no longer comparable to studies that do not include these factors in their models (Olejnik & Algina, 2003). Therefore, when the other factors are (possibly) intrinsic, it is better to provide a non-partial effect size (Levine & Hullett, 2010). It is of course quite possible that a design contains both extrinsic and intrinsic factors, which complicates the choice between the standard and partial effect size. A solution is offered in generalized omega squared ($\widehat{\omega}_G^2$), which is explained by Olejnik and Algina (2003). They go into many different designs (including random and fixed factors), where some factors are regarded as extrinsic and others as intrinsic. As this results in a myriad of possible combinations, this goes beyond the scope of this paper. However, by sharing the full ANOVA tables, researchers can enable others to calculate generalized omega squared. In their paper on generalized effect

sizes, Olejnik and Algina also describe generalized eta squared ($\hat{\eta}^2_G$). Similar to generalized omega squared, this effect size offers solutions to the problem of extrinsic and intrinsic factors. Additionally, its calculation is less complicated than the calculation of generalized omega squared. However, like its non-generalized counterpart, generalized eta squared is a descriptor of the sample and not of the population, and it does not compensate for the overestimation of the variance due to treatment (Olejnik & Algina, 2003, p. 441).

It can be hard to decide whether a factor should be regarded as intrinsic or extrinsic. Another solution can be to report both partial and standard effect sizes (Cohen, 1973). We recommend reporting explicitly which factors are considered intrinsic and which extrinsic. Again we advise reporting the formulas you use for the effect size, as well as including an ANOVA table.

Confidence Intervals for Omega Squared for Between-Subjects Designs

The APA recommends reporting confidence intervals for statistics, including effect sizes (7th ed., section 3.7; see also Fritz et al., 2012). As with any statistic, the effect size we calculate from our data is an *estimate* of the true effect size in the population. This is why we put the hat symbol over the Greek letter omega. How confident should we be in this estimate of effect size? That is what the confidence interval tells us. The wider the interval, the less precise is our estimate; the narrower the interval, the more precise is our estimate. A 95% confidence interval around a sample statistic means that if we were to endlessly resample the population with the same sample size each time, and put that same confidence interval width around the calculated sample statistic each time, that interval would contain the true population parameter 95% of the time. Unfortunately, determining confidence intervals is not straightforward for effect sizes such as eta squared and omega squared (Fidler & Thompson, 2001; K. Kelley, 2007; Thompson, 2007). In fact, as of this

writing there is no consensus on how to correctly calculate omega squared confidence intervals for within-subjects factors (K. Kelley, personal communication, March 17, 2022). However, it can be done for between-subjects factors.

For statistics such as the mean, a confidence interval can be obtained using a single simple formula. This is not possible for statistics such as eta squared or omega squared. But there is a roundabout way to do it, which we will now explain.

Remember that each effect size is tied to an effect. In ANOVA designs, each effect is tested using an *F* test statistic. *F* is a ratio (e.g., MS_{effect}/MS_{error} in a one-way between-subjects design, also called $MS_{between}/MS_{within}$). When there is no effect, the numerator and denominator are equal, yielding $F = 1$. If there is truly no effect in the population, then we would expect a sampling distribution of *F* values that is centered on $F = 1$, and skewed with a long tail to the right, since a ratio cannot go below 0. This is called the "central" *F* distribution. When we conclude that an effect is statistically significant, we are saying that the *F* test statistic that we obtained from our sample has less than a 5% probability of coming from that central *F* sampling distribution expected from the null hypothesis.

So then, what sampling distribution *did* our *F* statistic most likely come from? A different *F* distribution that is shifted over to the right by a *noncentrality parameter*, which is called lambda (λ). For a particular *F* statistic, lambda can be estimated iteratively by computer software. Software packages essentially use very advanced lookup tables. As there is an enormous number of possible lambda tables, finding the right lambda is virtually impossible without specialized software. Furthermore, the same methods can give us confidence intervals for lambda.[13]

[13] For more detailed discussion of noncentral distributions and their role in determining confidence intervals, see: Cumming and Finch (2001), Smithson (2001), and Steiger and Fouladi (1997).

Why is this relevant to omega squared? Because, for between-subjects designs, omega squared is related to lambda by a simple formula:

$$\omega^2 = \frac{\lambda}{\lambda + N_{\text{total}}} \tag{25}$$

Thus, we can calculate the confidence intervals for lambda, then convert those to the confidence intervals for omega squared.[14]

The steps for this procedure are laid out in Steiger (2004, p. 168), including additional details such as determining when one or both limits of the CI would be set to zero. K. Kelley (2007, section 4) covers similar procedures for CIs for R^2 in a regression context (which is eta squared in an ANOVA context). Unfortunately, as of this writing, such procedures for determining CIs are not easily available in popular statistical software such as SPSS.

Ambitious researchers may refer to Fidler and Thompson (2001, pp. 592-593) and Smithson (2001, Appendix), for syntax-based methods to calculate CIs in SPSS. Related procedures have been implemented in Excel (Cumming & Finch, 2001; J. B. Nelson, 2016), but do not provide a ready-made solution for researchers who just want to determine a CI with minimal struggle.

However, Ken Kelley has written a package of functions that can do what we need, called MBESS (K. Kelley, 2007, 2022), which can be used with the popular free statistical software called R. For the many researchers who are not well-versed in R, its use can be daunting and its documentation cryptic. We have faced our own confusion in using R, so we want to help readers by explaining the exact steps necessary to use R

[14] In addition to the noncentrality parameter approach to constructing confidence intervals, K. Kelley (2005) proposed two bootstrapping approaches, which were further studied by Algina et al. (2006) and Finch and French (2012). We do not address those approaches here.

to determine CIs for omega squared. This same procedure works for partial omega squared.

1. Go to https://cran.r-project.org/ and download R for the appropriate operating system for your computer (e.g., Mac OS, Windows). Use the downloaded file to install R.
2. Open R on your computer. You will see that it is a command line interface, in which you must type specific commands to tell the program what to do. Whenever you type a command, you must press the enter or return key to run the command. There are graphical user interfaces that can be installed for R, such as R Studio and R Commander, but they are not necessary.
3. Enter the following command: `install.packages("MBESS")`
   This command will retrieve the MBESS package of functions from R servers and then install that package. You only need to do this step once; the next time you use R, MBESS will already be installed.
4. Enter the following command: `library("MBESS")`
   This command will load the MBESS package so that its functions are ready to use. You will have to do this step every time you use R in order to use the MBESS package. Optionally, if you would like to read the manual for this package, you can visit https://cran.r-project.org/web/packages/MBESS/index.html or enter the following command in R: `help("MBESS")`
5. You will need to have the following values ready: F for the effect of interest, degrees freedom for the numerator of F (df1, aka df_{effect}), degrees freedom for the denominator of F (df2, aka df_{error}), and the total sample size of the design (N). Note that you do not even need the value of omega squared.

6. Now you will use the function ci.omega2 [added to the MBESS package in version 4.9.0]. Optionally, if you would like to read the manual for this function, enter the command:
   ```
   help(ci.omega2)
   ```
   To actually show you how to run the function, we will use values from the one-way between-subjects design example data (Finley et al., 2017) available in the supplemental materials of this paper (see Appendix D).
7. Enter the following command:
   ```
   ci.omega2(F.value=2.698, df.1=16, df.2=536,
   N=553)
   ```
8. The output gives separate values for the lower and upper limits of the confidence interval, and looks like this:
   ```
   $lower_limit_omega2
   [1] 0.0150022
   $upper_limit_omega2
   [1] 0.09235316
   ```
9. Omega squared from this example was separately calculated as .047 (Figure 2.1). We would report as follows: $\widehat{\omega}^2$ = .047, 95% CI [.0150, .092].

The confidence interval coverage is set to .95 (95%) by default for the `ci.omega2` function. If you wanted to instead calculate a 90% CI, you would add conf.level=.90 so that you would enter the following command:

```
ci.omega2(F.value=2.698, df.1=16, df.2=536,
N=553, conf.level=.90)
```

Note that if a value "NA" is output for one or both limits, just use 0 in place of NA. Here is an example using the main effect of Actual Test Format from our 2-way between-subjects example data (Finley & Benjamin, 2012):

Command:

```
ci.omega2(F.value=31.496, df.1=1, df.2=1, N=100)
```

Output:

```
$lower_limit_omega2
[1] NA
$upper_limit_omega2
[1] 0.6201419
```

Partial omega squared for this effect was separately calculated as .234 (see the online supplemental materials). We would report as follows: $\widehat{\omega}_p^2$ = .234, 95% CI [0, .620].

Confidence Intervals for Designs Including Within-Subjects Factors

When calculating confidence intervals, matters are more complicated for designs that include one or more within-subjects factors. As far as we know, no research into CIs for within-subjects and split-plot designs has been published. Whenever a within-subjects factor is included, omega squared is not directly related to *F*, as the error due to subjects is not part of the formula for *F*. Subsequently, omega squared is also not directly related to lambda. Perhaps future researchers could solve this problem by calculating a correction for lambda, or by developing a formula for omega squared that does derive from *F* even when within-subjects factors are present. Any such work that would provide more clarity on CIs for within-subjects designs would necessitate a Monte Carlo simulation study (K. Kelley, personal communication, March 21, 2022). This goes beyond the scope of this paper.

Quick Guide: Calculating, Reporting, and Interpreting Omega Squared

When reporting omega squared, we advise the following:

1. Use Appendix B to find the appropriate formula for your ANOVA design and the factor of interest. For each component in the formula, retrieve the value from the ANOVA table(s) output by the statistical software you are using. Examples of using SPSS output are shown in Figures 2. 1-5b, along with data and syntax in the supplemental materials. Use a tool such as Excel to perform the actual calculation of the formula using the component values; examples of this are also found in the supplemental materials.
2. Report omega squared with three decimals, as even a value as small as .008 can make a considerable difference in power analysis, especially for smaller effects (Lakens, 2015). Do not include a leading zero (e.g., 0.008), because that is never necessary for statistics that cannot exceed a value of one.
3. It is possible to find a negative effect size for omega squared. This happens when the within group error variance is so high that any treatment effects are either absent or impossible to detect (Keppel & Wickens, 2004). It should be noted that a negative value for omega squared does not indicate a negative effect, but signifies the absence of effect. It could be argued that when omega squared is negative, it should be set to zero. However, for clarity we advise reporting the actual value for omega squared, even if it is negative, accompanied by an interpretation in which the absence of effect is stated (Okada, 2017).
4. Always report the formula used to calculate the effect size, accompanied by a reference to the source of the formula (e.

g., cite this paper; Kroes & Finley, 2023). You can conveniently do this in the beginning of your Results section, or as a footnote.

5. Include a confidence interval for the effect size if possible (i. e., for between-subjects designs).
6. Include the complete ANOVA table(s) in your paper, Appendices, or supplementary materials to enable other researchers to calculate a different effect size when preferred. You may wish to reformat the tables output by SPSS to be consistent with APA style. Note that for within-subjects designs, SPSS outputs the subject factor (labeled "error") in a separate table called "Tests of Between-Subjects Effects" so be sure to include those values.

When interpreting the effect size, some guidelines exist on what constitutes a small, medium, and large effect (Kirk, 1996, p. 751). However, evaluating the effect size in such a manner should be done with much caution. Most importantly, these standardized classifications say little about the actual practical importance and practical significance of an effect (Keppel & Wickens, 2004; Maxwell et al., 2018). A small effect size could be extremely valuable if it, for instance, describes the effect of a life-saving drug. A large effect size could indicate that an effect is already commonly known and therefore not useful to investigate (Keppel & Wickens, 2004). Additionally, effect sizes vary across specific areas of research (Maxwell et al., 2018). What constitutes a groundbreaking discovery in one field, may be inconsequential in another. In most fields it is actually unclear what the smallest effect size of interest is (Lakens, 2022).

Conclusion and Discussion

Reporting measures of effect size has been a growing practice over the past decades, with eta squared and partial eta squared being reported most frequently for ANOVA designs (Alhija & Levy, 2009; Finley et al.,

2017; Fritz et al., 2012; Kirk, 2012; Zhou & Skidmore, 2017). Eta squared and partial eta squared are problematic because they are positively biased and thus tend to overestimate the population value (Albers & Lakens, 2018; Keppel & Wickens, 2004; Keselman, 1975; Lakens, 2015; Levine & Hullett, 2002; Yigit & Mendes, 2018). Omega squared and partial omega squared are less biased (Field, 2017; Keppel & Wickens, 2004; Lakens, 2013, 2015; Tabachnick & Fidell, 2007a; Yigit & Mendes, 2018) but are underused due to inconvenience, lack of guidance and unfamiliarity (Kirk, 1996; Zhou & Skidmore, 2017). There is lack of clarity about the varying formulas for different designs, especially for designs including within-subjects factors (Keppel & Wickens, 2004; Maxwell et al., 2018; Olejnik & Algina, 2000; Tabachnick & Fidell, 2007a).

To help with this lack of clarity we have provided formulas that can be used for between, within and split-plot ANOVA-designs with fixed factors (for a quick overview, see Appendix B). We have provided a guide to calculate confidence intervals for between-subjects designs, and we encourage researchers to investigate confidence intervals for designs that include within-subject factors. Disagreement exists whether variance due to systematic differences between subjects should be partialled out in designs that include within-subjects factors (Maxwell et al., 2018; Tabachnick & Fidell, 2007a). We agree with Maxwell et al. (2018) that it seems inadvisable to disregard variance caused by subjects, and we have presented formulas that include these differences. We argue that this offers an acceptable solution to the problem of non-additivity. Another problem is deciding whether factors in a model should be seen as intrinsic or extrinsic, and whether standard or partial omega squared should be used. A solution could be to present both (Cohen, 1973) or to calculate an alternative, like generalized omega squared (Olejnik & Algina, 2003).

SPSS and other statistical software packages have a great influence on what effect sizes are being used (Fritz et al., 2012; Kirk, 1996; Levine & Hullett, 2002). Statistical software developers can play a helpful role by including more options for omega squared, especially for the fixed

between designs, where there is no controversy over the formulas. Furthermore, the software should state the exact formulas being used. Since the APA started requiring reporting effect size, the use of effect sizes in general has increased considerably (Alhija & Levy, 2009; Finley et al., 2017; Peng et al., 2013; Zhou & Skidmore, 2017). The APA could play a facilitating role by offering more guidance about effect sizes, and by encouraging publishing ANOVA tables. We recommend researchers to always report the formulas used to calculate the reported effect sizes, and if possible to share the ANOVA table(s). This provides others the possibility to partial out specific factors, and it facilitates power calculations and meta-analyses. Most importantly, this aids in cumulative science, as it clarifies the methods used and offers researchers opportunities to make replications and gain deeper understanding of investigated effects.

Chapter 3

Sexual Double Standards?
How Parents Talk to Their Child About Adolescents Who Date and Kiss

Antoinette D. A. Kroes, Tessa M. Van de Rozenberg, Lotte D. Van der Pol, Anja Van der Voort, Marleen G. Groeneveld, Judi Mesman

Submitted for publication.

Abstract

According to the sexual double standard (SDS), men are praised for engaging in sexual activity, while women are penalized. However, research findings on the perpetuated existence of the SDS are inconsistent. Informed by cognitive social learning theory and sexual script theory, we aimed to shed light on the existence of the SDS within a relevant social setting. Within parent-child dyads, we examined parents' responses to vignettes describing adolescents who engage in multiple successive relationships and kissing strangers. Participating fathers and mothers came from 199 Dutch families. Through thematic content analysis, five themes emerged: normative sexual development, promiscuity and impulsivity, dependent and romantic creatures, social penalties and praises, and ambiguous and unclear answers. The most common scripts involved normalization of the adolescents' behavior, for boys and girls alike. The findings confirm endorsement of the SDS in some parents' responses by referring to common sexual scripts and gendered expectations or by instilling social penalties and praise. Girls were positioned as less agentic and more romantic than boys were, especially by parents who spoke to their daughters. Some parents called girls a slut, while boys were regularly labeled as players. We advise additional research into the transmission of the SDS and other sexual scripts, using more diverse samples.

Keywords: adolescents; script theory; parent-child communication; social learning theory; sexuality-related attitudes

“If you look back in history
It's a common double standard of society
The guy gets all the glory, the more he can score
While the girl can do the same and yet you call her a whore
I don't understand why it's okay
The guy can get away with it, the girl gets named”
(Christina Aguilera & Lil’ Kim, 2002)

In the song ‘Can’t hold us down’, Christina Aguilera and Lil’ Kim draw attention to how men gain status by having multiple sexual partners, while women receive backlash for the same behavior. As they point out, men being allowed more sexual freedom than women is a common conception. In research this phenomenon is part of the so-called (hetero)sexual double standard (SDS) (Reid et al., 2011; Zaikman & Marks, 2017). The SDS has many adverse effects. Endorsement of the SDS is associated with less sexual pleasure for both men and women (Laan et al. , 2021; Sanchez et al., 2012). By adhering to the traditional roles prescribed by the SDS, the ability to engage in authentic and rewarding sexual behavior is harmed. The SDS pressures men to hook up with (multiple) women, to prove that they are ‘real men’ (Currier, 2013). Men who do not adhere to the SDS risk homophobic victimization (Hird & Jackson, 2001; Pascoe, 2007). Women and girls especially experience adverse effects of the SDS. The possibility of being shamed for sexual behavior can, for instance, have negative effects on their mental and physical health (Goblet et al., 2021), and decrease condom use (Lefkowitz et al., 2014). According to the SDS, women and girls are less sexual and more vulnerable than men and boys (Endendijk et al., 2020), which leads to less access to health care services and information (World Health Organization, 2002), and less sexual autonomy (Emmerink et al., 2018). Because of these adverse effects, it is important to investigate contemporary sexual scripts of the SDS. Nonmarital and casual sex have become more acceptable for both men and women over time, especially

in countries with higher levels of gender equality (Bordini & Sperb, 2013; Endendijk et al., 2020). According to some recent studies, the SDS is no longer present (Kettrey, 2016), or has even reversed in certain contexts (Hensums et al., 2022; Kreager et al., 2016; Papp et al., 2015). Simultaneously, evidence for the SDS is still being found across cultural contexts (e.g., Endendijk et al., 2022; Hendrickx et al., 2022; Wesson, 2022; Wu et al., 2021). The inconsistencies in findings may be clarified by the use of a theory-based approach, and by improving methodologies (Zaikman & Marks, 2017).

We aim to investigate whether the SDS is endorsed by parents of adolescent children, through a thematic content analysis of reactions to vignettes concerning behaviors relevant to the SDS.

Defining the SDS

A first possible explanation for inconsistencies in SDS research, is that the concept is sometimes oversimplified (Zaikman & Marks, 2017). The SDS was first introduced and defined by Reiss (1956, 1964) within the context of premarital sex. According to Reiss, the SDS entails that premarital sex is wrong for women and permissible for men. In contemporary research, the SDS is often defined as men being rewarded for sexual behavior, while women are penalized (Zaikman & Marks, 2017). A broader range of sexual behaviors has since been studied, including one night stands, threesomes, (Bordini & Sperb, 2013), casual sex (Fjær et al., 2015; Kettrey, 2016; Reid et al., 2011), dating at a young age (Axinn et al., 2011), and kissing/making-out (Farvid et al., 2017). Rewards and penalties for sexual behavior have been measured with constructs like popularity among peers (Kreager et al., 2016; Kreager & Staff, 2009), ratings of traits like dominance, likeability, intelligence, and success (Marks et al., 2019; Marks & Fraley, 2007; Young et al., 2016), and social rejection (G. Smith et al., 2008). In qualitative studies among students, men stated that hookups gave them a better reputation and made them more popular (Currier, 2013), men with multiple partners were called ‘players’, and hooking up was competitive:

having more sexual partners meant winning (Fjær et al., 2015). Being a 'player' can be seen as a way to gain social, masculine, heterosexual status (Wilkins, 2012). Conversely, female students engaging in the same behavior were at risk of being called a 'slut', which was something they feared and avoided (Currier, 2013; Fjær et al., 2015), indicating its low status.

However, defining the SDS as men being rewarded for sex and women being penalized is an oversimplification (Tolman et al., 2015; Zaikman & Marks, 2017). The SDS also concerns the gendered expectations of societal roles, behaviors, and characteristics for men and women in relation to sexuality (Axinn et al., 2011; Zaikman & Marks, 2017). Especially important in these expectations is sexual agency. Men are expected to be sexually agentic, which entails being dominant, powerful, and assertive, and the initiator of sex (Endendijk et al., 2020; Muehlenhard & McCoy, 1991). When men do not adhere to these expectations, a common penalty is being referred to with derogatory terms like 'fag' or 'wuss' (Pascoe, 2007). Conversely, women are expected to be passive, submissive, and reactive (De Meyer et al., 2017; Endendijk et al., 2020; Zaikman & Marks, 2017). This corresponds to the idea of women as "naturally" romantic beings (Elliott, 2010; Farvid & Braun, 2014), while men supposedly only want sex and are uninterested in relationships (Pascoe, 2007; Schalet, 2011). While male sexuality is seen as incessant, dominant, and uncontrollable, female sexuality is framed by gatekeeping sex (Aubrey et al., 2020; Hird & Jackson, 2001; Tolman et al., 2015) which means that women are supposed to refuse and deter men (Aubrey et al., 2020; Bay-Cheng, 2015; Tolman et al., 2015). Giving permission too easily, or only because a man wants it, can lead to being called a slut (Armstrong et al., 2014; Bay-Cheng, 2015), while not giving enough permission may lead to being called a prude, which is negative too (Dajches et al., 2021; Tolman et al., 2015). So, while society generally expects men to be sexually active, and penalizes those who are not, women are expected to navigate a thin line between being too active and not being active enough (Currier, 2013).

The SDS can thus be defined as follows: the differential expectations, treatment, and evaluations of (heterosexual) men and women based on their sexual behavior.

Persistence and transmission of the SDS

The gendered differential expectations, treatment, and evaluations that make up the SDS can be explained by social cognitive learning theory. According to this theory, individuals learn by observing and imitating behavior of others, as well as by observing the social reinforcement and restrictions that follow from behavior (Bandura & Walters, 1963; Bussey & Bandura, 1999). When the SDS is endorsed by important socialization agents, like the media, peers, and parents, adolescents are likely to be socialized into the endorsement and perpetuation of the SDS (Bandura & Walters, 1963; Bussey & Bandura, 1999; C. Connell & Elliott, 2009; Kreager et al., 2016; Zaikman & Marks, 2017). The normative social construction of sexual behaviors is further explained by sexual script theory (Simon & Gagnon, 1984; Wiederman, 2015). Sexual scripts are (metaphorical) scripts that inform individuals on the meaning and appropriateness of sexual behavior and sexual roles. Sexual scripts exist on three levels: cultural, interpersonal, and intrapsychic. The three levels of sexual scripts are dynamically interrelated and constantly co-act in behavior and social interaction. Cultural sexual scripts are the most abstract scripts. They are often conveyed by institutions like the media, government, religion, and schools. By legalizing and stimulating certain sexual behaviors, and by forbidding and warning against others, cultural scripts set the stage for what sexual behaviors are praised and what sexual behaviors are punished (Wiederman, 2015). For instance, media like television programs often communicate the SDS through scripts of men being fixated on sex, sexually dominant, and uncontrollable, and gaining status from sexual relationships (Aubrey et al., 2020; Dajches et al., 2021; Kim et al., 2007). Conversely, women are positioned as responsible for deferring men, and they risk social penalties when they have sex in the wrong way.

Interpersonal scripts are the ways in which individuals adapt and divert from cultural scenarios in concrete situations. According to sexual script theory, people will only apply cultural scripts when they are convinced that this will yield a socially acceptable and positive outcome (Simon & Gagnon, 1986). For instance, while parents are aware of the cultural scripts, they may modify or completely divert from these scripts when discussing sexual behavior with their offspring. They could do this by encouraging their daughters to ward off suiters because "boys only want one thing", or by applauding women on television who initiate pleasurable sexual encounters. Intrapsychic scripts are scripts at the most personal level, involving the way in which an individual (re)organizes cultural and interpersonal scripts and relates those to personal wishes and desires (Simon & Gagnon, 1984). These scripts include fantasies, feelings of arousal, memories, and mental rehearsals (Wiederman, 2015). For example, an adolescent girl's fantasies about and desire to have casual sex.

Thus, sexual scripts can cover many gendered sexual expectations. Therefore, investigating these scripts is important to clarify to what extent the SDS still exists within relevant social settings, gain understanding into what scripts are part of the SDS, and to explain research inconsistencies. To further resolve inconsistencies, it is important to discuss the ways in which research methodologies can be improved.

Study design aspects when investigating the SDS

Besides a clear definition and a theory-based approach, inconsistencies in SDS may be remedied by improving methodology (Bordini & Sperb, 2013; Endendijk et al., 2020; Zaikman & Marks, 2017). We will discuss three methodological aspects that can explain inconsistent results: the instruments used, the sexual behavior that is studied, and the cultural context. The issues raised in relation to this aspect are used to determine our research aim and design.

First, instrument choice is important when investigating the SDS. In quantitative research, Likert-type scales are used often. These consist of statements related to the SDS, for which participants have to indicate their level of (dis)agreement. Examples of these statements are "Girls should act in a more reserved way concerning sex than boys" (Emmerink et al., 2017, p. 1707) and "I kind of admire a guy who was had sex with a lot of girls" (Muehlenhard & Quackenbush, 2013, p. 200). In other quantitative studies, participants are presented with a description of a person who engages in sexual behavior and are then instructed to rate the person on concepts like intelligence, dominance, and success (Marks & Fraley, 2007; Young et al., 2016). Using these types of instruments means that specific sexual scripts are selected a priori, and tested deductively, which ignores possible shifts in perception and makes discovery of new scripts impossible (Elliott, 2010; Simon & Gagnon, 2003). This approach may result in an operationalization of the SDS that is too narrow, and therefore could lead to under-detection. Another problem with some instruments is that they cannot distinguish between an egalitarian view and a reversed SDS, because they only include items that confirm the SDS (e.g.: "It is up to the man to initiate sex."), and no items in the opposite direction (e.g.: "It is up to the woman to initiate sex.") (Caron et al., 1993; Emmerink et al., 2017; Endendijk et al., 2020). Also, Likert-style instruments are of an explicit nature, meaning that participants are quite conscious of what is being measured, which may elicit social desirable answers (Endendijk et al., 2020). Social desirability is especially problematic when investigating attitudes towards controversial and sensitive topics like gendered sexual attitudes (Axinn et al., 2011). Because gender attitudes have generally become more egalitarian (Endendijk et al., 2022; Liben, 2016), participants may find that endorsing the SDS is controversial, increasing their inclination to give social desirable answers (Axinn et al., 2011). Meta-analysis has revealed that in studies using these types of scales the SDS is generally not found (Endendijk et al., 2020). Instruments that measure SDS endorsement (relatively) implicitly can be

more useful. Implicit instruments elicit responses that are more automized, involve less conscious decision-making, and less awareness of what is being measured. These implicit instruments are found to be a better predictor of behavior and are less prone to social desirability (Endendijk et al., 2020). Quantitative research is commonly conducted outside of relevant interpersonal social contexts, like questionnaires conducted among university students (Bordini & Sperb, 2013; Endendijk et al., 2020; Kreager et al., 2016), which may underestimate the endorsement of the SDS (Marks & Fraley, 2007). For example, while one might desire to hold men and women to the same standards and will indicate that when asked to respond to an explicit statement, one may endorse a double standard in conversation with their pastor, partner or child (Kreager et al., 2016; Kreager & Staff, 2009; Wiederman, 2015). Thus, research on the SDS can be enriched and informed by the use of more implicit instruments and a more qualitative approach (Bordini & Sperb, 2013), specifically in finding changes to sexual scripts and possibly detecting new scripts.

Second, the cultural context affects the endorsement of the SDS. Evidence suggests that there is lower endorsement in countries with a higher level of gender equality (Bordini & Sperb, 2013; Bosson et al., 2022; Endendijk et al., 2020). As most studies on the SDS are conducted in the U.S., research in other countries can contribute to gaining a better understanding of the SDS (Bordini & Sperb, 2013; Endendijk et al., 2020; for examples of research in other countries see De Meyer et al., 2017; Hendrickx et al., 2022; Wu et al., 2021). We will investigate the SDS in the context of Dutch (the parents of) adolescents. In the past, some researchers have compared how adolescents and parents discuss and negotiate sexuality in the U.S. and the Netherlands (Schalet, 2000, 2010, 2011; Wouters, 2013). In these studies, Dutch parents are described as more lenient and approving of developing sexuality, compared to their U. S. counterparts. Dutch parents were also more likely to describe adolescents' sexuality and romantic relationships as normal and

something that adolescents can (learn to) self-regulate than U.S. parents. More recent meta-reviews of parental sex education and sex communication indicate that U.S. parents remain more conservative than European parents, although direct comparisons have not been made in recent years (Ashcraft & Murray, 2017; Flores & Barroso, 2017; Noorman et al., 2022).

Third, it is important to focus on sexual behaviors that are particularly relevant to the era and studied population, as inconsistent evidence for the SDS may be due to a focus on sexual behaviors that are not fitting with participants' life phase (Bordini & Sperb, 2013; Kreager et al., 2016). According to large-scale research among Dutch adolescents the median age for initial sexual behaviors has been rising over the past decades (De Graaf et al., 2017). The median age for first 'French' kissing increased from 13.5 years in 2005 to 14.4 years in 2012 and to 15.4 years in 2017. For sexual intercourse, the median age increased from 17.1 years in 2005 and 2012 to 18.0 years in 2017 (De Graaf et al., 2005, 2012, 2017). More recent data is not (yet) available, but as a later sexual debut is for instance associated with social media use, it can be expected that the increase in median age for partnered sexual behavior will persist (De Graaf et al., 2022). Romantic relationships are common from childhood onward, and adolescents find relational aspects of their developing sexuality more important than physical aspects (Boelhouwer et al., 2014). Studying sexual behavior like romantic relationships and kissing is therefore appropriate when investigating the SDS in the context of (parents of) adolescents (Kreager et al., 2016).

Adolescence and parents as socializing agents

Adolescence is particularly interesting when investigating the SDS, because it is when individuals explore, develop and refine their sexual scripts (Martinez et al., 2021; Simon & Gagnon, 1984, 1986; Wiederman, 2015; World Health Organization, 2002). While there has been an increase of research into the SDS among adolescents over the past decade (see for

instance, De Meyer et al., 2017; Emmerink et al., 2018; Endendijk et al., 2022; Hendrickx et al., 2022; Hensums et al., 2022), it is a relatively understudied life phase (Bordini & Sperb, 2013; Endendijk et al., 2020; Kreager et al., 2016). Parents are important socializing agents to their (adolescent) children in general and of gender norms and sexual scripts specifically (Bandura & Walters, 1963; C. Connell & Elliott, 2009; Kågesten et al., 2016; Martinez et al., 2021). Additionally, while a significant body of work exists on the sexual socialization of children by their parents (see for instance Astle et al., 2022; Astle & Anders, 2022; C. Connell & Elliott, 2009; Elliott, 2010; Evans et al., 2020; Holman & Koenig Kellas, 2018; Schaafsma, 2022; Scull et al., 2021; Solebello & Elliott, 2011; E. K. Wilson & Koo, 2010), few studies focus on parents as socializing agents of the SDS specifically (but see Axinn et al., 2011; Miller, 2012). As sexual scripts change throughout the lifecycle, and in relation to which societal roles people have (Simon & Gagnon, 1984, 1986), investigating parents of adolescents is worthwhile.

Many parents provide at least some explicit lessons on sexuality (Astle & Anders, 2022; Scull et al., 2021), and all parents provide their children with implicit lessons on sexuality in various ways, for instance by setting rules for dating (Endendijk et al., 2022), and by communicating certain expectations, beliefs, and attitudes indirectly for instance by displaying discomfort when children mention sexual topics, or by making assumptions about the sexual orientation of children (Astle & Anders, 2022; Solebello & Elliott, 2011). During communication about sex, parents focus mostly on topics related to risks, like sexually transmitted diseases, and (unwanted) pregnancies (Evans et al., 2020; Schaafsma, 2022; Scull et al., 2021; E. K. Wilson & Koo, 2010). Positive topics, like pleasure, desire, and romantic relationships are discussed less (Astle & Anders, 2022; Evans et al., 2020).

Parents have been found to show several facets of the SDS in sexual communication. Regarding sexual behavior, like consuming sexual content media and dating at a young age, parents are more lenient and

approving towards their sons than towards their daughters (Axinn et al., 2011; C. Connell & Elliott, 2009; Endendijk et al., 2022; Kuhle et al., 2015; E. K. Wilson & Koo, 2010). Parents position their daughters as less sexually agentic than their sons, with daughters being more safeguarded than sons (Elliott, 2010; Kuhle et al., 2015; Solebello & Elliott, 2011). This difference relates to daughters being seen as more vulnerable than sons; parents are more concerned about harmful sexual consequences for them (E. K. Wilson & Koo, 2010). This perception of vulnerability may explain why parents are generally found to talk more about sex with their daughters than their sons (Evans et al., 2020; Schaafsma, 2022; E. K. Wilson & Koo, 2010).

Fathers and mothers differ in their sexual communication. In general, mothers feel more comfortable and report higher levels of self-efficacy in sexual communication with their children (Scull et al., 2021). Fathers generally communicate less about sexuality with their offspring and cover a smaller range of topics than mothers (Astle et al., 2022; Evans et al., 2020; Schaafsma, 2022; Scull et al., 2021). They have been found to be more permissive of sexual behaviors for their children than mothers (Axinn et al., 2011; Smit et al., 2022). Some studies show that men endorse the SDS slightly more than women (Petersen & Hyde, 2010), while others show no difference between men and women in endorsement (Endendijk et al., 2020). Further research is therefore warranted.

Research aim

In this study we focus on the Dutch cultural context to explore contemporary sexual scripts regarding the SDS. We investigate the social context of parents talking to their adolescent child. We use vignettes describing sexual behaviors, embedded in a larger list of vignettes concerning various situations (e.g. failing a class, being rejected by a friend). This way, participants are less likely to be aware of the research aim. Moreover, it can be easier for parents and adolescents to talk about

sexual behavior indirectly. For instance, when discussing sexuality, parents and their adolescents regularly talk about others (Afifi et al., 2008). When talking about sexual behaviors of others, parents communicate expectations, beliefs, and attitudes (Astle & Anders, 2022). Talking about (hypothetical) others can therefore give novel insights into the SDS, as has been shown in research among young adult women (Fjær et al., 2015).

We aim to investigate what sexual scripts are communicated by Dutch parents regarding sexual behaviors of adolescents. To investigate whether the SDS is present within these scripts, we examine differences between reactions concerning girls and boys. Furthermore, we investigate differences between reactions given to daughters and sons, and between reactions given by fathers and mothers.

Methods

Sample

This study is part of the longitudinal project Girls in Science, which examines adolescents' gender socialization in the family and school context. This study reports on data from the first wave of the project. Families with opposite-sex parents from the Western part of the Netherlands with at least two children were recruited through municipal records. Exclusion criteria were severe physical or mental disabilities of a family member, divorced/separated families, single-parent families, families with a non-biological parent, and parents raised outside the Netherlands. Families were eligible for participation when their eldest child was approximately 17 years old and their second-born child was approximately 15 years old. Based on municipal records, 5463 families were eligible and sent an invitation for the research project via mail. In total, 1075 families responded, out of which 328 families indicated willingness to participate. In total, 73 families were excluded and 56 families were unable to participate due to scheduling issues or illness. A

total of 199 families (each consisting of a father, a mother and a second-born adolescent) participated. In each family, the second-born adolescent (mean age 15.45 years, *SD* = 0.33; 123 female) conducted a structured interview with each of their parents separately. Fathers were on average 50.87 years old (*SD* = 4.40) and mothers 48.53 (*SD* = 3.49). Most parents were highly educated, 72% of mothers and 73% of fathers obtained a bachelor degree or higher. Most families (70%) had a gross annual income exceeding €60.000.

Procedure

Families were visited at home by trained (under)graduate students. During these visits, families completed various tasks, including the structured interview. Several tasks were filmed. After giving instructions, the researcher left the room. Each participating family member received a gift certificate of €20. We obtained informed consent from each participating family member. Ethics approval was granted by the host Institution.

Instrument

To recreate a relevant dyadic social interaction, namely a parent responding to their child, the adolescent conducted a scripted interview with each parent separately. The adolescent was given a booklet containing 19 short vignettes, each followed by a question. Each vignette described a hypothetical named individual in a situation that adolescents might encounter, like social rejection, failing a class, participating in sports, and experiencing same-sex attraction. The adolescent was instructed to read out the vignettes and accompanying questions (one question per vignette) to their parent in the presented order and to not reply to the questions themselves, nor to react to their parents' answers. Parents were instructed to give one-sentence answers, and adolescents were instructed to continue swiftly to the next vignette after each answer due to time restraints. This was done to elicit implicit attitudes. However,

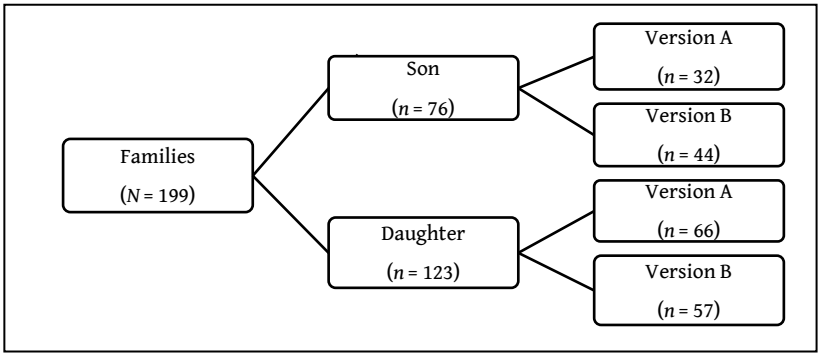
some parents gave longer answers. It was explained to the parent-adolescent dyads that the vignettes concerned adolescents of the same age as the adolescent participating in the task. In this study, we focused on two one-sentence vignettes with accompanying questions that concerned romantic relationships and kissing:

1. Thijs is in a relationship with his sixth girlfriend. What type of person do you think Thijs is?
2. Lieke kissed someone she doesn't know while going out. What type of person do you think Lieke is?

The vignettes were translated from Dutch by the first author. Thijs (male) and Lieke (female) are common Dutch names. In Vignette 1, the Dutch word *verkering* is used for relationship. This word indicates a romantic relationship, and is sometimes translated as courtship or 'going steady'. *Verkering* is commonly used when describing a romantic relationship between children or adolescents, while (young) adults more commonly use the word *relatie* (literally: relationship). The presented vignettes above represent Version A of the task. Approximately half of the families received Version A, whereas the other half received Version B, in which the gender markers and names were switched (i.e., Lieke had a relationship and Thijs kissed someone while going out (see Figure 3.1)). Fathers and mothers from the same family received the same version of the vignettes. As there were 199 families, in which a father and a mother each answered two questions, there were 796 responses in total. Exactly half of the responses (n = 398) were about a girl, the other half about a boy. Lastly, 492 responses were given to a daughter, and 304 to a son.

Figure 3.1

Number of Participants per Condition of the Task



Note. In Version A Thijs (male), is described as having a relationship and Lieke (female), is described as kissing someone she doesn't know. In Version B, the names and gender markers are switched.

Analytic Process

The videos were transcribed into text files. Notes were added when the adolescent had interrupted or otherwise intervened when the parent answered the question. Answers were then coded using ATLAS.ti (Version 9, 2009). Three of the authors (AK, TV, and LV) were involved in coding, while one author (AV) acted as a methodological advisor. We used a multi-stage approach of content analysis (Hickey & Kipping, 1996; Maykut & Morehouse, 1994; Memon et al., 2017). To ensure the trustworthiness of our research, the three coders coded parents' answers during multiple rounds of (re-)reading. After each round of coding, all codes were compared, discussed and coding-rules were recorded and refined in a codebook. In the first round, Coder AK and TV first coded 20% and in the second round they coded 40% of the data. In the third round, Coder LV coded the same 40% of data. In the fourth and final round, Coder AK and TV individually coded the entire dataset, and Coder LV a subset of 20%. All codes were compared, remaining disagreements and discrepancies were discussed until agreement was reached. We consulted formal and

informal dictionaries (e.g., websites about sex and relationships aimed at adolescents) to clarify the meaning of certain answers and to identify which words and phrases fit into specific codes. For instance, according to the dictionary, the word reveler is synonymous to libertine, and it is defined as someone who is frivolous, superficial, or loose. All responses that contained words like these, received the code *reveler.* Informal dictionaries indicate that the word *mooiboy* (literally: pretty boy) does not just refer to the way a boy looks, but also that it is synonymous to player, so it received the code *player*. When answers were unclear, videos were checked to see if intonation, irony, and body language could clarify which code to allocate. If an answer remained unclear after reviewing the video, it was coded as unclear (4 out of 796 responses). In total, 52 codes were developed, e.g., *insecure*, *adventurous*, *pleasure*. Codes were grouped into 15 categories, each consisting of one to six codes (for the full list of categories and codes, see Appendix C). This was done by writing all codes on cards and by sorting the cards according to common elements. This aided in further refining the coding rules (for more information see Maykut & Morehouse, 1994, pp. 123–131). For instance, the codes *insecure, can't be alone,* and *craving attention* were grouped into the category *dependent*, which we described as: "The adolescent is described in terms that signal dependence; often with submissive emotions like insecurity, or by describing the adolescent as needy or an attention seeker. The motivation for the behavior (appears to) come(s) from insecurity/neediness." Frequencies for each category were calculated.

In the supplemental materials, we share the (Dutch) codebook, an in-depth description of the coding process, and the anonymized raw data. All reactions presented in this paper were translated to English by the first author. Complete reactions are presented in order to maintain integrity of the data. Each reaction is preceded by the letter M or F, to indicate a mother or father, and a unique participant number. Laughter and interruptions by the child are indicated with square brackets. Following each quotation, we indicate whether this was said to a son or a daughter,

about a boy or a girl, and in reaction to Vignette 1 (sixth relationship) or 2 (kissing a stranger). For example:

F65: Eh Thijs enjoys life. (To son, about boy, V2)

Findings

Following each vignette, parents were asked the question "what kind of person do you think Thijs/Lieke is?" We discuss the categories that arose from parents' responses, which are further sorted into five themes based on their content. For instance, the categories *dependent* and *love* are discussed within the theme *dependent and romantic creatures*, alluding to the gendered expectations of girls being passive, dependent, and naturally romantic. We present the following five themes: 1) normative sexual development, 2) promiscuity and impulsivity, 3) dependent and romantic creatures, 4) social penalties and praises, 5) ambiguous and unclear answers. For each theme, we discuss whether evidence for the SDS or (other) gender differences were present. Total frequencies and percentages are presented in the text, while frequencies and percentages per group (mother/father, daughter/son, Lieke/Thijs) and per vignette are presented in Table 3.1A and Table 3.1B.

Table 3.1A

Frequencies and Percentages of Coded Categories for Vignette 2 (Kissing a Stranger) per Dyad

			Mother (n = 199)								Father (n = 199)							
			Daughter (n = 123)				Son (n = 76)				Daughter (n = 123)				Son (n = 76)			
			Girl		Boy		Girl		Boy		Girl		Boy		Girl		Boy	
Theme	Category	n	57		66		44		32		57		66		44		32	
			n	%	n	%	n	%	n	%	n	%	n	%	n	%	n	%
1	Normalization		3	*5.3*	2	*3.0*	1	*2.3*	2	*6.3*	3	*5.3*	2	*3.0*	1	*2.3*	2	*6.3*
	Age-related		0	*0.0*	2	*3.0*	3	*6.8*	1	*3.1*	0	*0.0*	1	*1.5*	2	*4.5*	0	*0.0*
	Exploration		11	*19.3*	7	*10.6*	5	*11.4*	5	*15.6*	14	*24.6*	6	*9.1*	4	*9.1*	4	*12.5*
	In search of		5	*8.8*	6	*9.1*	7	*15.9*	3	*9.4*	11	*19.3*	13	*19.7*	4	*9.1*	3	*9.4*
	Pleasure		10	*17.5*	7	*10.6*	1	*2.3*	6	*18.8*	2	*3.5*	8	*12.1*	2	*4.5*	5	*15.6*
2	Promiscuity		12	*21.1*	10	*15.2*	7	*15.9*	9	*28.1*	18	*31.6*	13	*19.7*	9	*20.5*	15	*46.9*
	Impulsivity		2	*3.5*	1	*1.5*	1	*2.3*	0	*0.0*	1	*1.8*	0	*0.0*	1	*2.3*	0	*0.0*
3	Dependent		26	*45.6*	6	*9.1*	4	*9.1*	2	*6.3*	12	*21.1*	1	*1.5*	1	*2.3*	1	*3.1*
	Love		9	*15.8*	2	*3.0*	5	*11.4*	3	*9.4*	8	*14.0*	2	*3.0*	4	*9.1*	0	*0.0*
4	Slut		5	*8.8*	1	*1.5*	3	*6.8*	1	*3.1*	6	*10.5*	1	*1.5*	3	*6.8*	0	*0.0*
	Player		2	*3.5*	17	*25.8*	0	*0.0*	9	*28.1*	2	*3.5*	16	*24.2*	1	*2.3*	12	*37.5*
	Positive		6	*10.5*	6	*9.1*	7	*15.9*	6	*18.8*	7	*12.3*	5	*7.6*	6	*13.6*	8	*25.0*
	Negative		1	*1.8*	3	*4.5*	0	*0.0*	1	*3.1*	4	*7.0*	4	*6.1*	0	*0.0*	0	*0.0*
5			9	*15.8*	8	*12.1*	2	*4.5*	3	*9.4*	15	*26.3*	6	*9.1*	2	*4.5*	2	*6.3*

Note. Vignette 1: "Thijs is in a relationship with his sixth girlfriend. What type of person do you think Thijs is?" Names and gender markers were changed for half of the sample. Themes: 1) normative sexual development, 2) promiscuity and impulsivity, 3) dependent and romantic creatures, 4) social penalties and praises, 5) ambiguous and unclear answers. Percentages are relative to their column. For example: out of 57 mothers who talked to their daughter about a girl, 3 used a normalization script, which equals 5.3 percent.

Table 3.1B

Frequencies and Percentages of Coded Categories for Vignette 2 (Kissing a Stranger) per Dyad

			Mother (n = 199)								Father (n = 199)							
			Daughter (n = 123)				Son (n = 76)				Daughter (n = 123)				Son (n = 76)			
			Girl		Boy		Girl		Boy		Girl		Boy		Girl		Boy	
Theme	Category	n	66		57		32		44		66		57		32		44	
			n	%	n	%	n	%	n	%	n	%	n	%	n	%	n	%
1	Normalization		16	24.2	18	31.6	10	31.3	7	15.9	14	21.2	14	24.6	6	14.0	9	20.5
	Age-related		10	15.2	7	12.3	5	15.6	7	15.9	3	4.5	6	10.5	0	0.0	1	2.3
	Exploration		19	28.8	22	38.6	15	46.9	11	25.0	27	40.9	25	43.9	15	34.9	8	18.2
	In search of		0	0.0	1	1.8	0	0.0	0	0.0	1	1.5	2	3.5	1	2.3	0	0.0
	Pleasure		8	12.1	12	21.1	8	25.0	8	18.2	8	12.1	6	10.5	8	18.6	4	9.1
2	Promiscuity		4	6.1	4	7.0	2	6.3	4	9.1	7	10.6	7	12.3	3	7.0	2	4.5
	Impulsivity		8	12.1	7	12.3	3	9.4	2	4.5	8	12.1	6	10.5	6	14.0	3	6.8
3	Dependent		8	12.1	3	5.3	2	6.3	0	0.0	1	1.5	2	3.5	1	2.3	0	0.0
	Love		1	1.5	1	1.8	1	3.1	2	4.5	2	3.0	1	1.8	1	2.3	1	2.3
4	Slut		1	1.5	5	8.8	4	12.5	1	2.3	6	9.1	2	3.5	1	2.3	1	2.3
	Player		0	0.0	2	3.5	0	0.0	1	2.3	0	0.0	3	5.3	1	2.3	1	2.3
	Positive		0	0.0	5	8.8	2	6.3	5	11.4	9	13.6	8	14.0	6	14.0	6	13.6
	Negative		7	10.6	3	5.3	2	6.3	3	6.8	1	1.5	11	19.3	2	4.7	2	4.5
5			21	31.8	6	10.5	9	28.1	2	4.5	8	12.1	7	12.3	5	11.6	4	9.1

Note. Vignette 2: "Lieke kissed someone she doesn't know while going out. What type of person do you think Lieke is?" Names and gender markers were changed for half of the sample. Themes: 1) normative sexual development, 2) promiscuity and impulsivity, 3) dependent and romantic creatures, 4) social penalties and praises, 5) ambiguous and unclear answers. Percentages are relative to their column. For example: out of 66 mothers who talked to their daughter about a girl, 16 used a normalization script, which equals 24.2 percent.

Theme 1: Normative sexual development

The most common theme in the responses of parents related to normative sexual behavior and development (n = 412 out of 796 responses, 52%). Normative in this context means normal, natural, and expected (Tolman & McClelland, 2011; Van De Bongardt et al., 2017; World Health Organization, 2002). Overall, we found no clear evidence for the SDS within these scripts, and only a few gender differences, which we will discuss below.

Normalization

Parents regularly used scripts that indicate normalization of sexual behavior (n = 110, 14%). For instance, parents used the word *gewoon* often. *Gewoon* translates to normal, usual, regular, or common. Additionally, it contains a normative component: it also means acceptable and right (Schalet, 2011). Parents additionally used words like standard, normal, and healthy. In other cases, parents described the situation as something that 'can just happen'.

M260: *Gewoon* a teenager, that can happen. [Laughs] (To daughter, about girl, V2)
F109: Eh, a *gewone*, healthy boy. (To daughter, about boy, V2)

One gendered script stood out. When the vignette concerned a boy, the adolescent was labeled as just 'a boy' in a few cases, without any other adjectives:

F111: That's a boy. (To daughter, about boy, V2)
M127: Well, *gewoon* a boy. (To daughter, about boy, V2)

This seems to imply that kissing a stranger is normal behavior for boys. No parent described the same behavior of girls by stating that the adolescent was 'a girl'. However, generally parents used normalizing scripts for boys and girls equally. Parents used normalizing language

mainly to describe the adolescent who kissed a stranger. In some cases, parents described the adolescent as *gewoon* drunk, implying that being drunk and kissing a stranger is normal behavior.

Another way in which normalcy of the behavior in both of the vignettes was implied, was through relating the behavior to the age of the adolescent. Parents did this by saying that the adolescents in the vignettes were young, and by calling them a *puber*. *Puber* is a Dutch word meaning a teenager or adolescent who is going through puberty. Parents used *puber* to relate the behavior to the adolescent's developmental stage:

> M164: I assume, I imagine that Thijs is still kind of a puber or at least a young adult and yes, sometimes you kiss with someone that you don't exactly know who they are. I have no judgment. (To daughter, about boy, V2)
>
> F81: I assume a young person, a young person who meets someone else, so just a normal boy, yes. (To son, about boy, V2)

Answers containing references to youngness and (being a) *puber* were mainly given in relation to kissing a stranger, and more so by mothers than fathers.

Exploration

The most common category within parents' responses of normalization, were descriptions of the adolescents as curious explorers or adventurers, or as wanting to experiment and investigate (n = 198, 25%). While some parents only used descriptors related to exploration, others explicitly added that the adolescents were finding out what they desire, or that exploration is normal, or good and pleasurable:

> F358: Yes she sometimes tries something out, and she thinks like "I should get enough experiences before I go off with

the real one", that's very sensible of Lieke. (To daughter, about girl, V1)

M122: Eh Thijs is someone who goes out on discovery, and eh is experimenting maybe, but mainly is discovering. He's a very healthy *puber*. (To son, about boy, V2)

Scripts related to exploration were mainly used in relation to the vignette about kissing a stranger. Parents also referenced the process of exploration by describing the adolescents as searching, not (yet) knowing what they like or want. These scripts were often more closely related to sexual and relational preferences and identity.

F114: Thijs he eh, can't really eh, find, and doesn't know what he eh is looking for in a girlfriend. (To daughter, about boy, V1).

M139: Eh, I think that Lieke is a little bit well searching for validation that she's all right, that she is searching for who she is. (To daughter, about girl, V1)

The kissing-vignette was formulated in such a way that the gender of the stranger with whom the adolescent kissed was unknown. While most parents did not reference sexual orientation, some parents assumed that the adolescent kissed someone of the opposite gender:

F233: Hm, a girl who dares to make a move towards boys and eh and dares to experiment with love. Or with, well, love, that's not really love of course, but eh with eh eh things that have to do with love, with attraction to the other sex. Yes. (To son, about girl, V2)

In only two cases, parents spontaneously addressed the possibility of the adolescent not being heterosexual:

F181: Eh, Lieke likes boys. Well how should I know if she kissed with a *boy*, but Lieke likes kissing. (To son, about girl, V2; emphasis added)

F221: Well Lieke is eh, searching for a steady boyfriend or girlfriend. (To daughter, about girl, V2)

Pleasure

In response to both vignettes, fathers and mothers regularly addressed pleasure and desire (n = 103, 13%). They did this mainly by describing the adolescent as someone who likes or loves activities like kissing, parties, and drinking. In other instances, they described the adolescent as someone who likes or loves (cute) boys or girls.

M71: Eh a eh popular girl who is crazy about boys. (To daughter, about girl, V1)

M76: Someone who likes a party and pleasure. (To son, about girl, V2)

F305: Well, a person who eh likes to be infatuated and to be in a relationship with many many girls. (To daughter, about boy, V1)

Descriptors like pleasure-lover, bon vivant, and party animal were used as well. Responses that received codes related to pleasure were virtually always positive. References to pleasure and desire were made in response to both vignettes, although they were made somewhat more often in response to the adolescent kissing a stranger.

Theme 2: Promiscuity and impulsivity

The term promiscuity is generally used to describe not being restricted to one sexual partner. It also signals being indiscriminate, casual, and careless. Parents regularly used scripts like these (n = 126, 16%), including words like unreliable, capricious, and unstable, or describing the adolescent as someone who gets bored quickly, or who has commitment issues. In some cases, these scripts had negative connotations, while in others they appeared to be neutral:

F62: Eh, a capricious, eh, well someone who doesn't really know what he wants. Who can't commit to someone, yet. (To son, about boy, V1)

M11: I think she easily gets tired of boys [laughs], she likes variation. (To daughter, about girl, V1)

More often, parents used words like reveler, floater, butterfly, and gallivant. These words were also used to describe promiscuous behavior, but suggest being care free rather than careless, and unsure rather than unreliable:

F89: Thijs is, Thijs is a bit of a gallivant, I think. But eh, he does, quite good eh, discover the world. (To daughter, about boy, V1)

M106: A butterfly. [Child: a butterfly?] Yes, she flutters from one to the next, like a happy-go-lucky-person. (To son, about girl, V1)

These descriptions of indiscriminate sexual behavior appear to be generally neutral or positive. Both types of scripts were used mainly to describe the adolescent engaging in successive relationships. Fathers used these scripts more often than mothers, especially when talking to their son about a boy in his sixth relationship. The indiscriminate character of promiscuity also relates to impulsivity, which was a script that was present in some of the responses (n = 49, 6%). Parents sometimes described the adolescents as impulsive or thoughtless, almost exclusively in the context of kissing a stranger. This appeared to be negative in some cases:

F163: I think that eh, in advance Lieke doesn't eh think very well about whether it is the right person for her, and then gets carried away every time. And that maybe, because of that, eh it's a disappointment again and again. (To daughter, about girl, V1)

M312: Yeah that's not very wise to kiss someone you don't know. That's a bit of a rash deed of Lieke. (To daughter, about girl, V2)

However, similar to the descriptors that relate to promiscuity, parents often used words and phrases that have more positive connotations, like being spontaneous and intuitive:

M179: Someone who trusts his gut and takes risks. (To daughter, about boy, V2)
F363: Hm, also an inquisitive person I think. Spontaneous, he thinks, "Well okay it comes on my path so I just do it." (To daughter, about boy, V2)

Fathers and mothers used scripts of impulsivity equally, and to sons and daughters alike. We found no indications for the SDS in scripts related to promiscuity and impulsivity.

Theme 3: Dependent and romantic creatures

Some parents used scripts related to being insecure, needy, wanting attention, having trouble being alone, and/or lacking agency (n = 70, 9%). These scripts fit the gendered expectations often held for women as part of the SDS.

M9: Eh, oops. I think she's a little insecure. And that she doesn't really watch out for herself and only wants to be with someone else. (To daughter, about girl, V1)
M176: Eh maybe someone who doesn't want to be alone, or maybe wants to comply with eh eh wishes, or eh, what her girlfriends do as well. (To daughter, about girl, V1)

These scripts were not used often in general, but they were used more regularly by parents who were talking about a girl to their daughter, especially in the context of the vignette describing successive

relationships. For mothers talking about a girl to their daughter, it was one of the most common scripts. This indicates the endorsement of the SDS by parents, and especially mothers, talking to their daughter.

Parents sometimes used scripts related to love (n = 43, 5%). Especially girls were described as looking for love, or falling in love (*too*) quickly, too early, or too much. These scripts were used more often context of being in the sixth relationship.

F231: Eh, yeah she is, looking for the one, her true boyfriend. (To son, about girl, V1)
M235: Someone who eh falls in love quickly and who needs that, a relationship. (To daughter, about girl, V1)

Theme 4: Social penalties and praises

An important facet of the SDS is formed by the gendered differences between social penalties and praises of sexual behavior (Reiss, 1956; Tolman et al., 2015; Zaikman & Marks, 2017). Through the use of explicit penalties and praises, the SDS became apparent in the answers of some parents (n = 224, 28%). Parents sometimes used words like 'slut' (n = 41, 5%), which is a social penalty and characteristic of the SDS (Bay-Cheng, 2015; Currier, 2013; Endendijk et al., 2022). Some parents used more euphemistic terms for slut, like easy (Armstrong et al., 2014).

F70: Geez, a slut. [laughs] (To daughter, about girl, V1)
M277: She's easy, too clingy. (To son, about girl, V2)

These scripts were used equally across vignettes and not often in general, but they were used slightly more often to describe girls than boys. In a few cases, when parents used words like slut to describe boys, it was accompanied by a caveat, e.g., "A *boy* hussy" (emphasis added). This implies that sluts and hussies are usually girls, and that this boy is an exception. So even when a boy is described as a slut, the SDS is still enforced.

One type of script that was used by parents more often in terms of social penalties and praises, involved words like player (n = 67, 8%). The word player is used to indicate a man who has multiple sexual partners, or juggles brief and sometimes overlapping intimate relationships. For responses concerning boys, this was one of the most common scripts indicating the SDS. While most parents who used this script did not add any further evaluative remarks, others questioned the motivations of the boy in the vignette:

F329: Why did those other five, why did those relationships end? Is it because he didn't find the right one yet because it was all eh, or eh, or is he a player? I don't know. (To daughter, about boy, V1)

M132: Eh, don't know, I can't judge that, I don't know how often he does that. If he does that every time, you would call him a player. (To daughter, about boy, V2)

Some other parents used negative words that are synonymous to player, but that include connotations of consuming or destructing the other, like *vrouwenverslinder* (literally: women devourer), or man-eater. Player scripts were mainly used when discussing the vignette concerning the sixth relationship. On some occasions, parents apparently referred to the SDS when using terms like slut and player:

F376: [Laughs] A player. O no, those are boys. No, I think an insecure girl. (To daughter, about girl, V1)

M148: Well with boys you don't say slut huh, I believe [Laughs]. [Child: Player] Eh. Player, oh. (To son, about boy, V2)

The sexual scripts described in the previous three sections (normative sexual development, promiscuity and impulsivity, dependent and romantic creatures) also contained various social penalties, like being described as unreliable or rash. Likewise, the scripts contained various social praises, including being called an explorer or a free spirit. Parents

sometimes used other scripts that did not fit within the previous themes that indicated either penalty or praise. Parents signaled penalties by using derogatory terms, like loser, pinhead, or fool, or describing the adolescent as inconsiderate:

> F272: Eh eh eh very, eh, a very obnoxious eh little man. (To daughter, about boy, V2).
>
> F348: O. Eh what type of eh? Eh. His sixth girlfriend. I believe he's not really eh a steadfast person. He is a little eh, should see a little more if he can make something eh of it. What type of person? Too eh, not… He has too little care for the other. (To daughter, about boy, V1)

Negative scripts were not used often (n = 44, 6%), but they were used more towards daughters, especially by fathers describing boys. This relates to the gendered expectation of daughters being vulnerable and fathers being tasked to protect them, which is part of the SDS (Kuhle et al., 2015; Solebello & Elliott, 2011). One father quite explicitly referred to this idea:

> F69: A curious person. Unless eh it's you eh who kissed with a eh. Then I think, then I think something different. [Laughs] Then I will beat that guy up. (To daughter, about girl, V2)

Adolescents were praised through the use of positive descriptors (n = 92, 12%). Many positive descriptors were used just once or twice, like sweet, flexible, creative, and confident. *Leuk* (cute, nice, pleasant) and *gezellig* (companionable, enjoyable, cozy) were used slightly more often, with no remarkable gender differences. More regularly adolescents were described as popular or social:

> F220: Sixth girlfriend? Well then Thijs is someone who, eh, by nature exerts attraction on girls, and maybe on people in general. (To son, about boy, V1)
>
> F22: Someone who likes to kiss someone, in any case very social, open. (To son, about girl, V2)

These positive scripts were used equally across vignettes, slightly more by fathers than mothers were. Both parents used slightly more positive descriptors when talking to sons than to daughters, which indicates an indirect form of the SDS.

Theme 5: Ambiguous and unclear answers

Parents regularly gave ambiguous answers or avoided answering altogether in response to both vignettes (*n* = 109, 14%). Parents did this in various ways. Firstly, and most commonly, parents regularly stated that their answer depended on unknown information, and/or shared multiple, often opposing scripts. In the context of the sixth relationship, some parents said that it mattered how long the relationships had lasted or how much time was between them. In the context of kissing a stranger, parents wondered whether this was something that the adolescent did regularly or if it was a single incident. When some parents shared opposing evaluations about girls, they used scripts related to sexual agency, and they did not do this when they spoke about boys:

> M271: Yes, someone who likes to take risks and go to the limit or something, I think. Or she is very insecure, that's also possible. (To son, about girl, V2)
>
> M330: Well that kind of depends. If she did it because she felt like doing it herself, than I think well eh, well that is someone who just eh… Someone she doesn't know? Yes. What type of person is that. Just a girl that wants to have fun. It is about the intention right, did she do it because she wanted to or because he wanted to? (To daughter, about girl, V2)

By second-guessing the agency of girls in this way, parents allude to the SDS. The second way in which answers were unclear, involved parents saying that they did not know or have an answer. Still, in most cases, they added an evaluation:

F24: Eh… phew, what type of person is Thijs, I don't know that. Thijs, Thijs, Thijs is a happy boy. Yeah well how should I know, I can't say that based on a name of someone. (To daughter, about boy, V2)

M110: Enterprising, I think, but eh I don't know actually, find it difficult. (To daughter, about girl, V2)

Lastly, parents indicated that they had no opinion or did not want to judge. Some parents elaborated on this:

M32: Well, this I really find a very hard question, what kind of person that is. Well, I really wouldn't know, there's not a right or wrong answer, it's about my answer and I can't answer this. Eh someone who eh often has different boyfriends or girlfriends, is that a type of person? No, I don't think that's a specific type, he often met the wrong person then. (To daughter, about boy, V1)

M238: A girl that kissed someone she doesn't know when she was going out. I don't think I should put Lieke in a box, if you ask me. I don't think she's 'a type of person'. A girl who goes out and kisses someone she doesn't know. (To son, about girl, V2)

Both parents gave these various kinds of ambiguous answers, more so when they talked about girls than boys, and especially when they were talking to their daughter. The gendered differences in the use of these scripts could be due to the SDS.

Discussion

In this study, we aimed to investigate contemporary sexual interpersonal scripts regarding the SDS within a sample of Dutch parents. Through thematic content analysis, we found five central themes in the sexual scripts used by parents in response to vignettes concerning adolescent sexual behavior: normative sexual development, promiscuity and

impulsivity, dependent and romantic creatures, social penalties and praises, and ambiguous and unclear answers. We will discuss the findings along three points. First, the most commonly used scripts related to normative sexual development and did not contain evidence for the SDS. Second, in the other themes indications of the SDS were present by referring directly or indirectly to gendered sexual scripts. Lastly, there were some remarkable differences between the reactions to the vignette about successive relationships and the vignette about kissing a stranger.

More than half of the responses included at least one script related to normative sexual development, indicating endorsement of adolescent sexuality as normal and right, with (sexual) exploration as a normal component. Other common scripts related to promiscuity and impulsivity. These were regularly phrased neutrally or positively, which could mean that, to some extent, parents see promiscuity and impulsivity in the context of dating and kissing as normal and perhaps age-appropriate too. These scripts did not show evidence of the SDS either. While this fits with how adolescent sexuality is defined by health organizations, researchers, and governmental institutions, this stance has often been found not to be supported by parents in earlier research, where parents appear to see adolescents as either asexual or as uncontrollable and therefore not ready for sex (Elliott, 2010; Kuhle et al., 2015; Schalet, 2011; Solebello & Elliott, 2011; Tolman & McClelland, 2011; World Health Organization, 2002). Additionally, while talking about sex in a positive way is associated with various components of sexual wellbeing, like increased sexual self-efficacy and self-esteem (Astle & Anders, 2022), past studies have found that most parents focus on negative topics (Evans et al., 2020; Schaafsma, 2022; Scull et al., 2021; E. K. Wilson & Koo, 2010). The differences between previous studies and our findings may be related to cultural context. This high level of approval and endorsement of sex-positive scripts can be explained partly by the high levels of gender equality in the Netherlands (Bosson et al., 2022; Endendijk et al., 2020). Most of the aforementioned research was conducted among U.S.

participants. European samples generally show higher levels of sex-positivity than U.S. samples (Noorman et al., 2022). This emphasizes the importance of cross-cultural research.

Endorsement of the SDS occurred in nearly one in four responses. Firstly, parents used scripts containing gendered behavioral expectations. Indirect evidence in the responses of parents can be seen in differences between answers given to daughters and sons. When talking about girls, and especially when talking to daughters, scripts related to dependency and love were often used, alluding to the idea of girls as passive, reactive, and naturally romantic creatures. This is in concordance with earlier studies (Endendijk et al., 2020; Farvid et al., 2017; Schalet, 2011; Zaikman & Marks, 2017). Mothers used these scripts more often than fathers did. Because these gendered scripts are found to apply to women in general (De Meyer et al., 2017; Endendijk et al., 2020; Zaikman & Marks, 2017), these scripts are likely more relevant to mothers than fathers, which can explain their higher usage of these scripts. Some parents used words like 'slut' and 'player', which are very characteristic of the SDS. 'Slut' was more often used to describe girls than boys, and especially 'player' was used far more often to describe boys than girls. In research, 'player' is seen as a term with positive connotations (Currier, 2013; Endendijk et al., 2020; Fjær et al., 2015; Kreager & Staff, 2009; Papp et al., 2015; Staiger, 2005; Tanenbaum, 2000; Zaikman & Marks, 2017). However, some responses indicate that parents used 'player' as a negative script. Some studies have shown that a higher number of sexual partners leads to more negative evaluations, regardless of their gender (Marks & Fraley, 2007). This begs the question whether the script of player is (still) an unequivocal label of status, especially from the mouths of parents. Some research indicates that the term 'player' has decreased in status (Staiger, 2005), but more research is needed to clarify how the meaning of 'player' has changed over the years, and how it differs between social and cultural contexts. Additionally, parents, especially fathers, more often used negative scripts when talking to their daughters, primarily when

describing a boy. This relates to the idea that girls are vulnerable and need protection from their fathers (Kuhle et al., 2015; Solebello & Elliott, 2011). Parents also used slightly more positive scripts when talking to their son than to their daughter, and fathers did more so than mothers. This fits with previous findings of gendered differences between boys' and girls' experiences of sexual communication with their parents, in which girls also received more sex-negative messages than boys (Kuhle et al., 2015; E. K. Wilson & Koo, 2010), and in which fathers were more permissive than mothers (Axinn et al., 2011). Indirect evidence of the SDS can also be seen in the ambiguous and unclear answers that were given. Parents gave more ambiguous and unclear answers when talking about girls, especially when directed to daughters, and mothers did more so than fathers. In some cases, this appeared to stem from doubts about the sexual agency of the girl in the vignette. Other reasons for ambiguous and unclear answers could stem from the idea that girls are more vulnerable and sensitive, resulting in more careful or elaborate formulations. Furthermore, it is possible that talking about sexual behaviors that do not fit gendered expectations of girls led to discomfort, resulting in less clear answers. So, while most parents do not use derogatory labels like 'slut', differential gendered expectations and evaluations may be communicated in much subtler ways, like ambiguity and

Lastly, our findings revealed substantial differences between the two vignettes. In the context of kissing a stranger, parents used scripts of normalization, age-relatedness, exploration, and pleasure far more often than in the context of a sixth relationship. This indicates that kissing a stranger fits within the idea of normative sexual development, in which it is normal and acceptable for an adolescent to kiss a stranger. Parents appeared to be more negative about adolescents who engage in successive relationships. Parents far more often connected the script of player to the vignette about relationships, but they used the script of slut equally across both vignettes. This appears to be in line with previous studies, in which it was found that being a slut is closely related to being easy and giving

guys permission for sex, and less to the actual number sexual partners (Armstrong et al., 2014; Bay-Cheng, 2015; Tolman et al., 2015). Therefore, by allowing a stranger to kiss her a girl can be seen as a slut, but a boy kissing a single stranger does not make him a player.

Strengths, limitations, and research implications

The strength of our study lies firstly in the development and use of a relatively new instrument, which had various advantages. The use of vignettes combined with open-ended questions enabled us to find which sexual scripts are used and which are most salient in Dutch parents' responses to adolescents' sexual behavior. The most salient scripts, those related to normative sexual development, are rarely used in SDS research. The behaviors described in the vignettes are not controversial or extreme, but they have been found to be age appropriate for Dutch adolescents (De Graaf et al., 2017). This increases the ecological validity of the instrument. If we had used more extreme vignettes, we might have provoked more SDS-related scripts. However, finding these scripts in the context of relatively mild and common behaviors, which are seen as appropriate and acceptable by many, shows that the SDS persists at a basal level to some extent.

Vignettes can be seen as a (relatively) implicit instrument, which avoids social desirability and awareness of the research aim (Axinn et al., 2011; Endendijk et al., 2020). Still, parents were aware that they were being filmed, resulting in possible observer effects which can cause social desirability bias in responding even though parents were not aware of the exact research goal. By embedding the vignettes in a larger list of vignettes describing various other situations, and by using different vignettes instead of the same vignette for both genders we aimed to further decrease this awareness. Additionally, by instructing parents to give quick and brief answers, they were less likely to inhibit (unconscious) biased attitudes. However, it is possible that if parents had been given the opportunity to give a longer response, additional scripts would have been

used, and that other evidence of the SDS would have surfaced. Future research could benefit from allowing parents to give a more elaborate answer. More importantly, future research should either involve a broader range of sexual scripts, including scripts of normative sexual development, or it should allow for the discovery of new scripts.

Using a scripted interview setting and focusing solely on parents enabled us to investigate a large number of responses within a relevant social interaction in a relatively large sample of families. However, this is not an authentic naturalistic interaction. The task instructions (e.g., parents were instructed to give one-sentence answers, adolescents were not allowed to respond) contributed to the setting being non-realistic. It is perhaps impossible to investigate authentic social interactions in which parents use interpersonal sexual scripts in their communication with their adolescent child, without interfering or affecting the conversation. This is especially true when the aim is to investigate a larger number of participants. Finding a balance between decreasing awareness of the research aim and approximating an authentic naturalistic setting more closely is important in future improvement on existing methods.

We chose to investigate parents of adolescents, who are important socializing agents (Bandura & Walters, 1963; Bussey & Bandura, 1999; C. Connell & Elliott, 2009; Kreager et al., 2016; Zaikman & Marks, 2017), which rarely has been done before in research focusing on the SDS (Axinn et al., 2011; Miller, 2012). Investigating the SDS within specific social settings is important, as sexual scripts change throughout the lifecycle, and differ across social contexts (Kreager et al., 2016; Kreager & Staff, 2009; Marks & Fraley, 2007; Simon & Gagnon, 1984, 1986; Wiederman, 2015). Adolescence is especially interesting when investigating the SDS, as this is the life phase in which individuals explore, develop, and refine their sexual desires, ideas, roles, and meanings (Martinez et al., 2021; Simon & Gagnon, 1984, 1986; Wiederman, 2015; World Health Organization, 2002).

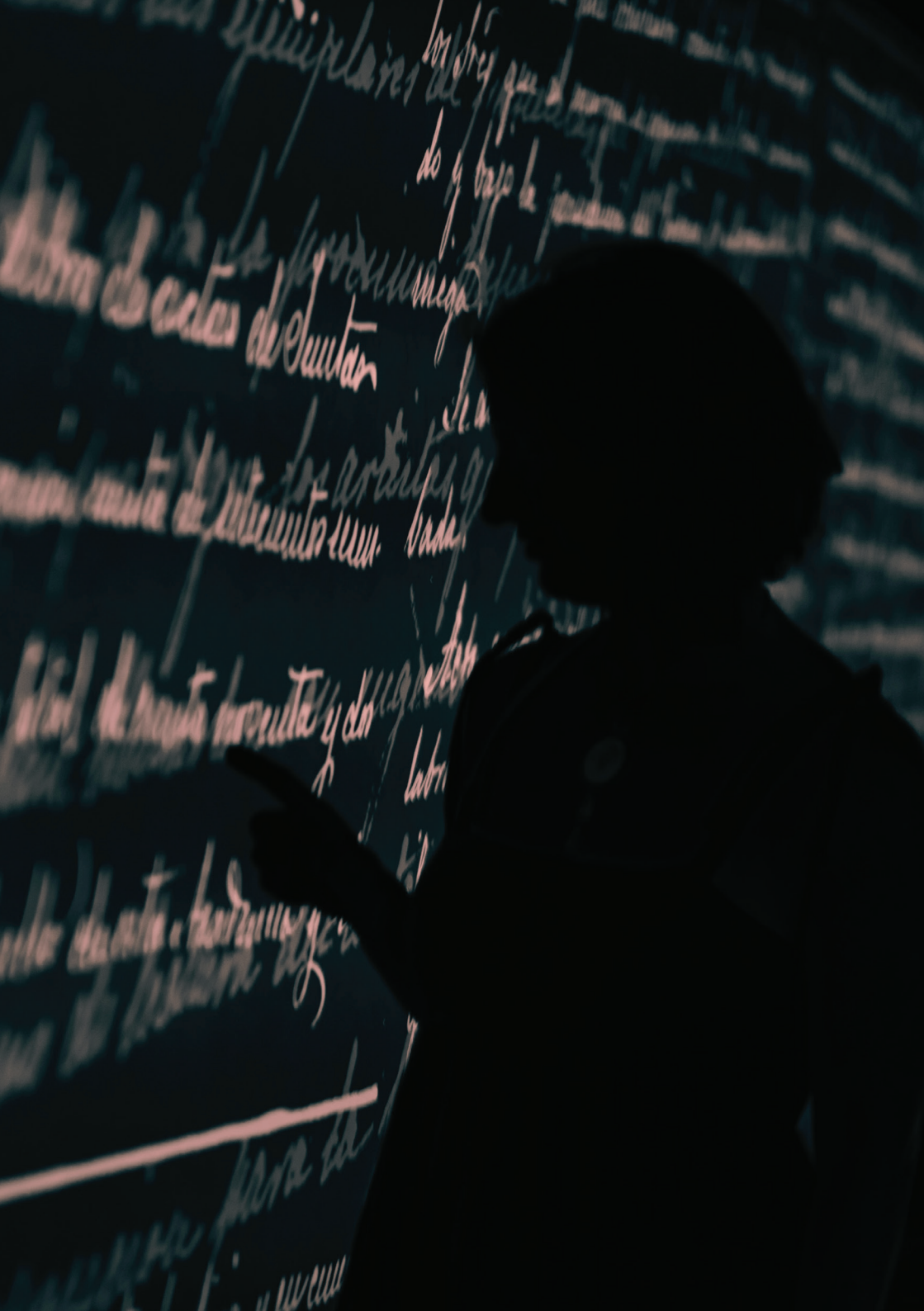
Another strength of our study lies in the focus on the specific cultural context of the Netherlands, as most studies on the SDS have been focused

on the United States (Bordini & Sperb, 2013; Endendijk et al., 2022). However, our sample was rather homogenous, consisting of people who mostly hold privileged social positions. Therefore, our findings can only be generalized to parents with a Dutch ethnic majority background, who have a relatively high socio-economic status, and who are in a heterosexual relationship with the parent of the focus-child. People from lower socio-economic backgrounds have been found to show higher endorsement of the SDS (Armstrong et al., 2014; Kreager & Staff, 2009). Ethnicity has been shown to affect both the content of gendered expectations and the endorsement of the SDS (Bordini & Sperb, 2013; Pascoe, 2007; Reiss, 1964; Staiger, 2005; Wilkins, 2012). Little is known about the endorsement of the SDS by non-heterosexual people. Future research could benefit by expanding the focus on gendered sexual expectations beyond the heteronormative SDS, for instance by using vignettes that explicitly describe non-heterosexual and/or gender nonconforming persons. Including participants from underrepresented and marginalized communities can further our knowledge about sexual scripts and gendered socialization mechanisms, as their experiences, attitudes, and scripts differ from dominant social groups, and they often face more substantial backlash for deviation from gendered expectations (Fine & McClelland, 2006; Liben, 2016).

Conclusion

Even though the parents in this study are part of a demographic that commonly shows low endorsement of the SDS in research, and the behaviors described in the vignettes is not controversial, we found the SDS in (some of) their reactions. This is evidence for the persistence of the SDS, albeit in a subtle way. As the SDS has various adverse outcomes, especially for girls (Emmerink et al., 2018; Goblet et al., 2021; Laan et al., 2021; Lefkowitz et al., 2014; Sanchez et al., 2012; World Health Organization, 2002), this topic remains important for studies on parenting, sexual development, and gender equality. Our study

contributes by exploring which sexual scripts are used by parents when talking to their adolescent child. While the SDS was endorsed by some parents, it was more common for parents to use sexual scripts related to normative sexual development. Additional research into the SDS and other sexual scripts, using a more diverse sample, is necessary.



Chapter 4

Gender and Ethnicity: Bias in Written Evaluations of Primary School Pupils

Antoinette D. A. Kroes, Marleen G. Groeneveld, Meike M. de Boer, Lotte D. Van der Pol, Judi Mesman

In preparation for publication.

Abstract

Teacher bias can affect pupils' (future) wellbeing and achievements, but research findings on the existence of bias based on pupil characteristics such as gender and ethnicity have been inconsistent. The aim of this study was to investigate whether gender and ethnicity are associated with differences between written evaluations by primary school teachers on report cards. Written evaluations were coded for valence, performance-relatedness, and language abstraction. In Study 1 (N = 203) on ethnic majority pupils, results indicated that girls were evaluated more positively than boys in relation to performance, while boys were evaluated more negatively than girls in relation to other attributes, like character traits and behavior. In Study 2 (N = 136), results indicated that ethnic majority girls were evaluated more positively than all other groups. Ethnic minority girls received more negative performance-related evaluations than other pupils did. Results are discussed in terms of their implications for the school context.

Keywords: linguistic intergroup bias, gender, ethnicity, primary school, teacher bias

Primary school pupils regularly receive evaluations from their teachers through common practices like grading, verbal feedback, and written comments. These evaluations are an expression of teachers' expectations which affect pupils' (future) achievement, socio-psychological, and behavioral outcomes, making them an important research area (Wang et al., 2018). Besides past achievement, pupil evaluations can be informed by various pupil factors, like (perceived) effort, and pupil characteristics (Rubie-Davies, 2015). When teacher expectations are based on pupil characteristics independent of their actual achievement, this indicates bias. This has for instance been found consistently for socioeconomic status (SES), with teachers having lower expectations of pupils with low SES than of those with high SES (Timmermans et al., 2018). However, findings on the relation between expectations and pupil gender and ethnicity have been inconsistent (Geven et al., 2018; Wang et al., 2018), indicating that more research is necessary. In this paper, we will investigate Dutch teachers' written pupil evaluations for possible gender and ethnicity-based bias. We will examine differences between boys and girls, and between ethnic minority and majority pupils along three components of written teacher comments. Firstly, the proportions of positive and negative remarks. Some studies show that the proportions of positive and negative evaluations are affected by pupil gender and ethnicity (Ni & Li, 2013; Rojek et al., 2019). Secondly, the proportion of remarks related and unrelated to performance. Previous research has shown that teachers use more performance-related feedback when they have high expectations of a pupil (Gentrup et al., 2020). Lastly, we investigate language abstraction, which has been found to be a good measure of implicit bias (Beukeboom, 2014; Menegatti et al., 2017; Semin & Fiedler, 1988).

Bias in teacher expectations

There is a long history of research into teacher expectations and their effect on pupil achievement and wellbeing (Wang et al., 2018). Teacher

expectations can be defined as "inferences that teachers make about the present and future academic achievement and general classroom behavior of students" (Good & Brophy, 1997, p. 79). According to this definition, teacher expectations are expressed in many different forms, and teacher bias has been found in many of these expressions, like in implicit and explicit attitudes (Glock & Klapproth, 2017; Glock & Kleen, 2017), and in teacher behaviors and school processes like verbal feedback and turn giving (Bašaragin & Savic, 2019; Denessen et al., 2020; Gentrup et al., 2020), track recommendations (Geven et al., 2018; Timmermans et al., 2018), and evaluations of pupil achievement and behavior (Shepherd, 2011; Van Ewijk, 2011). Teacher expectations are positively related to pupil outcomes like achievement, motivation, and perceived self-efficacy, and this relation is especially strong for high expectations (Gentrup et al., 2020; Rubie-Davies, 2015; Wang et al., 2018). For instance, a pupil that receives more demanding feedback gets more opportunities for optimal development (Denessen et al., 2020), this can simultaneously increase their motivation, which in turn indirectly improves performance (Jussim, 2009). There is evidence for long-term effects of teacher expectations on pupil achievement, indicating a self-fulfilling prophecy effect (De Boer et al., 2010; Gentrup et al., 2020). This is especially true for pupils from marginalized groups, as they are impacted more by the effects of teacher expectations (Wang et al., 2018). It is therefore important to stimulate high expectations in teachers (Rubie-Davies, 2015), as Weinstein, Gregory, and Strambler (2004) have stated: "all but high expectations are, by definition, inaccurate, given that achievement scores reflect what has been taught, potential is not measurable, and the goal of education is to foster academic growth" (p. 513).

Differences in teacher expectations can stem from different sources, like institutional settings, classroom and school composition, and teacher traits (Geven et al., 2018). Biases that exist in a broader societal context, like gender and ethnicity based bias, can be (unconsciously) reproduced by teachers (Myhill & Jones, 2006). Pupil demographics have therefore

often been studied in relation to teacher bias; with gender, ethnicity, and socioeconomic status (SES) being studied the most (Wang et al., 2018). While positive effects for pupil SES on teacher bias are found quite consistently, the relations between teacher bias and, respectively, pupil gender and ethnicity are less clear (Geven et al., 2018; Wang et al., 2018).

Gender bias

Findings for gender bias in teacher expectations have been inconsistent internationally (Wang et al., 2018), and in the Netherlands specifically (Geven et al., 2018). Gender stereotypes about behaviors and characteristics may constitute a basis for biased teacher expectations. On the one hand, girls are perceived to have better work habits and to be more motivated than boys (Glock & Klapproth, 2017; Myhill & Jones, 2006; Timmermans et al., 2016, 2018). On the other hand boys are seen as more disruptive, competitive, dominant, and outspoken (Frawley, 2005). The majority of teachers takes behavioral pupil characteristics into account in their evaluations (Inspectorate of Education, 2014; McMillan, 2019). This may be why girls generally receive higher grades than boys on non-standardized tests, independent from actual achievement (Voyer & Voyer, 2014). Conversely, classroom observations have shown that while boys are disciplined more, they have a higher status than girls, which is reflected in being allowed more speaking time and receiving more helpful feedback (Bašaragin & Savic, 2019; Bassi et al., 2016; Frawley, 2005).

Gender bias is likely affected by cultural differences. The Netherlands score quite high on various international indices of gender equality (Dutch Central Bureau of Statistics, 2022a), possibly causing a decrease in gender bias. For example, a longitudinal Dutch study on track recommendations between 1995 and 2014 shows a decrease in positive bias towards girls (Timmermans et al., 2018). In the final year of primary school, Dutch pupils receive a track recommendation by their teacher, which assigns pupils to one of six secondary school tracks. These tracks are based on cognitive skills, running from low level vocational education

to pre-academic education (for a more extensive explanation of the Dutch educational system, see Stevens et al., 2019). More recent data shows that gender bias in track recommendations appeared to be absent after 2014, and in the last few years (2019-2021) has actually reversed, with boys receiving higher recommendations than girls, independent from achievement (Dutch Central Bureau of Statistics, 2022a). The inconsistencies and changes over time in findings of gender bias in teacher expectations indicate the necessity of further investigation.

Ethnic bias

The ethnic/racial background of pupils may be another characteristic that leads to bias in teacher expectations. Bias could be fueled by stereotypical beliefs and attitudes about ethnic outgroups. To what extent teachers show ethnic bias has been the subject of ongoing research debates (Geven et al., 2018; Mason et al., 2014; Stevens et al., 2019). Systematic reviews indicate that most studies find negative evaluation bias against ethnic minority groups in North America and Europe when compared to the ethnic majority (Childs & Wooten, 2023; Tenenbaum & Ruck, 2007; Wang et al., 2018). In a smaller number of studies from both regions, no negative ethnic bias is reported. Several studies indicate that pupils from ethnic minority groups are not evaluated more negatively or positively on average, but that teachers are found to be less accurate in their evaluations of ethnic minority pupils than in those of ethnic majority pupils (Geven et al., 2018; Mason et al., 2014; Tenenbaum & Ruck, 2007; Wang et al., 2018). In the Netherlands, various large scale quantitative studies have shown no ethnicity-based bias against ethnic minority pupils in teachers' expectations (Stevens et al., 2011, 2019). However, smaller scale studies that used implicit measures, did find ethnicity-based bias (Van den Bergh et al., 2010; Van Ewijk, 2011), as did ethnographic research in Dutch classrooms (Weiner, 2015, 2016). This indicates that research may benefit from the investigation of a larger variety of practices, using qualitative and/or implicit measures.

In various studies that found no ethnicity based bias in teacher expectations, negative bias against ethnic minority pupils was explained by socio-economic factors (De Boer et al., 2010; Rubie-Davies & Peterson, 2016; Stevens et al., 2019). However, SES and ethnicity show high levels of overlap (Stevens et al., 2019). Ethnicity and SES indicators intersect and interact, and treating these categories as distinct and separate does not offer a true representation of social processes (Gillborn et al., 2018; Stevens et al., 2019). The overlap of SES and ethnicity is illustrated by differential effects of SES on the attainment of academic proficiency for ethnic minority pupils compared to ethnic majority pupils, which has been recorded in many countries (OECD, 2018). Attempting to disentangle indicators of SES from ethnicity and immigration status in statistical analyses, while these concepts are intertwined, overlooks socially constituted injustice (Gillborn et al., 2018; Stevens et al., 2019). In the current study, we will therefore match pupils on parents' educational level.

Research outcomes for both gender and ethnicity bias in teacher expectations are inconclusive; information on the intersection of ethnicity and gender is even less clear. Some studies have found a similar gender gap for ethnic minority pupils, with girls being evaluated more positively than boys (Farris & de Jong, 2014), while others have found the opposite, with negative bias against ethnic minority girls (Kleen & Glock, 2018). Several studies indicate that pupils from ethnic minority groups are evaluated more negatively than the ethnic majority group, regardless of gender (Glock & Klapproth, 2017; Menegatti et al., 2017). More research may provide insights into how pupil characteristics relate to teacher bias.

Written evaluations

In this paper, we will focus on written evaluations of pupils, which are an expression of teacher expectations (Wang et al., 2018). Including written evaluations (also called teacher comments, written feedback, or narrative evaluations) on report cards is common practice, especially for primary

school pupils (Hollingsworth et al., 2019). These comments can convey information that goes beyond grades, clarifying what pupils learned, the progress made, and what is left to learn (Hattie & Timperley, 2007). According to previous studies, there are several important characteristics that written evaluations should meet to stimulate development for all pupils, independent from past performance. Firstly, evaluations should mainly be aimed at performance, including specific tasks and task processes (Guskey, 2019; Hattie & Timperley, 2007). Evaluations concerning personal attributes, like "good girl" and "great effort", are unlikely to be effective, although they are more commonly given than performance-related evaluations (Hattie & Timperley, 2007). Secondly, pupils benefit from evaluations that include statements with positive and negative valence (Guskey, 2019; Hattie & Timperley, 2007; Hyland & Hyland, 2006). Evaluations with positive valence (e.g., "She is diligent"), can contribute to motivation, interest, self-efficacy, and teacher-pupil relationships. However, positive evaluations can also have negative effects. For instance, pupils can interpret praise as an expression of low teacher expectations (Hattie & Timperley, 2007). Evaluations with negative valence (e.g., "She is not able to multiply numbers over 12") are beneficial when they are constructive, related to specific tasks and task processes, indicate what needs to be improved, and providing guidance on how to improve (Guskey, 2019; Hattie & Timperley, 2007; Hyland & Hyland, 2006). However, in a qualitative analysis of written evaluations, it was found that there are often very few negative comments on report cards, the few negative comments relate mainly to behavior, and they are often formulated in very general terms, like "try harder" or "a consistent effort is needed" (Hattie & Peddie, 2003). These types of evaluations offer little guidance for pupils and their parents on how to improve on specific tasks. Moreover, pupils interpret these evaluations as impersonal, and can find them unhelpful, confusing, and discouraging (Hyland, 2013). When evaluations are not made in relation to a specific task, both positive and negative evaluations may imply that performance is not the result of the

pupils' efforts, but of a stable ability or trait (Cimpian, 2010). This can lower motivation and effort, thus affecting (future) performance (Hattie & Timperley, 2007). In summary, written evaluations should primarily be performance-related, consist of guidance for improvement, and should include positive and negative remarks for all pupils, independent from the actual performance and expectations of future achievement (Gentrup et al., 2020; Guskey, 2019; Hattie & Timperley, 2007).

Subtle forms of bias in written evaluations appear to be added by evaluators outside of awareness (Beukeboom, 2014; Rojek et al., 2019). This makes written evaluations interesting for the investigation of implicit teacher bias. Written evaluations are forms of archival and authentic data (Ni & Li, 2013), which gives them a higher ecological validity than measures that include fictional students, and (digital) implicit bias tests. However, not many studies into written evaluations on report cards exist (Hollingsworth et al., 2019; Ni & Li, 2013). In a study focusing on written evaluations of non-academic areas, teachers showed ethnic bias (Ni & Li, 2013). Compared to White pupils, Black pupils received significantly more negative comments, while Asian pupils received significantly more positive comments on behavior. Gender was not considered. To the best of our knowledge, similar quantitative research on performance-related written comments does not exist. In a qualitative study on teacher comments, it was found that lower performing pupils received less comments related to performance overall, and more related to effort, resulting in a similar overall number of positive comments (Hattie & Peddie, 2003). Effects of gender and ethnicity were not investigated. In studies involving classroom observations, verbal teacher feedback has been investigated. In one study, boys received more negative comments than girls (Bassi et al., 2016), in another study teachers gave more performance-related and more positive feedback to boys (Bašaragin & Savic, 2019). In a third study, teachers were more likely to give performance-related verbal feedback to pupils of whom they had inaccurate (i.e., not based on actual performance) high

expectations in general, and more positive performance-related feedback specifically, compared to pupils of whom they had inaccurate low expectations (Gentrup et al., 2020). Effects of pupil gender and ethnicity were not investigated. Therefore, investigating whether the proportion of performance-related teacher comments differ based on gender and ethnicity can help in determining whether bias is present.

Linguistic bias

Language, for instance in written evaluations, is a tool that regulates cognitive and motivational processes between a sender and a receiver (Semin, 2000). Cognitive processes entail the transfer of meaning, for instance, when an event is described, or instructions are given. Motivational processes entail the way in which messages are affected by wishes and desires, like the desire to portray someone as positive or negative (Beukeboom, 2014). These processes affect the words people choose to use, and take place largely outside of awareness. Hence, people's choice of words shows their implicit bias, which makes written evaluations an interesting source for bias research.

Bias in word choice on written evaluations has been found for both gender and ethnicity (Biernat et al., 2012; Rojek et al., 2019). For instance, in a study comparing written evaluations of medical students, it was found that ethnic majority students were more often than ethnic minority students described with positive competency-related words like "outstanding" and "impressive", and with positive words related to personal attributes like "mature" and "sophisticated" (Rojek et al., 2019).

Besides focusing on the meaning of words, linguistic bias can be investigated by the degree of linguistic abstraction that is used. Imagine a boy scribbling on a worksheet, and compare the following two sentences: "he scribbled on his paper" and "he is messy". The first sentence is more concrete, signaling an incident, or temporary event, while the latter is more abstract, signaling a stable, typical characteristic of the boy (Beukeboom, 2014; Maass et al., 1989; Menegatti et al., 2017; Wigboldus et

al., 2000, 2005). Language abstraction, like communication, is thought to be affected by both cognitive and motivational mechanisms (Beukeboom, 2014; Semin, 2000). The cognitive mechanism is driven by expectations; people tend to use language that is more abstract when behavior is expectancy-consistent, while expectancy-inconsistent behaviors are described at a lower level of abstraction. The motivational mechanism is formed by the desire to protect social identities, for instance by portraying a member of a marginalized group as negative, even when this is in contrast with cognition. Linguistic abstraction can be investigated with the Linguistic Category Model (LCM), which sorts verbs and adjectives in four categories with different levels of abstraction (see the method section of Study 1). Language abstraction is considered a sound implicit measure of bias, and has been used to that end in various contexts (e.g., Menegatti et al., 2017; Menegatti & Rubini, 2017; Prati et al., 2015; Schoel et al., 2014). According to cognitive mechanisms underlying the LCM, when teachers have biased expectations about girls (e.g., "girls are better behaved than boys"), they will use more abstract words when evaluating a girl who is well behaved. According to the motivational mechanism, teachers will be more likely to use abstract language for positive evaluations of the dominant social group, while using more concrete language for positive evaluations of marginalized and underrepresented groups (Beukeboom, 2014; Semin, 2000).

By investigating language abstraction, possible implicit bias in teacher evaluations can be detected, and previous inconsistencies in research findings can be informed using this method. In an Italian context, Menegatti et al. (2017) showed that girls were described with more abstract positive terms and more concrete negative terms than boys, revealing that positive attributes of girls were expected whereas negative attributes were seen as exceptional. The same was observed for students with non-migrant origins over students with migrant origins. Reversely, boys and students from migrant origins were more often described with abstract negative language and concrete positive language, implying

stable negative and situational positive characteristics. By investigating similar associations in a Dutch sample, we can both add to the existing body of knowledge on linguistic bias and teacher bias, and we can see whether cultural differences between Italy and the Netherlands arise.

Study 1

The aim of Study 1 was to investigate the relation between teacher evaluations and pupil gender. Because positive bias towards girls has been reported regularly (Wang et al., 2018), especially when concerning attributes like behavior and effort (Glock & Klapproth, 2017), and because report cards have been found to contain mainly evaluations in relation to behavior and effort, we expected that: 1.1A) girls receive more positive evaluations than boys, both related and unrelated to performance; 1.1B) girls receive more performance-related evaluations than boys, both with positive and negative valence; 1.2) girls are evaluated with more abstract positive terms and more concrete negative terms than boys.

Method

Sample

This study is part of the longitudinal project "Girls in Science", which examines adolescents' gender socialization in the family and school context. This study reports on data from the first wave of the project. Opposite-sex couples with at least two children from the Western part of the Netherlands were eligible for participation. Two groups of families were recruited, with children in two different age groups. Exclusion criteria were severe physical or mental disabilities of a family member, divorced/separated families, single-parent families, families with a non-biological parent, and parents raised outside the Netherlands. The first group had participated in a previous longitudinal project (Boys will be Boys?; see Hallers-Haalboom et al., 2017) and were selected from

municipality records when their second-born child was approximately 1 year old. At the moment of first participation, there were 1,249 eligible families, out of which 390 participated (31%). For the current study, families were invited to participate again when their second-born was approximately 10 years old. Due to dropout during the previous project, 345 families were invited. In total, 233 families were willing to participate, six of which were excluded due to divorce, emigration, or decease of a family member. Ultimately, 144 families participated (66%). Data collection for this group took place in 2020 and 2021, and had to be paused for several months due to COVID-19, likely affecting the rate of participation. The second group was recruited for the current study and were eligible when their second-born was approximately 12 years old. The maximum age difference with the first-born was 36 months. Municipality records were used to select 2,988 families who received a written invitation by mail. In total, 164 families (5%) in this group fully participated. In total across both groups, 308 families participated.

Children were eligible to participate in the current study if they were in primary school and the family could provide a report card which included written evaluations by the teacher (N = 213). Some children (n = 8) were excluded because no written remarks were codable, 1 child was excluded because the report card was from a lower grade, and 1 child was excluded because the remarks on the report card were unreadable, resulting in a final sample of 203 (104 boys). The sample contained 33 sibling-pairs. The pupils attended 155 different schools. For pupils who attended the same school, we checked whether the report cards could have been written by the same teacher. There were 5 duos of pupils whose report cards were written by the same teacher, one of these duos was a sibling pair. Removing these duos from the sample did not make any significant changes to the results, so the duos were included in the data analysis. All pupils were in sixth, seventh, or eight grade of primary school. These are the final grades of Dutch primary school; pupils are

typically aged 9-13 in these grades. Most parents received higher education (mothers: 88%, fathers: 78%).

Procedure

Each family was visited once by one or two trained graduate or undergraduate researchers. Report cards were first collected during home visits, where they were photographed by the visiting researcher. Due to COVID-19, the home visits were continued online in 2020 and subsequently report cards were collected through digital uploading by the parents of the children. Families completed several questionnaires and observational tasks before, during and after these visits. Families received gift certificates after completing the tasks. Informed consent was obtained from all participating children and their parents. Ethical approval was provided by the research ethics committee of the researchers' host institute.

Instruments

Written teacher evaluations were transcribed from pseudonymized photographs into Microsoft Excel files. The evaluations were then coded. In the coding process, separate units were identified for each written evaluation. Units were defined as (parts of) sentences containing a (implied) subject and a predicate. For instance, sentences like: "You are sporty and know a lot about exercise" were separated into two units: 1) "You are sporty", 2) "(and) know a lot about exercise", even when the subject was not repeated. Each unit was coded in three aspects: (1) performance-relatedness, (2) valence, and (3) language abstraction. Units were excluded from coding when they: (A) described a neutral situation ("You had music lessons on Tuesday"); (B) related to someone else or to multiple pupils ("The class made a nice play"); (C) characterized the teacher ("I look forward to next year"). In total, 24% of units were excluded from coding on language abstraction.

Valence was coded for each unit. Positive ("You are very good at collaborating") and negative ("She often doubts herself") units were further analyzed. Neutral units were excluded. A small number of units was excluded because they were ambivalent (< 1%; "You do not enjoy being in the spotlight very much"). The relative number of positive units was calculated as the percentage of positive units out of the total number of included units.

Performance-relatedness was a binary category. A unit received a score of 1 if the unit directly related to scholastic skills ("You are good at spelling"). If a unit related to other categories, like behavior ("You do your best") or perceived characteristics ("You are sweet"), a score of 0 was given. The relative number of performance-related units was calculated as the percentage of performance-related units out of the total number of included units.

Language abstraction was coded using the LCM (Semin & Fiedler, 1988), adapted for the current purpose in accordance with Menegatti et al. (2017) and Watson and Gallois (2002). This model consists of four categories, three of which are verb categories, and the last and most abstract category consists of adjectives. All units in which the (inferred) subject was or concerned the pupil were coded as one of the four categories. If a unit contained an *adjective* (ADJ), the unit was coded as the most abstract category ("You are an incredible boy"; "You are a creative thinker"). If a unit did not contain an adjective, the main verb was coded. The most concrete and objective verb category in the LCM is formed by *descriptive action verbs* (DAVs), which refer to observable actions with a beginning and end ("You wrote one book report"; "She giggles in class"). The next, slightly less concrete verb category are the *interpretative action verbs* (IAVs). IAVs are again actions with a beginning and end, but are slightly less objective than DAVs in the sense that they are not unambiguously observable and have a positive or negative connotation, i.e., require interpretation ("You help the teacher"; "You ridicule others"). The next and least concrete verb category, *state verbs* (SVs), do not refer to an action

but rather express an enduring emotional or cognitive state ("She loves her friends"; "He understands it"). Each unit received a score for language abstraction accordingly (DAV = 1, IAV = 2, SV = 3, ADJ = 4). An average score of language abstraction was calculated for positive and negative units, that either were related or unrelated to the student's performance, resulting in four mean scores. The possible range was 1 to 4, with a higher score representing a higher level of abstraction. Additional examples of the four linguistic categories, combined with valence and performance-relatedness, can be found in Table 4.1.

Table 4.1

Examples of Positive and Negative, Performance-Related and Performance-Unrelated Units at each Abstraction Level.

	Abstr. level	Positive	Negative
Performance-related	1 - low	You watched the videos.	You did not finish your bookmark.
	2	You anticipate your opponent.	You forget the letter 'n'.
	3	You know the times tables.	You do not understand it.
	4 - high	Your essay was very strong.	Your spelling is sloppy.
Performance-unrelated	1 - low	You raise your hand in class.	You did not bring your presentation.
	2	You show perseverance.	You can't concentrate.
	3	You enjoy school.	You dislike working together.
	4 - high	You are a funny boy.	You should have a more confident attitude.

Note. The abstraction levels relate to 4 linguistic categories: 1) Descriptive Action Verbs (DAV), 2) Interpretative Action Verbs (IAV), 3) State Verbs (SV), 4) Adjectives (ADJ). The relevant verbs and adjectives have been underlined.

The first and third author developed a codebook (in Dutch, available through the supplemental materials, see Appendix D). Subsequently four graduate students were trained and independently coded a subset of data

(15 report cards, consisting of 284 units, 5% of the total number of units). The coders first selected the units and coded the performance-relatedness (0/1) and valence (+/-) for all report cards. Agreement with the head coders was >95%. Coders then coded abstraction (1-4), for which intercoder reliability was satisfactory (Cohen's κ = .77 - .88).

Data analysis

A priori power analysis with G*power 3.1 (Faul et al., 2007) showed that with power of .80 and α set at 0.05 a total sample of at least N = 82 was required to detect the smallest effects of interest found in previous studies, i.e., an effect size of η_p^2 = .07 (Gentrup et al., 2020; Menegatti et al., 2017). Further analyses were carried out using SPSS 27.

Welch's t-tests and a split-plot ANOVA were used to investigate the effect of gender on the relative number of positive units and performance-related units. Part of the sample received solely positive units (n = 71 pupils, 40 girls), resulting in missing values for the mean abstraction score of negative units. Excluding these pupils not only decreases the power of the study, but it also excludes precisely those pupils that were evaluated most positively. Following previous studies, missing values for negative abstraction have therefore been replaced with a score of zero, which signifies a very concrete level of negative abstraction (Menegatti et al., 2017; Rubini, Moscatelli, Albarello, et al., 2007; Rubini, Moscatelli, & Palmonari, 2007). A split-plot ANOVA was used to investigate the effects of gender (boy, girl), and valence (positive, negative), on language abstraction in written evaluations. Part of the sample (n = 60, 37 boys) received only units unrelated to performance, 4 girls received only performance-related units. These scores were not replaced. As effect size we report $\widehat{\omega}_p^2$, formulas and full ANOVA tables can be found in the supplemental materials (see Appendix B and D). Marginal effects were investigated through pairwise comparisons of the estimated marginal means. SPSS 27 calculates an adjusted p-level for multiple comparisons, so that significance can be accepted when $p < .05$.

All variables were inspected for outliers, for groups $n \geq 100$ defined as values more than 3.3 *SD* above or below the mean, and for groups $n < 100$ defined as values 2.58 *SD* above or below the mean (Tabachnick & Fidell, 2007a). Outliers on main variables were winsorized, meaning they were brought closer to the distribution while maintaining the same rank. Results after winsorizing are reported.

Results

The mean proportions of positive, negative, performance-related, and performance-unrelated units can be found in Table 4.2. The effect of gender on the relative number of positive units and performance-related units was investigated. The assumption of homogeneity of variances was violated for both positive units and performance-related units, as assessed by Levene's test for equality of variances ($p < .001$, and $p = .023$ respectively), so Welch's *t*-tests were used. Girls' report cards contained a significantly higher percentage of positive units than boys report cards, mean difference = 7.81, 95% CI [3.06, 12.55], $t(180.79) = 3.24$, $p = .001$, Cohen's $d = .45$, 95% CI [0.17, 0.73]. Girls' report cards also contained a significantly higher percentage of performance-related units than boys' report cards, mean difference = 6.90, 95% CI [1.02, 12.78], $t(184.18) = 2.32$, $p = .022$, Cohen's $d = .33$, 95% CI [0.05, 0.60]. Because each unit was either positive or negative, the effect of gender for negative units was practically the same as the effect for positive units. Similarly, the effect of gender was practically the same for performance-related and performance-unrelated units.

To investigate possible interactions of gender, valence, and performance-relatedness, a split-plot ANOVA was used. No three-way interaction was found. Direct effects of valence ($F(1, 201) = 799.98$, $p < .001$, $\widehat{\omega}_p^2 = .797$), and performance-relatedness, ($F(1, 201) = 383.32$, $p < .001$, $\widehat{\omega}_p^2 = .653$) indicated that report cards contained more positive than negative units, and more performance-unrelated than performance-related units.

There was a significant two-way interaction between performance relatedness and valence, $F(1, 201) = 143.55$, $p < .001$, $\widehat{\omega}_p^2 = .413$, which is not further discussed because it is not relevant for the hypotheses of this paper. Significant two-way interactions between gender and valence ($F(1, 201) = 10.22$, $p = .002$, $\widehat{\omega}_p^2 = .043$), and between gender and performance ($F(1, 201) = 5.81$, $p = .017$, $\widehat{\omega}_p^2 = .023$), confirmed the findings from the t-tests. Pairwise comparisons of marginal effects revealed that girls received significantly more positive performance-related units than boys, $F(1, 201) = 8.36$, $p = .004$, $\widehat{\omega}_p^2 = .035$, mean difference = 7.66, 95% CI [2.44, 12.88]. Additionally, girls received significantly less performance-unrelated negative units than boys, $F(1, 201) = 10.14$, $p = .002$, $\widehat{\omega}_p^2 = .043$, mean difference = -7.09, 95% CI [-11.47, -2.70]. There was no significant difference between boys and girls for positive performance-unrelated units, nor for negative performance-related units.

Means and standard deviations of positive and negative abstraction scores are reported in Table 4.3. A split-plot ANOVA was carried out to investigate the effects of gender (boy, girl), and valence (positive, negative), on language abstraction in written evaluations. In line with previous research (Menegatti et al., 2017), a main effect for valence was found, indicating that positive units had a higher mean abstraction score than negative units ($F(1, 201) = 233.16$, $p < .001$, $\widehat{\omega}_p^2 = .534$). For our hypothesis, an interaction effect of gender and valence was relevant, but no interaction was present. There was a main effect of gender, $F(1, 201) = 3.91$, $p = .048$, $\widehat{\omega}_p^2 = .014$; indicating that units on boys' report cards were slightly more abstract on average, regardless of valence. Pairwise comparisons indicated that for negative units, the effect of gender was in the expected direction, with boys receiving more abstract negative units than girls. However, this effect was not significant ($F(1, 201) = 3.06$, $p = .082$, $\widehat{\omega}_p^2 = .010$), and the 95% CI for the mean difference of 0.37 included zero, [-.05, 0.80].

Table 4.2

Mean (Standard Deviation) Proportions for Positive and Negative Units Related and Unrelated to Performance (Study 1)

		Boys		Girls			Total	
	n	104		99			203	
Total	positive	79.85	(20.23)	87.66	(13.55)	**	83.66	(17.69)
	negative	20.12	(20.20)	12.20	(13.42)	**	16.27	(17.64)
	perf.-related	17.67	(18.28)	24.57	(23.70)	*	21.03	(21.33)
	perf.-unrelated	82.30	(18.28)	75.32	(23.63)	*	78.90	(21.30)
Positive	perf.-related	13.73	(14.65)	21.39	(22.46)	**	17.46	(19.21)
	perf.-unrelated	66.01	(25.07)	65.74	(24.75)		65.88	(24.86)
Negative	perf.-related	3.63	(6.67)	2.86	(5.42)		3.35	(6.09)
	perf.-unrelated	16.29	(18.85)	9.20	(11.90)	**	12.83	(16.20)

Note. The differences between the top four rows have been tested with Welch's *t*-tests. The differences between the bottom four rows have been tested through the estimated marginal means. Significant differences between boys and girls have been denoted with asterisks.
* $p < .050$, ** $p < .010$

Table 4.3

Mean (Standard Deviation) Abstraction Scores for Positive and Negative Units for Male and Female Pupils (Study 1)

Variable	Boys	Girls
n	104	99
Positive units	3.64 (*0.29*)	3.60 (*0.31*)
Negative units	2.13 (*1.49*)	1.76 (*1.55*)

Note. A higher score signifies a higher level of abstraction. Scores for negative units could range between 0 and 4, scores for positive units could range between 1 and 4.

Conclusion Study 1

Girls received more positive units and more performance-related units than boys did. Girls especially received more positive performance-related evaluations than boys, partially confirming Hypotheses 1.1A and 1.1B. We expected girls to also receive more performance-related negative evaluations than boys, but there was no significant difference. Additionally, we expected girls to also receive more positive performance-unrelated evaluations than boys, but this was not the case. However, boys did receive more negative performance-unrelated evaluations, which could be seen as other side of the same coin. These results indicate positive bias towards girls. We expected girls to be evaluated with more abstract positive terms and more concrete negative terms than boys, but we did not find this in the data. Instead, we found that boys were evaluated slightly more abstract in general. Hypothesis 1. 2 was therefore rejected.

Study 2

In this second study, we investigated the relation of pupil ethnicity with teacher evaluations, and explored the intersection of ethnicity and gender. We will use the term ‘ethnic majority’ to indicate White people who were born in the Netherlands, and whose parents and grandparents were born in the Netherlands. Majority refers to group size, as this is the largest ethnic group in the Netherlands. We will use the term ‘ethnic minority’ to indicate people who were not born in the Netherlands, and/or whose (grand)parents were not born in the Netherlands; following the definition used by the Dutch Central Bureau of Statistics (2016). Ethnic minority pupils do not form one homogenous ethnic group, but rather belong to many different smaller ethnic groups. However, ethnic minority pupils share the status of a (often marginalized) outgroup in relation to the dominant majority. Any person from an ethnic minority background can be seen as 'the Other' by members of the dominant ethnic group,

without distinctions regarding specific backgrounds (Jensen, 2011). This generalized otherness of ethnic minority individuals is also reflected in governmental categorization of citizens, when people are seen as either 'native' or 'non-native' (Van Schie, 2018). The simple fact of being an ethnic minority, regardless of which specific one, has been shown to be relevant for opportunities, performance, and wellbeing across different domains of education (e.g., Kleen & Glock, 2018; Menegatti et al., 2017; OECD, 2018).

While findings on the effect of pupil ethnicity on teacher bias have been inconsistent, negative bias towards ethnic minority pupils has been most common (Wang et al., 2018). Therefore, we hypothesized that: 2.1A) ethnic majority pupils would receive more positive evaluations than ethnic minority pupils, both related and unrelated to performance; 2.1B) ethnic majority pupils would receive more performance-related evaluations than ethnic minority pupils, both positive and negative; 2.2) ethnic majority pupils are evaluated with more abstract positive terms and more concrete negative terms than ethnic minority pupils. Additionally, because findings on the relation between pupil gender and ethnicity have been inconsistent (compare for instance, Farris & de Jong, 2014; Glock & Klapproth, 2017; Kleen & Glock, 2018; Menegatti et al., 2017), we explored the possible interactions between ethnicity and gender, and the main variables.

Method

Sample

For the ethnic majority group, a subsample of the Study 1 participants was used (n = 92, 47 boys), including 15 sibling-pairs. Parents of children from ethnic minority groups were recruited for participation through social workers, (weekend) schools, social media, and the personal network of the researchers. To participate, parents were asked to confirm that their child belonged to an ethnic minority group (the exact questions and definitions used can be found in the questionnaire in the supplemental materials).

Parents were eligible to participate if they had a child of whom they could share a report card that was written during sixth, seventh, or eight grade of primary school. The report card had to be written during the same period as the ethnic majority group (2016-2022). Out of 81 parents who indicated to be willing to participate, 51 were able to send a report card and fill in the questionnaire. Some participants were excluded because the report card was from a lower grade (n = 4), and some because none of the written remarks were codable for language abstraction (n = 3). This ultimately resulted in a sample of 44 (24 boys). The total sample size of Study 2 was 136. In the ethnic minority sample, adoption was not an exclusion criterion; 2 of the pupils were adopted. This sample contained 3 sibling-pairs. The education level of the parents in the ethnic minority sample was slightly higher than the distribution in the Dutch population, with 12% of parents having received low, 58% medium, and 30% higher-level education. The ethnic majority sample was selected based on the educational level of mothers and fathers, matched as closely to the ethnic minority sample as possible. Differences were assessed with t-tests. There were no significant differences between the two samples for parental education level ($t(58.20) = 1.63$, $p = .109$). The entire sample of pupils attended 111 different schools. For pupils who attended the same school, we checked whether the report cards could have been written by the same teacher. This was the case for one pair of (nonrelated) ethnic majority pupils. Excluding one of these pupils from the analyses did not result in any significant differences, so results are reported including both pupils.

Procedure

The data of ethnic minority pupils was collected through a short online questionnaire in Dutch. Some questionnaires were administered on paper or by telephone, for parents who were unable to complete the questionnaire online due to language or digital barriers (n = 5). Report cards were photographed and sent digitally. Participants received a small gift after participation was completed. Ethical approval for this research

was provided by the research ethics committee of the researchers' host institute. As the ethnic majority pupils were a subsample of Study 1, the procedure for them was identical as the one described there.

Instruments

Written teacher evaluations were coded for valence, performance-relatedness, and language abstraction as described in Study 1. As names were deleted from the text, coders were unaware of ethnicity while coding.

One of the report cards included a small portion written in Frisian, which was translated by a native speaker. Part of the sample received solely positive units (n = 40), resulting in missing values for the mean abstraction score of negative units. In the ethnic minority group 8 pupils did not receive negative remarks (4 girls), and in the ethnic majority group 32 pupils (19 girls). More pupils in the ethnic majority group received only positive units than pupils in the ethnic minority group (assessed with Welch's t-test; $t(102.22) = 2.15$, mean difference 17%, CI 95% [0.1, 31.9], $p = .034$, Cohen's $d = .37$, 95% CI [0.01, 0.73]).

Ethnicity was a binary variable (i.e., belonging to the ethnic majority or not). Ethnicity was assigned by self-determination, as well as by the birth country of the pupil, their parents, and grandparents. Following the governmental definition of having a migration background at the time of the study, pupils were included if they or at least one of their (grand)parents was born outside of the Netherlands (Dutch Central Bureau of Statistics, 2016; Van Schie, 2018). There were 6 pupils with a first-generation migration background (i.e., born outside the Netherlands), 31 with a second, and 5 with a third-generation migration background; 2 pupils were adopted. Countries and regions of origin apart from the Netherlands were Suriname (n=13), Morocco (n=7), Turkey (n=6), South Asia (n=6), Middle East (n=5), East Asia (n=5), Europe (n=4), Southeast Asia (n=3), North America (n=1), Lesser Antilles (n=1), and West Africa (n=1).

Data Analysis

A priori power analysis with G*power 3.1 (Faul et al., 2007) showed that with power of .80 and α set at 0.05 a total sample of at least $N = 72$ was required to detect the smallest effects of interest found in previous studies, i.e., $\eta_p^2 = .04$ (Gentrup et al., 2020; Menegatti et al., 2017). Further analyses were carried out using SPSS 27. The effects of ethnicity and gender on the main variables were investigated with between-group and split-plot ANOVAs. Full ANOVA tables can be found in the supplementary materials. Marginal effects were investigated through pairwise comparisons of the estimated marginal means. SPSS 27 calculates an adjusted *p*-level for multiple comparisons, so that significance can be accepted when $p < .05$ and no further adjustments have to be made to the *p*-value.

All variables were inspected for outliers, for groups $n < 100$ defined as values more than 2.58 *SD* above or below the mean (Tabachnick & Fidell, 2007a). Outliers on main variables were winsorized, meaning they were brought closer to the distribution while maintaining the same rank. Results after winsorizing are reported. According to the distinction between "western" and "non-western" migrant origins (Roeleveld et al., 2011; Van Schie, 2018), 5 ethnic minority pupils (3 girls) had a 'western' background. Out of these pupils, 4 did not receive any negative remarks on their report cards (2 girls). This group is too small to be included as a separate level for ethnicity in the ANOVAs. However, we did run all the analyses with these 5 pupils excluded, yielding similar results. Full results excluding "western" ethnic minority pupils can be found in the supplemental materials.

Results

The effects of ethnicity and gender on the proportion of positive units were investigated with a between-group ANOVA. Means and standard deviations are reported in Table 4.4. There was a main effect of ethnicity

Table 4.4

Mean (SD) Proportions of Positive and Performance-Related Units on Report Cards of Male and Female Pupils from Ethnic Minority and Ethnic Majority Backgrounds (Study 2)

		Ethnic minority						Ethnic majority					
		Boys		Girls		Total		Boys		Girls		Total	
	n	24		20		44		47		45		92	
Total	positive	73.39	(20.35)	73.04	(21.56)	73.23	(20.66)	78.04	(20.92)	88.79	(13.02)*	83.30	(18.23)
	negative	26.61	(20.35)	26.75	(21.55)	26.68	(20.66)	21.95	(20.92)	10.96	(12.91)	16.58	(18.23)
	performance-related	19.61	(16.97)	25.22	(19.91)	22.16	(18.36)	17.61	(16.58)	23.19	(23.39)	20.40	(20.39)
	performance-unrelated	80.39	(16.97)	74.57	(19.78)	77.74	(18.32)	82.28	(16.87)	76.03	(24.83)	79.22	(21.26)
Positive	performance-related	14.95	(12.90)	16.17	(15.37)	15.50	(13.92)	14.55	(14.83)	20.86	(22.82)	17.64	(19.31)
	performance-unrelated	58.42	(25.45)	56.87	(21.26)	57.71	(23.39)	63.50	(25.76)	67.35	(25.31)	65.38	(25.48)
Negative	performance-related	4.61	(7.56)	9.05	(11.24)*	6.63	(9.56)	2.98	(5.64)	2.02	(3.81)	2.51	(4.83)
	performance-unrelated	21.90	(18.29)	17.70	(18.02)	19.99	(18.08)	18.75	(19.88)	8.62	(10.77)*	13.80	(16.78)

Note. Values marked with * differ significantly from the other values in the same row.

on the total number of positive units, $F(1, 132) = 8.86$, $p = .003$, $\hat{\omega}_p^2 = .055$, with ethnic majority pupils receiving more positive units than ethnic minority pupils, mean difference = 10.21, 95% CI = [3.42, 16.99]. Pairwise comparisons revealed that ethnic majority girls received more positive units than all other groups ($p < .010$, $\hat{\omega}_p^2$ ranging between .046 and .061, and mean differences ranging between 10.74 and 15.75). There were no differences between the other groups. This indicates an interaction between gender and ethnicity, but this interaction was not significant ($F(1, 132) = 2.62$, $p = .108$, $\hat{\omega}_p^2 = .012$). As the effect size could be interpreted as a small effect (Kirk, 1996), non-significance could be due to lack of power. The effects of ethnicity and gender on the proportion of performance-related units was investigated with a between-groups ANOVA as well, there were no significant effects.

To further investigate possible interactions between ethnicity, gender, valence, and performance-relatedness of the written evaluations, a four-way split-plot ANOVA was used. There were no significant four-way or three-way interactions. Like in Study 1, there were significant main effects of valence ($F(1, 132) = 279.55$, $p < .001$, $\hat{\omega}_p^2 = .676$) and performance ($F(1, 132) = 244.61$, $p < .001$, $\hat{\omega}_p^2 = .646$), indicating that report cards contained more positive than negative units, and more performance-unrelated than performance-related units. There was a significant two-way interaction between performance relatedness and valence, $F(1, 132) = 143.55$, $p < .001$, $\hat{\omega}_p^2 = .413$, which is not further discussed because it is not relevant for the hypotheses of this paper.

There was a significant two-way interaction between valence and ethnicity, $F(1, 132) = 9.01$, $p = .003$, $\hat{\omega}_p^2 = .057$, confirming the results from the between-group ANOVA. Pairwise comparisons revealed that ethnic minority girls received significantly more negative performance-related units than all other groups ($p < .05$, mean differences ranging between 4.44 and 7.03). These are negative remarks that concern performance (e.g., "Your fine motor skills are weak"). Additionally, ethnic majority girls received less negative performance-unrelated units than all other groups

(p < .05, mean differences ranging between 9.08 and 13.28). These are negative remarks that concern mainly personal attributes (e.g., "You are a messy kid"). No other marginal effects were found.

A split-plot ANOVA was carried out for each model to investigate the effects of ethnicity (minority, majority), gender (boy, girl), and valence (positive, negative), on language abstraction in written evaluations. Means and standard deviations of positive and negative abstraction scores are reported in Table 4.5.

Table 4.5

Mean (SD) Abstraction Scores for Positive and Negative Units for Male and Female Pupils from Ethnic Minority and Ethnic Majority Backgrounds (Study 2)

	Ethnic minority		Ethnic majority	
Variable	Boys	Girls	Boys	Girls
N	24	20	47	45
Positive units	3.62 (0.38)	3.59 (0.24)	3.58 (0.33)	3.56 (0.33)
Negative units	2.74 (1.31)	2.35 (1.23)	2.15 (1.45)	1.60 (1.48)

Note. A higher score signifies a higher level of abstraction. Scores for negative units could range between 0 and 4, scores for positive units could range between 1 and 4.

There was a main effect of valence, indicating that positive units had a higher mean abstraction score than negative units in general, $F(1, 132) = 101.55$, $p < .001$, $\widehat{\omega}_p^2 = .287$. There was a significant main effect for ethnicity, $F(1, 132) = 7.54$, $p = .007$, $\widehat{\omega}_p^2 = .047$, which was qualified by the significant two-way interaction between ethnicity and valence, $F(1, 132) = 5.37$, $p = .022$, $\widehat{\omega}_p^2 = .017$. There was a significant simple main effect of ethnicity for negative units, but not for positive units. Negative units on report cards of ethnic minority pupils were more abstract than negative units on report cards of ethnic majority pupils ($F(1, 132) = 6.71$, $p = .011$, $\widehat{\omega}_p^2 = .040$, mean difference = 0.67, 95% CI [0.16, 1.18]). Pairwise comparisons revealed that the negative units received by ethnic majority girls were significantly

more concrete than those received by ethnic minority boys, $p = .002$, and ethnic minority girls, $p = .048$. Additionally, they appeared to be more concrete than those received by ethnic majority boys, but this difference was not significant, $p = .059$. There was no difference between ethnic minority girls and ethnic majority boys, nor between ethnic minority boys and ethnic majority boys. There were no significant effects for positive units.

Conclusion Study 2

The report cards of ethnic majority girls contained more positive units than those of all other groups, which indicates a positive bias towards them. This effect was present overall, but not at the marginal level. This indicates that the difference in positive units is not specifically driven by performance-related or -unrelated evaluations. Thus, Hypothesis 2.1A is partially accepted. Report cards of ethnic majority girls contained less performance-unrelated negative units than all other groups. Additionally, report cards of ethnic majority pupils more often contained solely positive units than report cards of ethnic minority pupils.

There was no difference in the number of performance-related units in general. Report cards of ethnic minority girls contained more performance-related negative units than all other groups. This result was unexpected, Hypothesis 2.1B is rejected.

We expected that ethnic minority pupils would be evaluated with more concrete positive terms, and more abstract negative terms than ethnic majority pupils. We found that negative units on ethnic minority pupils' report cards are indeed more abstract than those on ethnic majority pupils' report cards. Pairwise comparisons revealed that this difference was only significant when ethnic minority pupils were compared to ethnic majority girls. Hypothesis 2.2 is partially accepted.

General discussion

In the present studies we investigated teacher bias through the valence, performance-relatedness, and language abstraction of written pupil evaluations. Generally, patterns of teacher bias driven by pupil gender and ethnicity have indicated that girls and ethnic majority pupils are advantaged by positive bias, but these findings have not been consistent across different studies (Geven et al., 2018; Wang et al., 2018). Written evaluations have been found to be a useful but underused source to investigate intergroup bias (Beukeboom, 2014; Menegatti et al., 2017; Ni & Li, 2013). Our findings indicate positive bias towards ethnic majority girls.

Valence and performance-relatedness

According to previous studies, holding high expectations is beneficial for all pupils, but especially those from marginalized groups (Rubie-Davies, 2015; Wang et al., 2018; Weinstein et al., 2004). Effective evaluations should include both positive and negative comments, focusing on performance and offering clear guidance on how pupils can improve their achievements in the future (Guskey, 2019; Hattie & Timperley, 2007; Hyland & Hyland, 2006). If all report cards would satisfy these ideas, it could be argued that we should not find differences between pupils in the proportions of positive/negative and performance-related/unrelated remarks. However, this was not what we found.

In the current studies, we found that ethnic majority girls received more positive evaluations than all other groups. In Study 1, the difference between (ethnic majority) boys and girls for the proportion of positive statements was explained by the performance-related statements. In Study 2, ethnic majority girls received more positive evaluations overall, but not at the performance-related (e.g., "Your project had a logical structure") and -unrelated (e.g., "You are a clever gal") level specifically. While ethnic majority girls received both more positive performance-related and performance-unrelated evaluations than all other groups, these differences were not significant. This is likely due to lack of power.

Additionally, we found that ethnic majority girls received the least negative performance-unrelated evaluations (e.g., "You should wait your turn"). This fits the "good girls, bad boys" stereotype, which encompasses the idea that girls are better behaved, more motivated, and have better work habits than boys (Glock & Klapproth, 2017; Glock & Kleen, 2017; Myhill & Jones, 2006; Timmermans et al., 2016, 2018; Voyer & Voyer, 2014). Overall, positive bias towards girls in teacher expectations has been found regularly in the past (Myhill & Jones, 2006; Voyer & Voyer, 2014). Our findings suggest that positive gender-based bias only works for ethnic majority girls, which is consistent with some other studies (Glock & Klapproth, 2017; Menegatti et al., 2017), but not with studies in which ethnic minority girls were evaluated more positively than ethnic minority boys (Farris & de Jong, 2014), or vice versa (Kleen & Glock, 2018).

We found that report cards of ethnic minority girls contained more negative performance-related statements than the report cards of other pupils. This could be interpreted in two ways. Firstly, we could interpret this as an indicator of positive bias towards ethnic minority girls compared to other groups, as teachers tend to give more performance-related feedback to pupils of whom they have higher expectations (Gentrup et al., 2020), and negative feedback can be very beneficial for improving performance (Hattie & Timperley, 2007). However, as performance-related feedback is not given often on the report cards in our studies, and negative performance-related feedback is especially rare, it appears that the way Dutch teachers write evaluations does not concur with the guidelines that stem from research (Hattie & Peddie, 2003; Hattie & Timperley, 2007; Hyland, 2013; Hyland & Hyland, 2006). In that sense, the higher number of negative performance-related comments on the report cards of ethnic minority girls can be an indicator of bias against them.

Language abstraction

According to the Linguistic Category Model (LCM), a higher level of abstraction implies stable and (stereo)typical characteristics, suiting expectations held by the person who wrote the message, while a lower level of abstraction implies unexpected, situational, and temporary events (Beukeboom, 2014; Maass et al., 1989; Wigboldus et al., 2000). Based on the LCM, we expected that girls would be evaluated with more abstract positive terms and more concrete negative terms than boys; and that ethnic minority pupils would be evaluated with more concrete positive terms, and more abstract negative terms than ethnic majority pupils (cf. Menegatti et al., 2017). In Study 2, ethnic minority pupils were evaluated with more abstract negative terms than ethnic majority girls were. This linguistic difference can be driven by ethnicity-based biased expectations through cognitive linguistic processes that take place outside of teachers' awareness, and by motivational linguistic processes, because teachers may be implicitly motivated to retain ethnicity-based social structures (Beukeboom, 2014; Semin, 2000). We did not find any effect on statements with positive valence, nor did we find any gender effects. While the mean abstraction scores did indicate that ethnic majority girls were evaluated with more concrete negative terms than ethnic majority boys were, this difference was not significant. There are several possible explanations for our unexpected results.

Firstly, the differences could be explained by culturally specific gender bias and gender stereotypes. For example, Italians have been found to show a somewhat higher endorsement of some gender stereotypes, and a somewhat lower level of agreement with gender equality norms than the Dutch (Halman et al., 2022).

Secondly, it appears that Dutch teachers used more adjectives than Italian teachers, especially for positive evaluations. This is reflected in the very high mean abstraction score in our studies, which is around 3.60. In the studies by Menegatti et al. (2017), the mean abstraction score for positive evaluations is considerably lower, around 2.90. While it is known

that there are cultural differences in the use of adjectives and verbs, these differences are generally small between speakers/writers from Western countries like Italy and the Netherlands (Maass et al., 2006). It would be interesting to examine why Dutch teachers are more inclined to use adjectives when writing evaluations.

Lastly, the mean abstraction scores are strongly affected by the number of report cards that did not contain any statements of positive or negative valence. Following previous studies of language abstraction, when a pupil did not receive any statements in either category, they received a score of zero, indicating a very concrete score (Menegatti et al., 2017; Rubini, Moscatelli, Albarello, et al., 2007; Rubini, Moscatelli, & Palmonari, 2007). In line with the LCM, this means that a pupil who received no negative statements receives a negative abstraction score of zero, signaling that negative evaluations are unexpected. In our study, a considerable number of pupils received no negative statements, and this was true for significantly more ethnic majority pupils than ethnic minority pupils. In the Italian study, this pattern was present too, but additionally there were pupils who received no positive statements, and these were almost exclusively ethnic minority pupils, and ethnic minority girls more so than boys were. Furthermore, the differences in missing values between groups were far larger in the study by Menegatti et al. (2017), artificially increasing the effect of gender and ethnicity on language abstraction.

Implications for the school context

Effective evaluations focus on performance and include both positive and negative comments (Guskey, 2019; Hattie & Timperley, 2007; Hyland & Hyland, 2006). However, we found that in general, over 80% of the remarks on the report cards were positive, and around 80% of the remarks were unrelated to performance. The large number of positive statements, especially unrelated to performance, is not in the best interest of pupils. Many of these statements concerned personal attributes, a type of praise

that is not beneficial for performance (Hattie & Timperley, 2007). Both positive and negative statements on personal attributes, especially when they are made up of generic language, may foster the idea that performance and behavior are mainly related to stable traits and characteristics, and not to the effort made by pupils (Cimpian, 2010). This is especially true when these statements are made at the highest level of abstraction (Menegatti et al., 2017), which was generally the case in our studies. The notion that performance and behavior are mainly related to stable attributes can have adverse effects on the self-esteem and motivation of pupils. It is therefore important that teachers learn about these effects of language. While it is hard to inhibit linguistic bias in direct verbal interaction (Beukeboom, 2014), teachers can be trained to use more concrete language when writing evaluations (Menegatti et al., 2017). Moreover, the most effective evaluations contain both positive and especially negative evaluations, providing pupils with specific information on what they should improve and how they can improve it (Guskey, 2019; Hattie & Timperley, 2007).

The absence of performance-related statements and negative statements on many report cards indicates that perhaps teachers are not effectively prepared during teacher training to write effective evaluations. Additionally, the proportions of valence and performance-relatedness pose normative questions about what goals teachers have with their evaluations. Research shows that there are gaps in our knowledge about how teachers form their evaluations, and about teachers' beliefs and perceptions of what should be included in effective evaluations (McMillan, 2019). Research into teachers' evaluative processes could clarify how they employ written evaluations and what types of messages they aim to convey. Several studies indicate that pupils and parents regularly find that the standards to which the child is measured are unclear, and that teacher comments are not understood (Hattie & Peddie, 2003; Hyland, 2013; Tuten, 2007). When evaluative standards are vague, this can increase the effect of teacher bias (Quinn,

2020). Educational research, schools, teachers, pupils, and parents could benefit from discussions about how evaluations should be formed and what they mean (Hattie & Peddie, 2003; McMillan, 2019).

Besides improvements upon evaluative processes, gender- and ethnicity-based bias can be combatted through various types of interventions. For instance, bias has been shown to decrease through intergroup contact, as intergroup contact increases knowledge and empathy towards members of other groups (Martin et al., 2017; Pettigrew et al., 2011). Additionally, intergroup contact leads to seeing others more as individuals than as a homogenous group. Some research indicates that this also decreases linguistic bias (Prati et al., 2015).

Limitations and future directions

Our studies and analyses had some limitations. It was impossible to investigate the relation of valence, performance-relatedness, and language abstraction with actual performance. This is in part due to the Dutch system, in which schools are allowed to choose between a variety of standardized tests from multiple commercial providers, schools are free to decide how many standardized tests they want to conduct (Ministerie van Onderwijs, Cultuur en Wetenschap, 2014). The only mandatory test is a school leavers' test, which is made in the final year of primary school, and results in a track recommendation for secondary school. Research has shown that the test scores across test providers are incomparable (Van Baars, 2022). Additionally, not all schools share the standardized test results on pupils' report cards. While we did have data of the school leavers' test for some pupils in Study 1 (n = 68), these results came from five different test providers. For the most popular standardized test (n = 40), we found no correlations between test score and the main variables (valence, performance-relatedness, and language abstraction, p > .450), indicating that differences between written evaluations were not the result of differences in performance. In future studies, it would be beneficial either to conduct a standardized cognitive

test among participating pupils, or to recruit report cards that include results on standardized tests by the same provider.

The sample was too small to find differences between different ethnic minority groups, or differences between pupils with a first-, second-, or third generation immigration background. Although ethnic minority groups are connected through their marginalized status, this may mask bias that is specific to certain ethnic minority groups (Gillborn et al., 2018). We did run the analyses excluding "western" ethnic minority pupils (e.g., European, North American), yielding similar results, but with larger effect sizes for the proportions of positive and performance-related evaluations (which can be found in the supplemental materials). These results indicate that teachers may not (only) differentiate between pupils based on majority-minority status, but also based on specific ethnic minority background. In the future, it may be worthwhile to employ a different recruiting strategy, for instance by recruiting through schools instead of through families. This could result in a larger and more representative sample.

We did not investigate the content of evaluations beyond their valence and performance-relatedness. This means we can only draw limited conclusions about what our results mean. In Study 2, ethnic minority girls received the highest number of negative performance-related remarks. This type of evaluation can be very useful for future achievement, but this is only true when the evaluations are clear and specific. However, studies have shown that teacher comments are usually formulated in vague and general terms, thus not benefiting the pupil (Hattie & Peddie, 2003; Hyland, 2013). Content analysis would be necessary to evaluate whether the negative performance-related statements written by the teachers were useful. Additionally, previous studies into written evaluations have shown that bias is present at the content level, through differential word use based on gender and ethnicity (Biernat et al., 2012; Rojek et al., 2019).

In future research, content analysis could be employed to investigate what types of performance-related and performance-unrelated aspects

are discussed exactly, and how these aspects are associated with gender and ethnicity. Teachers' perceptions of pupil behavior (Mason et al., 2014; Myhill & Jones, 2006) and reactions to pupil behavior (Bašaragin & Savic, 2019; Frawley, 2005) are known to show bias. Combining research on written evaluations with observational studies can help disentangle what parts of evaluations are based on bias and what parts are based on actual pupil behavior.

Conclusion

By using a combination of methods to study bias through linguistic habits of teachers, we have gained important new insights into how gender- and ethnicity-based bias is present in teacher evaluations. As ethnic majority girls were evaluated more positively in written teacher evaluations, our study indicates that boys and ethnic minority pupils are disadvantaged. Teacher bias can have detrimental as well as uplifting effects on many aspects of the (future) functioning of pupils. Our study emphasizes the need for continuous attention to the fair treatment of pupils.

Chapter 5

Big News Stories and Longitudinal Data Collection

A Prominent Child Sexual Abuse Case Negatively Affects Parents' Attitudes Towards Male Caregivers

Antoinette D. A. Kroes, Lotte D. Van der Pol, Marleen G. Groeneveld, Judi Mesman

Published in *International Journal of Behavioral Development*, 2021, *45*(6), 561-568, https://doi.org/10.1177/01650254211042418

Abstract

Consumption of news media can influence attitudes towards specific groups, but the influence of news media on longitudinal data collection has not yet been researched. We present a method to index media attention on a specific topic, as well as a case study on a big Child Sexual Abuse (CSA) story and its effect on parents' attitudes towards male childcare professionals in a longitudinal study with fathers and mothers of 207 Dutch families. Questionnaire data on attitudes toward gender-differentiated parenting were collected in four annual waves between 2010 and 2014. NexisUni® Academic database was used to index articles on CSA to chart patterns of media attention before and during that time span. There was an immediate increase in media attention, the amount of articles on CSA doubled, as well as a prolonged increase in attention which culminated during the second wave of the study. In the first wave 97 of the families participated before the CSA case became known, and 110 participated afterwards. Parents who participated after the first news about the case came out reported a more negative attitude towards hiring a male babysitter than those who participated before it. This effect was stronger for mothers. The negative effect on attitude endured during the subsequent waves for all fathers and for those mothers who participated before the news broke. Findings indicate that big news stories influence attitudes that lasts over time and can therefore influence longitudinal data. Further analysis suggests that the influence of news stories is gendered, as mothers showed a recovery in their attitudes over time, while fathers did not. We recommend further research on the effect of news on attitude and behavioral measures in longitudinal research.

Keywords: news media influence; longitudinal study; child sexual abuse; fear of pedophilia

Most people spend a significant and increasing amount of time consuming news (Wonneberger et al., 2012), which influences their behaviors (Slater, 2007) and can reinforce negative implicit and explicit attitudes towards specific groups (Arendt & Northup, 2015). A possible explanation is overrepresentation of crime in the news (Prieto Curiel et al., 2020; Windzio & Kleimann, 2009). Research suggests that the consumption of news is related to an increase in fear of crime and significantly elevates the perceived risk of crime (Callanan, 2012; Intravia et al., 2017). This effect can be crime-specific, like fear of terrorism (Nellis & Savage, 2012) and kidnapping (B. J. Wilson et al., 2005); and has been found to be stronger for women than for men (Callanan, 2012; Intravia et al., 2017; B. J. Wilson et al., 2005). Fear of crime is negatively associated with level of education (Callanan, 2012; B. J. Wilson et al., 2005). In recent years, online news consumption has greatly increased. Online news consumption shows a stronger relation with fear of crime than other forms of media (Intravia et al., 2017), as well as a stronger relation with non-neutral attitudes towards specific groups of people (Mertens et al., 2019). It could be theorized that big news stories influence longitudinal attitude research. In this paper, we present a time-efficient screening-method to gain insight into patterns of national media attention that a certain story or topic has received in a certain period and a case study on the impact of a big child sexual abuse (CSA) news story on parents' attitudes about male professional caregivers measured in a Dutch longitudinal research project.

News media and longitudinal research

News media can have an effect on attitudes and social-emotional behaviors over a longer period of time (Arendt & Northup, 2015; Slater, 2007). As attitudes and social-emotional behaviors are often topics of research in the social sciences, news coverage may inadvertently influence the outcomes of longitudinal research. This could subsequently be wrongfully attributed to other constructs. To the best of our

knowledge, there are no studies on how news media inadvertently influence the outcomes of longitudinal data collection on participants' attitudes or social-emotional behavior.

The influence of media messages on measured behavior has been conducted mostly in research on health-risk behaviors. Media messages are associated with an increase in attendance of cancer screenings and the use of contraceptives, and a decrease in the use of recreational drugs (Wakefield et al., 2010). Health related media messages have also been shown to have a positive effect on health-related attitudes, such as attitudes towards causes and treatment of depression (Wakefield et al., 2010). News stories can also affect attitudes towards specific groups. In news media, an increasing amount of attention is given to crime stories (Grosholz & Kubrin, 2007), even though most forms of crime have decreased over the past decades (Windzio & Kleimann, 2009). Crime stories can lead to more negative attitudes towards specific groups who are overrepresented and frequently portrayed in the media as (potential) perpetrators (Arendt & Northup, 2015; Windzio & Kleimann, 2009), such as juveniles (Grosholz & Kubrin, 2007) and immigrants (Windzio & Kleimann, 2009). The influence of media messages on attitudes and behaviors has been found in studies in English-speaking countries across the world, as well as countries in Latin-America, Asia, and across Europe (Arendt & Northup, 2015; Callanan, 2012; Grosholz & Kubrin, 2007; Nellis & Savage, 2012; Popović, 2018a; Prieto Curiel et al., 2020; Wakefield et al., 2010; Windzio & Kleimann, 2009), indicating that this influence is present in many cultural contexts.

Children are overrepresented in news media as crime victims (Grosholz & Kubrin, 2007), and CSA is overrepresented when compared to other forms of child abuse (Hove et al., 2013). CSA news stories often focus on the most heinous stories and the most extreme offenders (Hove et al., 2013; Popović, 2018b; L. R. Smith & Pegoraro, 2020; Weatherred, 2015, 2017). Research in the USA shows that there are big spikes in the number of newspaper articles on CSA attributable to specific cases and

perpetrators (Hove et al., 2013; Weatherred, 2017). There is evidence that parents associate male childcare professionals with pedophilia and CSA (Rentzou, 2011; Sak et al., 2019). A link between the attention for CSA in news media and a rise in suspicion towards male childcare workers among parents, especially mothers, has been theorized in research (Eidevald et al., 2018; Munk et al., 2013), but to the best of our knowledge this relation has not yet been studied.

As the coverage of CSA in the media can be capricious (Weatherred, 2017), a big CSA case that unfolds during a research period might cause an unexpected spike in attention for the topic among research participants. News messages have been found to perpetuate stereotypes about perpetrators (Cromer & Goldsmith, 2010a), as well as elicit stereotypical responses from the public (Popović, 2018a). Increased coverage of CSA in childcare could therefore lead to a more negative attitude towards men working in early childcare education (Eidevald et al., 2018; Munk et al., 2013). This may subsequently influence attitudes and behaviors that are being studied. An explanation of how media consumption influences attitudes and behavior is posed in the resonance thesis (Weitzer & Kubrin, 2004). This states that when the content of media coverage is closely linked to a person's everyday reality or vulnerability for that type of crime, media and personal experience jointly reinforce fear and perceived risk. This mechanism could explain why CSA news coverage would lead to an increase in parents' fear and perceived risk of CSA, and subsequently be negatively related to parental attitudes towards male childcare professionals. Parents of young children may be particularly sensitive to CSA in the news because of their personal vulnerability.

Case study: a big CSA news story during a longitudinal research project

The focus of this paper will be on the attitudes of parents who participated in the longitudinal project "Boys will be Boys?" towards letting a male caregiver mind their children before and after the first news stories about

a big CSA case were published. On December 7th 2010, a male childcare professional was arrested in Amsterdam, the Netherlands. This perpetrator was ultimately convicted of 67 counts of CSA concerning children aged 0 to 4, as well as the possession, creation, and distribution of child pornography. The perpetrator worked in formal and informal childcare. The case was later dubbed the biggest sex crime case of all time in the Netherlands by the media (Koch, 2012). However, to the best of our knowledge, the amount of media attention this case received has never been studied.

The aim of the longitudinal project Boys will be Boys? was to examine the relation between parental gender-differentiated socialization and the socio-emotional development in early childhood. The research project took place over the course of four years, consisting of four annual waves. During the first wave (June 2010 - June 2011), approximately half of the families participated before the Amsterdam CSA case first became known (December 7th, 2010) and the other half afterwards. As part of the project parents filled in a questionnaire measuring their gender attitudes, which contains an item on their attitude towards hiring a male babysitter. All families had children in the age range of the victims in the Amsterdam CSA case, and the majority used a form of childcare (11% only informal, 57.5% only formal, 21.5% both formal and informal). Consequently, the news of the Amsterdam CSA case could be considered close to their everyday reality. We expected that: (1) there would be a spike in media attention related to the perpetrator and CSA in general; (2) during Wave 1, fathers and mothers who participated after the news of the Amsterdam CSA case became public would show a more negative attitude towards hiring a male babysitter than those who participated before the news; (3) this effect in Wave 1 would be stronger for mothers than for fathers; (4) during the three subsequent waves both mothers and fathers would continue to have a more negative attitude towards hiring a male babysitter than during Wave 1; (5) the effect of date of participation in

Wave 1 did not occur for other aspects of parents' gender attitudes that did not relate to males in formal caregiving.

Method

Salience of CSA in the media

To index the frequency of yearly publications on topics related to CSA, the NexisUni® Academic database was searched. This database has been recommended as the most important source for content analysis of textual news data (Krippendorff, 2012; Weatherred, 2017), and has been used before to investigate patterns of media attention over time (Van der Meer et al., 2019). The five largest Dutch newspapers were analyzed, which had a combined daily circulation of 1.63 million copies during Wave 1 of the longitudinal study (Instituut voor Media Auditing, 2011). While our data collection took place in a period in which television was an important news source and online news consumption saw a large increase, we decided to focus on newspapers. Research has shown that the content of social media is greatly influenced by mass media (Prieto Curiel et al., 2020) and content found and shared on social media is often published by traditional (print) media outlets (Möller et al., 2020; L. R. Smith & Pegoraro, 2020). Searching the NexisUni® Academic database offers convenience and easy access to Dutch newspapers. Dutch television news broadcasters do not have an accessible archive.

To investigate the salience of topics related to CSA, the number of articles containing the word "*zedenzaak*" (sex crime) and the number of articles containing the word *"kinderporno"* (child pornography) were indexed per year, starting on January first 2006 and ending on December 31st 2014 (i.e., four years before and after the Amsterdam CSA case became known, and during the longitudinal study which took place between 2010–2014). Search results contained some duplicates which were deleted (n = 19). To further examine the connection between the Amsterdam CSA case

and newspaper content mentioning sex crimes and child pornography, articles containing the name of the perpetrator, Robert M., were also registered per year. As there are some notable persons who have a similar name (e.g., author Robert M. Pirsig), the heading of each article was read to ensure suitability for the study. When in doubt, a longer part of the article was read. Articles not related to the case were deleted (n = 36), as well as duplicate articles (n = 2). No articles were published about the perpetrator before December 7th 2010. There was an overlap between the articles that were included in the counts for sex crime, child pornography, and the name of the perpetrator. The overlap was registered as well. As the longitudinal study consisted of four annual waves that stretched from midway June until June the following year, these processes were repeated for the four waves. The total process of indexing took approximately 10 hours.

Longitudinal family study

Sample

This study is part of the longitudinal study "Boys will be Boys?" Families with a father, a mother, and two children were selected from municipality records in the Western region of the Netherlands. Families were included if the second-born child was around 12 months of age and the firstborn child was approximately 2 years older. Out of 1249 eligible families, 390 families participated in the study, which involved multiple home visits with video observations, questionnaires, and computer tasks. For more information on the sampling method and inclusion criteria study see Hallers-Haalboom et al., 2017. For the current report, families were selected if both parents filled out the Child Rearing Sex Role Attitude Scale (CRSRAS, see below) in each of the four waves. Due to the intensive data collection methods, in 47% of the families at least one parent did not fill in the questionnaire in at least one of the waves. This resulted in a subset of 207 families. The subset did not differ significantly in terms of age or educational level.

In Wave 1, 97 families participated before the news about the Amsterdam CSA case was first published and 110 families afterwards. Data collection took place annually as close as possible to the birthday of the second child. In Wave 1, mothers' ages ranged from 25.1 to 45.6 years (M = 34.19, SD = 3.94) and fathers' ages from 25.8 to 53.3 years (M = 37.33, SD = 5. 41). The children were on average 3.11 years (SD = 0.30) and 1.02 years (SD = 0.03) old. Most parents were highly educated: 81.7% of mothers and 76. 8% of fathers (International Standard Classification of Education (ISCED) level 6 or higher). In the first wave, 94.6% of the parental couples were married, registered partners or had a cohabitation agreement and 5.4% were unmarried. During the following waves, five of the married couples got divorced. The sibling gender combinations were as follows: 22.2% girl-girl, 23.7% boy-boy, 28% girl-boy, and 26.1% boy-girl.

Procedure

All families were visited in each of the four waves. Parents were asked to individually fill out a set of questionnaires, including the CRSRAS, before every visit. Families received a yearly gift of 30 Euros and small presents for the children. For more information about the procedure see (Hallers-Haalboom et al., 2017). Informed consent was obtained from all participating families. Ethical approval for this research was provided by the Research Ethics Committee of Commission Research Ethics Code of the Leiden Institute of Education and Child Studies (ECPW-2009/014).

Instruments

CRSRAS. The CRSRAS (Freeman, 2007; adapted from Burge, 1981) was used to assess fathers' and mothers' explicit attitudes toward gender-differentiated parenting. The questionnaire consisted of 19 items that were completed on a 5-point scale ranging from 0 (*strongly agree*) to 4 (*strongly disagree*). The internal consistency of the CRSRAS was adequate at each wave, with the Cronbach alpha coefficient ranging from .75 to .79 for fathers, and .72 - .75 for mothers. For this study, three items relating

to stereotypical attitudes about gender-specific roles of boys/men were selected, namely: 1) *I would not hire a male babysitter,* 2) *Parents should handle the behavior of boys and girls differently,* and 3) *I could buy my son a doll.* The third item was recoded, so a lower score on each item indicates a more stereotypical attitude. The first item was chosen as focus-item as its content relates closely to the Amsterdam CSA case. The other two were included to compare patterns of responses to the focus item.

Data analysis

Media attention related to Amsterdam CSA case was indexed using NexisUni®. Using Microsoft Excel; frequency tables were created and the data was subsequently used to create graphs for visual inspection in order to detect patterns and spikes in media attention. Further analyses of the data of the longitudinal family project were carried out using SPSS 25.0. The three CRSRAS items were inspected for outliers, defined as values more than 3.29 *SD* above or below the means (Tabachnick & Fidell, 2007b), no outliers were found.

We examined whether parents who participated after the Amsterdam CSA case scored lower than those who participated before the news in Wave 1 with independent samples *t*-tests. *T*-tests have been found suitable for examining single five-point Likert items (Norman, 2010). The tests were carried out separately for fathers and mothers to examine the role of parent sex. To investigate the effect of the moment of participation over time and the interaction effect with parent sex generalized linear mixed model (GLMM) analyses were conducted for each item. The within subjects factors parent sex (father, mother), and time (the four waves), and the between subjects factor timing of participation (before or after the news) were examined. The third item (doll for son, recoded) was non-normally distributed (standardized skewness ranged from -8.25 to -6.33 and standardized kurtosis from 1.00 to 4.20). The central limit theorem establishes that for sample sizes greater than 10 per group, regardless of the distribution of the dependent variable, the means are approximately

normally distributed (Norman, 2010). As the sample sizes per group in our study are 97 and 110, the tests are robust.

In this paper both η_p^2 and $\hat{\omega}_p^2$ are reported as effect sizes, because while η_p^2 is broadly reported, it is a biased measure of effect and it has been recommended to additionally report $\hat{\omega}_p^2$ (Yigit & Mendes, 2018). The formulas used to calculate partial omega squared can be found in Appendix B.

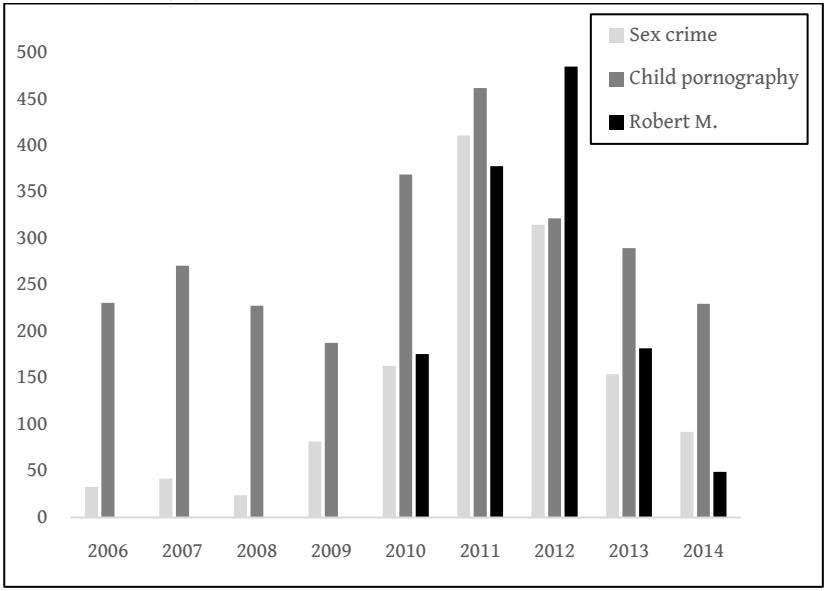
Results

Salience of CSA in the media

The frequency of articles containing the words sex crime, child pornography and the name of the perpetrator per year can be found in Figure 5.1. In the four years before the Amsterdam CSA case became known (2006-2009), on average 229.50 (*SD* = 33.91) articles per year contained the words 'child pornography', and 45.25 (*SD* = 25.58) 'sex crime'. In 2010, when the news about the case broke, a large increase is visible, with 369 articles mentioning child pornography and 163 mentioning sex crime. The news about the Amsterdam CSA case broke on December 7th 2010. Of the articles that contained the word child pornography that were published in 2010, 35.7% was published after December 6th, for articles containing the word sex crime this was 71.8%. The highest peak of articles that contained these words occurred in 2011, with 462 and 411 articles respectively. In 2012-2013 there were still noticeably more articles on sex crimes and child pornography than in 2006-2009. All articles containing the name of the perpetrator were published after December 6th 2010. The largest peak in articles mentioning him occurred in 2012, when the most important parts of his trials and sentencing took place. The number of published articles was strongly correlated with the number of articles containing the word sex crime, $r = .94$, $p < .001$, and the number of articles containing the word child pornography, $r = .76$, $p = .018$.

Figure 5.1

Number of Newspaper Articles Per Year

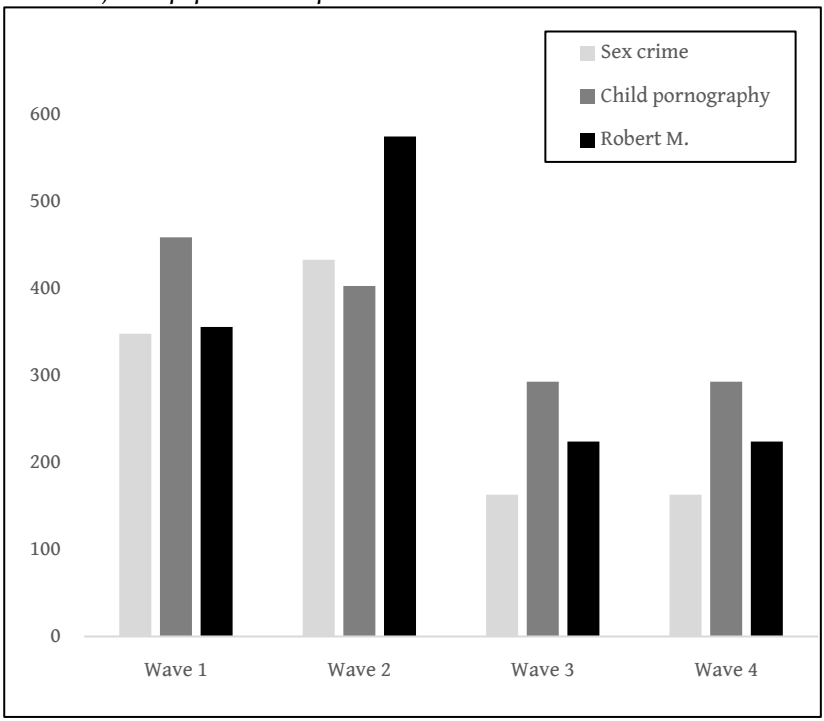


Note. Frequency of news articles per year in the five biggest newspapers in the Netherlands that contain the words sex crime (*zedenaak*), child pornography (*kinderporno*), and/or the name of the perpetrator (Robert M.).

When looking at the articles that were published per wave during the longitudinal study "Boys will be Boys?", between June of 2010 and June of 2014, it becomes apparent that the largest peak for all three investigated terms took place during the second wave of the study (Figure 5.2). Over the four waves, there was overlap between articles. Of the articles containing the word sex crime 12.5-21.8% per wave mentioned the name of the perpetrator, this was 3.9-6.6% for articles containing the word child pornography, and 5.4-11.1% mentioned all three.

Figure 5.2

Number of Newspaper Articles per Wave



Note. Frequency of news articles per Wave in the five biggest newspapers in the Netherlands that contain the word sex crime (zedenzaak), child pornography (kinderporno), and/or the name of the perpetrator (Robert M.). Each wave started in June and ended in June of the following year, with Wave 1 starting in 2010 and ending in 2011.

Longitudinal family study

Means and standard deviations on the three CRSRAS items selected for this study can be found in Table 5.1. The difference in attitude towards hiring a male babysitter between parents who participated before the initial breaking of the news about the Amsterdam CSA and parents who participated afterwards in Wave 1 was investigated (Figure 5.3). A significant difference was found between fathers who participated before and fathers who participated after the news first became known (mean

difference = 0.47, 95% CI = 0.10 to 0.83; $t(205) = 2.51$, $p = .013$; Cohen's $d = 0.35$). A significant difference was also found between mothers who participated before the news and those who participated afterwards (mean difference = 0.86, 95% CI = 0.51 to 1.21; $t(204.89) = 4.90$, $p < .001$ (two-tailed); Cohen's $d = 0.68$). Cohen's d indicates that the effect of the news breaking during Wave 1 was stronger for mothers than fathers.

Table 5.1

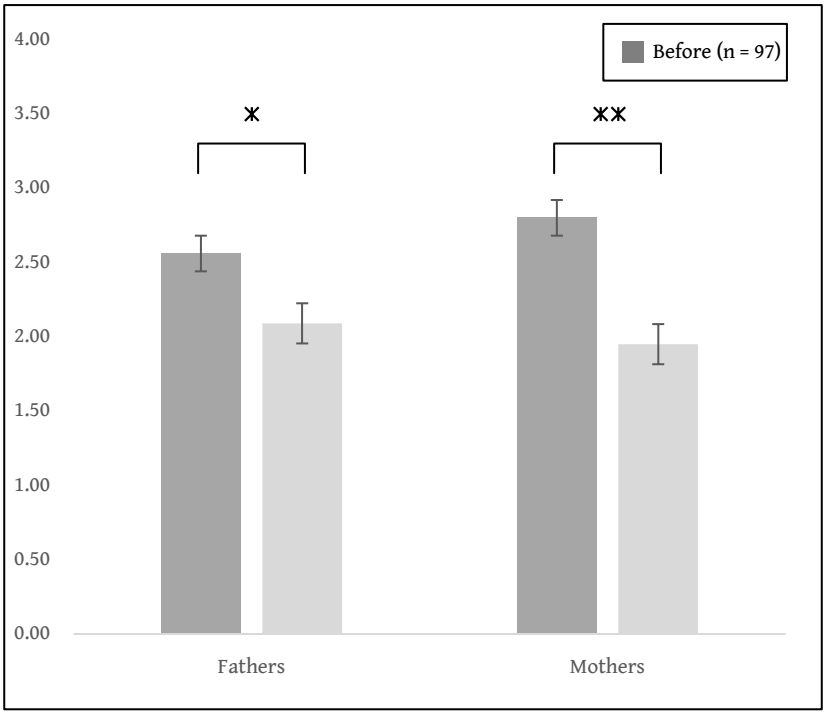
Means and Standard Deviations on three CRSRAS items for fathers and mothers (N = 207).

	Item 1		Item 2		Item 3	
	Fathers	Mothers	Fathers	Mothers	Fathers	Mothers
Wave	M(*SD*)	M(*SD*)	M(*SD*)	M(*SD*)	M(*SD*)	M(*SD*)
T1	2.31 (*1.35*)	2.35 (*1.33*)	2.28 (*1.28*)	2.43 (*1.18*)	2.26 (*1.31*)	3.00 (*1.10*)
T2	1.95 (*1.37*)	2.13 (*1.31*)	2.14 (*1.23*)	2.48 (*1.21*)	2.48 (*1.28*)	3.15 (*1.09*)
T3	2.07 (*1.36*)	2.17 (*1.35*)	1.95 (*1.27*)	2.43 (*1.20*)	2.45 (*1.30*)	3.14 (*1.03*)
T4	2.12 (*1.40*)	2.23 (*1.33*)	2.03 (*1.21*)	2.38 (*1.17*)	2.41 (*1.26*)	3.17 (*0.97*)

Note. Item 1 = I would not hire a male babysitter, Item 2 = Parents should handle the behavior of boys and girls differently, Item 3 = I could buy my son a doll. Items were completed on a 5-point scale ranging from 0 (*strongly agree*) to 4 (*strongly disagree*). The third item was recoded, so a lower score on each item indicates a more stereotypical attitude.

Figure 5.3

Differences in Fathers' and Mothers' Attitude Towards Hiring a Male Babysitter



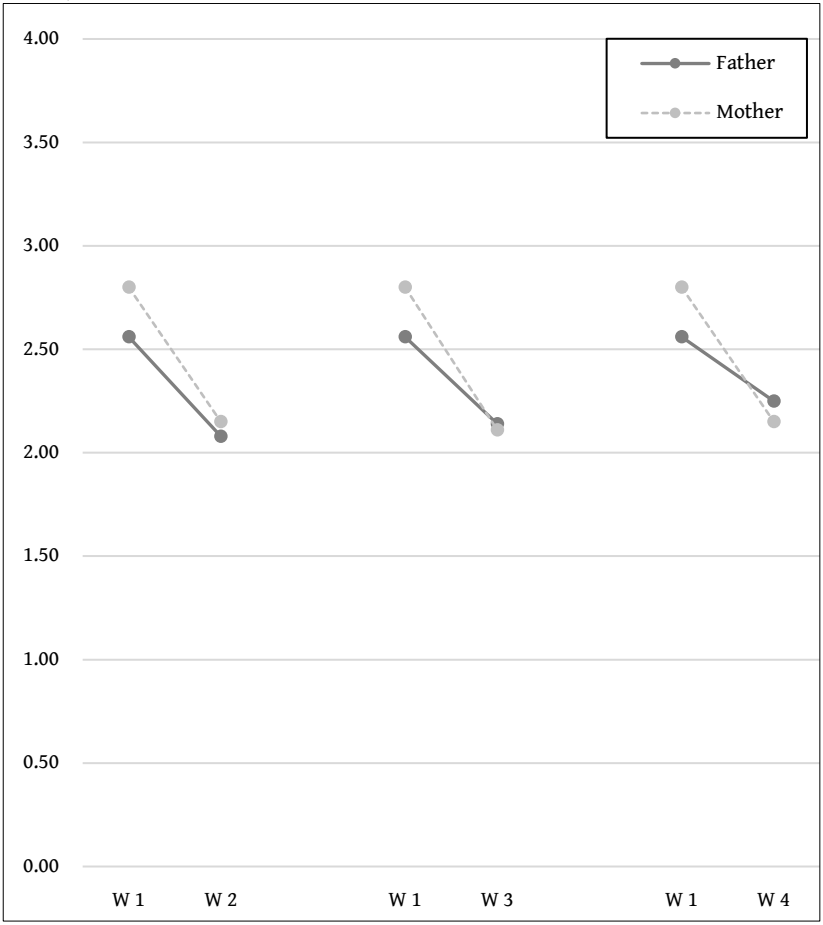
Note. Error bars: 95% CI. Changes in fathers' and mothers' mean attitude scores for the first item (I would not hire a male babysitter) in Wave 1. There was a significant difference between the parents who participated before the Amsterdam CSA case and those who participated after.

* p < .05 ** p < .001

GLMM analyses were conducted to examine parents' attitudes over time. The interaction effect of the moment of participation over time and the interaction effect with parent sex was investigated. Regarding Item 1 (hire male babysitter) Mauchly's Test of Sphericity indicated that the assumption of sphericity had been violated for the within-subjects factor time, $\chi^2(5) = 18.12$, $p = .003$, therefore all degrees of freedom in the following analyses were corrected using Greenhouse-Geisser estimates of sphericity ($\varepsilon = 0.94$; (Keppel, 1991). A significant three-way interaction effect was found between time, moment of first participation and parent gender, $F(2.95, 604) = 3.88$, $p = .010$, $\eta_p^2 = .020$, $\widehat{\omega}_p^2 = .005$. This indicates there was an effect of moment of participation over time, which was moderated by parent sex. Contrasts were used to break down the interaction, in which Wave 1 was compared to the subsequent waves. Parents who participated before the news broke in Wave 1, showed a lower inclination to hire a male babysitter in each subsequent wave after the news came out (Figure 5.4A). A sex specific effect emerged when looking at parents who participated after the news broke in Wave 1 (Figure 5.4B). Firstly, Wave 1 was compared to Wave 2, $F(1, 205) = 5.50$, $p = .020$, $\eta_p^2 = 0.03$, $\widehat{\omega}_p^2 = .005$. Fathers showed a lower inclination to hire a male babysitter in Wave 2 than Wave 1, while mothers showed no difference between these two waves. Secondly, comparing Wave 1 to Wave 3, $F(1, 205) = 7.43$, $p = .007$, $\eta_p^2 = 0.04$, $\widehat{\omega}_p^2 = .014$. Fathers showed no difference in their inclination compared to Wave 1, while mothers showed a higher inclination in Wave 3 than Wave 1. The same occurred when comparing Wave 1 to Wave 4, $F(1, 205) = 9.20$, $p = .003$, $\eta_p^2 = 0.04$, $\widehat{\omega}_p^2 = .020$. Fathers again showed a similar disposition in Wave 1 and Wave 4, while mothers showed a higher inclination to hire a male babysitter in Wave 4 than in Wave 1. Regarding item 2 (gender differentiated parenting of child), as well as item 3 (doll for son, recoded) independent *t*-tests showed no effect for the moment of participation in Wave 1. A main effect was found for time for Item 2 ($F(2.97, 608.11) = 3.11$, $p = .027$, $\eta_p^2 = .02$, $\widehat{\omega}_p^2 = .010$) and Item 3 ($F(2.78, 570.53) = 4.97$, $p = .010$, $\eta_p^2 = .02$, $\widehat{\omega}_p^2 = .019$). This indicates that over time parents

Figure 5.4A

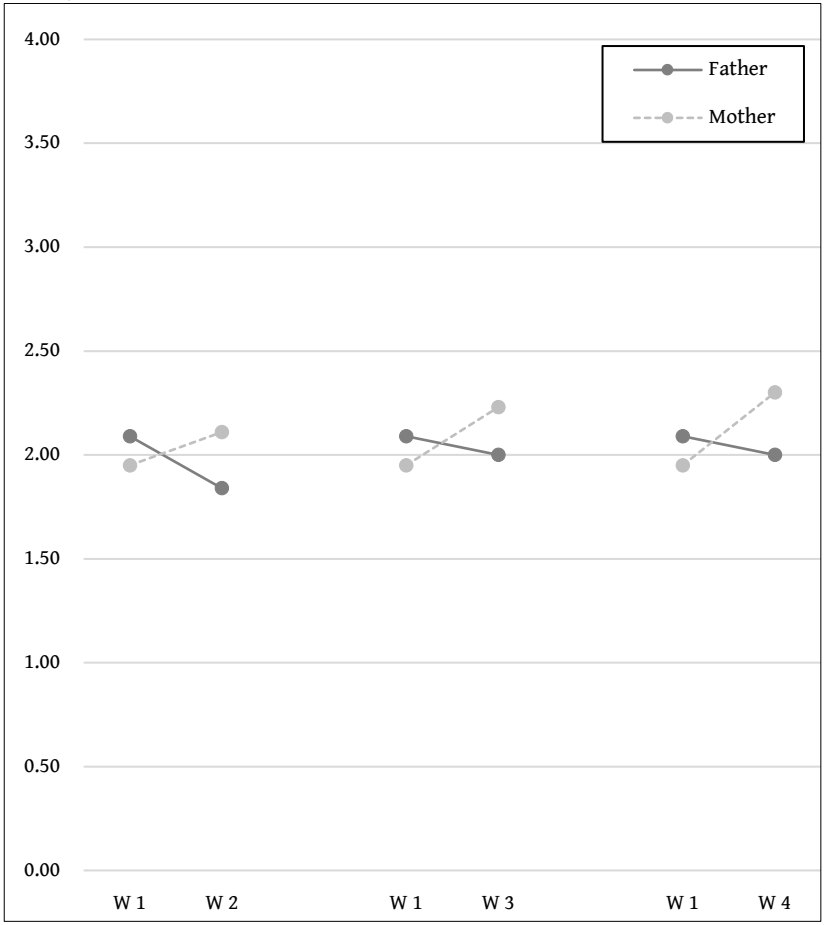
Parents' Attitude Towards Hiring a Male Babysitter in Wave 1 Compared to Subsequent Waves (N = 97)



Note. Differences in mean attitude scores on the first item (I would not hire a male babysitter), comparing Wave 1 to the subsequent waves for parents who participated **before** the CSA case became known in Wave 1.

Figure 5.4B

Parents' Attitude Towards Hiring a Male Babysitter in Wave 1 Compared to Subsequent Waves (N = *110)*



Note. Differences in mean attitude scores on the first item (I would not hire a male babysitter), comparing Wave 1 to the subsequent waves for parents who participated **after** the CSA case became known in Wave 1.

were less inclined to handle boys' and girls' behavior differently, and less inclined to buy their son a doll. For both items, neither two-way interaction effects between the moment of participation and time; nor three-way interaction effects between moment of participation, time and parent were found.

Discussion

In the current study we presented a way to index media attention on a specific topic, and investigated the possible influence of a big news story on longitudinal data collection. In concordance with research in the USA (Hove et al., 2013; L. R. Smith & Pegoraro, 2020; Weatherred, 2017) and our expectations, we found spikes in articles on CSA, compared to the four years before the Amsterdam CSA case became known. We also found prolonged attention to the perpetrator of the Amsterdam CSA case in the media. The attention culminated during the second wave of the data collection. We found that after the news of the case broke (Wave 1), parents showed a more negative attitude towards hiring a male babysitter than those who participated before the case. This effect was stronger for mothers than for fathers. This negative effect on parents' attitude towards hiring a male babysitter lasted over time and differed between fathers and mothers. None of the mentioned effects were found for aspects of parents' gender attitudes unrelated to male childcare professionals.

According to previous studies in varying cultural contexts (Arendt & Northup, 2015; Mertens et al., 2019; Popović, 2018a; Windzio & Kleimann, 2009), news coverage can negatively affect attitudes towards specific groups. The negative influence of news coverage of CSA on attitudes towards male childcare professionals has been suggested before (Eidevald et al., 2018; Munk et al., 2013). To the best of our knowledge this is the first time this effect has been studied. The found effect of the news about the Amsterdam CSA case on parents' attitudes could be explained by the resonance thesis (Weitzer & Kubrin, 2004). Having young children and

making use of childcare may play a role in parents' attitudes towards hiring a male babysitter. Consistent with earlier research (Callanan, 2012; Intravia et al., 2017; B. J. Wilson et al., 2005), mothers were affected more than fathers in Wave 1. This could be because women report higher levels of fright-related feelings in relation to crime and crime news (Callanan, 2012; B. J. Wilson et al., 2005), which could cause the bigger effect of the news on mothers' attitudes.

As expected, the negative effect on parents' attitude towards male babysitters lasted across the following waves. We found that the patterns of this enduring effect were different for mothers than fathers. Fathers remained less inclined to hire a male babysitter over the years, while mothers who participated after the case in Wave 1 showed an increasing inclination in Wave 3 and 4. The highest peak of media attention took place during Wave 2, when the perpetrator was on trial. During Wave 2, all fathers showed a lower inclination to hire a male babysitter, while mothers who were aware of the news during Wave 1 did not show a further decrease during Wave 2. Our findings indicate that while mothers' attitudes are affected more by big news stories initially, their attitudes may also recover faster or they are less sensitive to prolonged media attention to a certain crime. Further research into this mechanism is recommended. The sample in this study was quite homogenous in regard to the level of education, as over 76% of parents were highly educated. Participation bias in intensive longitudinal research is a well-known problem (Goodman & Gatward, 2008; Schmidt & Woll, 2017). Previous studies have shown that fear of crime is negatively associated with level of education (Callanan, 2012; B. J. Wilson et al., 2005), so with a more representative sample the effect could possibly be larger. Studying the effect of big news stories in a more heterogeneous sample is recommended to investigate such effects in more detail. Doing so within a less demanding study design could prevent participation bias.

Our findings relate to one specific component of explicit attitudes towards gender-differentiated parenting, so we can only make limited

conclusions. What impact big news stories could have on social behaviors and attitudes remains to be studied. However, our aim was foremost to present a case study to illustrate the possible influence of news media on longitudinal data collection and how researchers can include media attention in their work. We have demonstrated that this can be done in a low-cost and timely manner. To conclude, our findings provide evidence that big news stories can influence parents' gender attitudes over a longer period of time. As the consumption of news media has risen over the past decades (Wonneberger et al., 2012), and the increase of online news consumption has shown to elicit stronger and more polarized effects on attitudes (Intravia et al., 2017; Mertens et al., 2019; Popović, 2018a), we advise researchers to be mindful of, and where possible take into account, such effects in future studies.

Chapter 6

General Discussion

The aim of my dissertation was to gain deeper understanding into gender and ethnic bias in socialization by various socializing agents. In this general discussion, I will connect previous chapters to socialization theories and summarize the findings, present limitations and future directions, and I will end my dissertation with general conclusions and implications for parents, policy makers, and educators.

Theoretical reflection and summary of findings

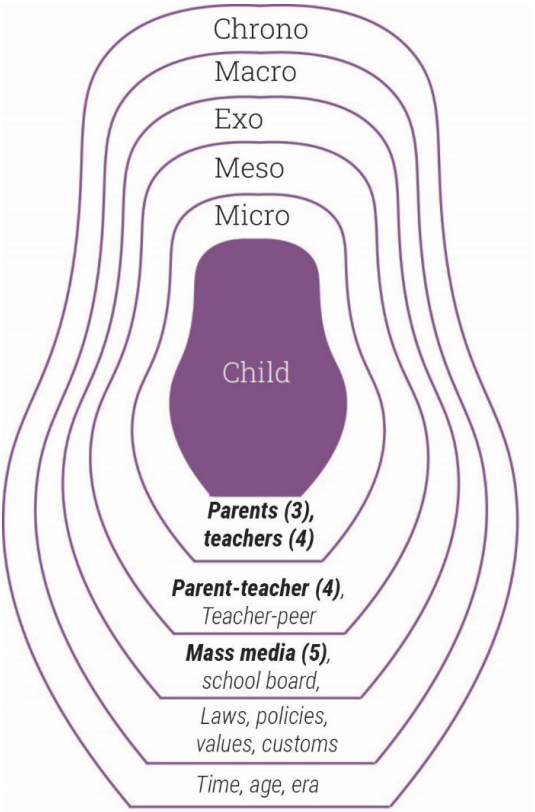
It is important for (behavioral and social) research to be based in theory. This allows for better research designs, greater integration between various research areas, and for a deeper understanding of how certain mechanisms work (Leaper, 2011; Zaikman & Marks, 2017). Therefore, in this section the previous chapters will be reconnected to the theories presented in the general introduction (Chapter 1).

According to the ecological model of development, constructivist theorists, and the social cognitive learning theory, learning and development take place within social settings and interactions (Bandura, 1986; Bronfenbrenner, 1977, 1979; Bussey & Bandura, 1999; Piaget, 1954; Vygotsky, 1978). In the lives of developing children, there are many important people and entities that form these settings and are part of direct or indirect social interactions. These people and entities can be seen as models of cognitions, attitudes, and behaviors. In social cognitive learning theory they are called socialization agents (Bandura, 1977, 1986; Bandura & Walters, 1963; Bussey & Bandura, 1999). The socialization agents can transfer biased attitudes and behaviors unto individuals, and they can affect the motivations, emotions, and values of individuals (Bussey & Bandura, 1999). In this dissertation, I focused on three different socialization agents: parents, teachers, and the media.

By focusing on various socializing agents, we also acknowledge the ecological perspective that children's development takes place within various systems. According to the ecological model, there are five types of systems that engulf the developing individual (see Figure 6.1 Bronfen-

Figure 6.1

Bronfenbrenner's Ecological Systems Model



Note. The model consists of five types of systems that engulf the developing child like nesting dolls. The names of the five types of systems are seen at the top of the model, examples of the systems are in italics at the bottom of the model. Examples in bold italics represent the socializing agents for each chapter in this dissertation, with chapter numbers in parentheses.

brenner, 1979). Bronfenbrenner, who established the ecological model, urged researchers to not just look at the individual level of the developing child, but also at other levels of the ecological model (1977, 1979). He argued that other systems, like the interactions between parents and teachers at the level of the mesosystem, and values and beliefs at the level of the macrosystem, should also be studied as they affect the lives of individuals. Below, I will place the socialization agents within the various types of systems, and summarize the findings of the previous chapters.

Parents

In **Chapter 3**, we focused on parents as socializing agents, and the messages that they communicate to their adolescent children about sexual behavior. We aimed to investigate whether the sexual double standard (SDS) was present. The SDS entails the way in which men and women are treated differently for the same sexual behavior (Zaikman & Marks, 2017). Based on gendered expectations of sexuality, men generally gain social status through sexual activity, while women are more likely to be derogated (Endendijk et al., 2020). Parents are part of a microsystem, in which they have direct and immediate interactions with their developing child. We investigated these direct interactions, in which parents answered questions about adolescents' sexual behavior read to them by their adolescent child. In previous research on the SDS, findings were inconsistent, which led researchers to questioning the perpetuated existence of the SDS (Bordini & Sperb, 2013; Zaikman & Marks, 2017). One reason for these inconsistencies was that the SDS was commonly investigated outside of real world contexts and relevant social settings (Marks et al., 2019; Marks & Fraley, 2007). By examining the microsystem of interactions between parents and their children, we increased the likelihood of finding authentic sexual scripts relevant to the participating parents and their adolescent children.

In our study, we specifically focused on interpersonal scripts used by parents. Interpersonal scripts involve the way in which broad social

norms are adapted and/or diverted from within social interaction (Wiederman, 2015). By using a structured interview, we were able to elicit various social scripts. This was important because another cause of inconsistencies could be the common use of closed-ended and Likert-type questionnaires in SDS research (Endendijk et al., 2020). Using these types of instruments increases the participants' awareness of the research aim, which in turn increases the likelihood of social desirability affecting the answers of parents. Moreover, these types of instruments not only lack relevant social context, but also do not allow the discovery of new social scripts (Elliott, 2010; Simon & Gagnon, 2003). Likert-type questionnaires often focus on specific cultural scripts, which commonly relate to promiscuity, assertiveness, passivity, the (perceived) romantic nature of women, and positive and negative evaluations (Caron et al., 1993; Endendijk et al., 2020; Muehlenhard & McCoy, 1991). Through content analysis, we found that some parents indeed use interpersonal scripts that were clearly related to these cultural scripts. Firstly, some parents related having six successive relationships and kissing a stranger to promiscuity. However, there were no gendered differences in the use of these types of scripts, indicating that there was no SDS present within these scripts. Secondly, parents used scripts that positioned boys as players and girls as sluts. This relates to gendered ideas and expectations of men as sexually agentic, dominant, powerful, and successful when having multiple partners (Fjær et al., 2015; Kreager & Staff, 2009; Reid et al., 2011). Woman who have multiple partners are more often penalized, which is reflected in the use of derogatory terms like slut (Armstrong et al., 2014; Farvid, 2010; Kreager et al., 2016). Thirdly, parents used scripts describing girls as dependent, passive, and romantic creatures, while these scripts were rarely used to describe boys. However, we found that the most common scripts related to normative sexual development, which refers to adolescents' sexuality as normal and acceptable, and adolescence as a time of sexual initiation, discovery, and experimentation into sexual activity (Tolman & McClelland, 2011; World Health Organization, 2002).

These findings indicate that, within social interaction in the microsystem of Dutch parents and children, parents most commonly used scripts that are often missing in SDS research. As such, our study adds to the body of knowledge on SDS, and offers a stepping-stone for further investigation into scripts involving sexual behavior of adolescents.

Teachers

In **Chapter 4**, we focused on teachers as socialization agents and whether bias based on gender and ethnicity was present in their written evaluations of pupils. The report cards we investigated in this chapter, can be seen as authentic and natural data (Ni & Li, 2013), that form an interaction between teachers and pupils. Teachers can therefore be seen as part of a microsystem, as they have direct and immediate interactions with their pupils. However, the report cards that were investigated in Chapter 4 regularly appeared to be aimed at the parents, making the report cards an interaction or interrelation between microsystems. This means that the report cards could also be seen as part of a mesosystem, in which teachers interact with parents (Bronfenbrenner, 1977). The findings of research on gender and ethnicity based teacher bias have been inconsistent (Wang et al., 2018). Specifically within the Dutch context, one explanation for these inconsistencies can be found in the methodological approaches. Some Dutch studies that were aimed at interactions and school processes through observations and ethnographic methods, do indicate bias (Denessen et al., 2020; Van den Bergh et al., 2010; Weiner, 2015, 2016). However, the most common research traditions consist of large scale quantitative studies, which mainly focus on describing differences in educational outcomes between groups of pupils, without addressing possible gender and ethnicity based bias in school processes (Stevens et al., 2011, 2019). These research traditions are fueled through funding by the Dutch government, and by close relationships between educational researchers and policymakers (Stevens et al., 2011, 2019). These relationships illustrate how macrosystems indirectly affect

developing children, as policies informed by these studies influence the daily school practice of pupils. In these large-scale quantitative studies, lower achievements are subsequently explained by characteristics of the pupils and their families, like the lack of certain resources that pupils from low SES and ethnic minority backgrounds experience. While these studies provide valuable information about achievements gaps, they disregard possible social injustices within school processes. Our focus on the authentic messages communicated by teachers to pupils and their parents therefore forms an important addition to the study of teacher bias.

We focused on the evaluations on report cards because it is very difficult for teachers to inhibit subtle and implicit forms of bias in written evaluations (Beukeboom, 2014). We examined differences between the messages that teachers wrote on report cards of boys and girls, belonging to the ethnic majority group or to ethnic minority groups. We found that ethnic majority girls received the most positive statements. Negative statements were formulated more abstractly towards ethnic minority pupils than towards ethnic majority pupils. This finding indicates that teachers showed negative bias towards ethnic minority pupils, because people tend to use higher levels of language abstraction when they describe things in concordance with their expectations (Beukeboom, 2014; Menegatti et al., 2017; Semin & Fiedler, 1988). Ethnic minority girls especially received more negative statements related to performance. Statements unrelated to performance mainly involved behavior and character traits. Boys received more negative performance-unrelated statements than girls did, and ethnic minority pupils received more of these than ethnic majority pupils did. This means that ethnic minority boys received most of these negative statements, and ethnic majority girls the least, further indicating bias based on gender and ethnicity. Our findings reinforce the idea that research in the Netherlands into educational inequalities could benefit from employing greater variation

of methodologies, including in-depth, qualitative, observational, and ethnographic studies (Stevens et al., 2011, 2019).

Media

In **Chapter 5** we focused on the media, which is an exosystem, as neither the child nor the parents are active participants in mass media (for studies that focus on Child Sexual Abuse (CSA) in social media, where individuals are active participants, see for instance Popović, 2018a). The media are often specifically mentioned as agents of socialization and communicators of cultural scripts (Brown & Stone, 2018; Bussey & Bandura, 1999; Endendijk et al., 2022; Wiederman, 2015; Zaikman & Marks, 2017). The content of magazines, newspapers, television programs, films, and advertisements have often been investigated, and bias and prejudice towards various groups have been found regularly (Aubrey et al., 2020; Coyne et al., 2014, 2016, 2021; Runderkamp et al., 2022; Signorielli, 2009; Weatherred, 2017). However, less is clear about the transmission of bias and prejudice by the media. It is generally accepted that the media affect attitudes, but it is unsure to what extent and whether these effects last over time (Easteal et al., 2015; Mutz & Goldman, 2010). While some studies indicate longitudinal effects on biased attitudes (Arendt & Northup, 2015; Coyne et al., 2014, 2016; Schiappa et al., 2006; Slater, 2007), others show none (Breuer et al., 2015; Shi et al., 2018) or only towards specific groups (Tukachinsky et al., 2017).

In concordance with previous research (Weatherred, 2015, 2017), it became clear that a single large CSA case had caused a prolonged peak of media attention to CSA. In this case, a male early childcare professional and babysitter, was eventually found guilty of the sexual abuse of a large number of children aged 0 to 4 years. The effect of the prolonged media attention to a notorious CSA case on the attitudes of parents towards hiring a male babysitter, adds proof to the theory of the media as agents of socialization, spreading cultural scripts of bias by emphasizing the unfitness of men in caring roles. Additionally, it illustrates how more

distant systems can affect children's development. Previous obscurities in research on longitudinal effects of media on attitudes may be explained by the resonance thesis. According to this thesis, bias and prejudice can be reinforced by the media, when media coverage is closely linked to an individual's everyday reality (Weitzer & Kubrin, 2004).

Subtle bias and small effects

The bias in the socializing messages in Chapters 3, 4, and 5 was mostly subtle. For instance, while we found that the media attention for the big CSA case in Chapter 5 was large and prolonged, newspaper articles did not state that men working in early childcare are pedophiles. Still, the news story affected parents' attitude towards hiring a male babysitter. Our findings of subtle expressions of bias is in concordance with other contemporary bias phenomena. In the past, bias was often overt and explicit, but nowadays, bias is mainly subtle and elusive (Dovidio et al., 2016). This makes the recognition and study of bias more complex. This is exemplified by the development of the SDS, as described in Chapter 3. The SDS was first introduced and investigated in the 1950s and 1960s, in the context of premarital sex (Reiss, 1956, 1964). In this era, the SDS was investigated with scales consisting of explicit statements. Results showed that people were on average more negative about women who engaged in premarital sex than they were about men. In more recent years, the SDS was no longer found consistently, especially in studies that employed similar methods (Bordini & Sperb, 2013; Endendijk et al., 2020; Kreager et al., 2016; Zaikman & Marks, 2017). By using innovative methods, focusing on people for whom the concepts are relevant, and collecting data that is as authentic as possible, the study of contemporary bias can be aided.

According to guidelines (Kirk, 1996), most of the effects found in the quantitative studies in Chapters 4 and 5 are small. The relevance of effect size has been discussed at length in Chapter 2. Most importantly, effect size measures provide insight into the practical significance of an effect. In theory, small effect sizes indicate low practical significance. So, does

this render our findings in Chapters 4 and 5 practically insignificant? There are several reasons to conclude otherwise. Firstly, small effects are interesting and practically significant for several reasons (Prentice & Miller, 1992), as established in Chapter 1. The most important reason may be the possible accumulation of adverse effects caused by bias. For instance, some of the effects of pupil gender and ethnicity on teachers' written evaluations are small, but they can accumulate with the effects of bias in other school processes. This can result in the widening of achievement and wellbeing gaps between specific groups of pupils (Rubie-Davies, 2015). This accumulation may be what makes the subtle expressions of bias particularly pernicious. While subtle discrimination can be hard to recognize and is often expressed unknowingly and unintentionally, it has a higher frequency and thus more chronic nature compared to overt discrimination (Jones et al., 2016). So while a small microaggression may not have a large impact on its own, due to cumulative effects these small subtle expressions of bias can lead to large unfair inequalities in the long run (Dovidio et al., 2016).

Secondly, the cut-off values that are used to decide whether an effect is small, medium or large are arbitrary (Cohen, 1988). Effect sizes cannot be seen as a one-size-fits-all, as they vary across specific areas of research (Maxwell et al., 2018; Primbs et al., 2022). It is often unclear what the smallest effect of interest within a specific field actually is (Lakens, 2022). While some guidelines exist (Anvari & Lakens, 2021; Primbs et al., 2022), studies into what constitutes a meaningful effect within the field of subtle bias and inequality in socialization are therefore needed.

When dealing with subtle issues, it is important to use correctly calculated effect sizes. When using variance analyses (ANOVA), like we did in Chapters 4 and 5, the most commonly used effect size is (η_p^2). However, η_p^2 structurally overestimates effect size, especially with smaller sample sizes. Therefore, we explained in Chapter 2 how to calculate less biased effect sizes for variance analyses (ANOVA). Our advice includes the sharing of effect size formulas and full ANOVA tables, which makes effect

sizes easier to interpret and enables other researchers to calculate a different effect size when desired.

Limitations and future directions

Ecological approach and agency of the learner

In the previous section, the socialization agents were placed within the various ecological systems defined by Bronfenbrenner. While most systems were present within this dissertation, we focused on quite different types of biased socialization within the separate chapters. For instance, in Chapter 3, we added a relevant social context to the existing body of literature by examining the interpersonal scripts used within the microsystem of immediate interactions between parents and their child. In most studies, the SDS is studied as the attitudes or cognitions of individual participants towards sexual promiscuity of men and women, often using Likert-style questionnaires (Endendijk et al., 2020). However, we did not look into the extent to which these sexual scripts are still present within exo-, marco-, and chronosystems. We assumed that the SDS is still present in contemporary society based on literature (Bordini & Sperb, 2013; Endendijk et al., 2020; Zaikman & Marks, 2017), but did not actually investigate this. In the future, it would be worthwhile to see how these scripts are present in systems like mass media, and in societal beliefs and values, and how these scripts may change due to the age of the child and due to changes in the zeitgeist. This can perhaps also offer insights into how these types of harmful attitudes can be impeded. Likewise, in the context of Chapter 4, it would be interesting to investigate settings and interactions in which the child is not involved, to gain further understanding into gender and ethnicity based bias in schools. For instance, by examining bias among the school boards the teachers work for and related policies implemented by the municipality the teachers work with. Lastly, in relation to Chapter 5, including the perspectives of (male) early childcare professionals, their educators, and the managers who are responsible for hiring policies would result in a more complete

framework. This framework could provide a deeper and broader understanding of the effects of CSA stories in the media on the attitudes of socializing agents from various social systems. Additionally, it would aid in examining how the effects of the media can be impeded, how bias against men in caring professions can be decreased, and in how the stark gendered division of labor in these professions can be combatted.

Another, related limitation lies in the focus on the socialization agents. By focusing on the socialization agents, the individuals that are socialized can appear as being relatively passive (Bem, 1983), especially in Chapters 3 and 4, in which the adolescents who are socialized are not investigated. This is a limitation, as people assimilate, accommodate, and adapt biased socialization, within social interactions and individually at the intrapsychic level (Bandura, 1977; Piaget, 1954; Simon & Gagnon, 1986; Vygotsky, 1978; Wiederman, 2015). Therefore, while we know more about the biased messages that were sent by parents and teachers, we do not know how these were received and formed into intrapsychic scripts, attitudes, and behaviors by the adolescents at whom they were aimed. Nor do we know how the scripts, attitudes, and behaviors of adolescents affected parents and teachers. Socialization is a reciprocal process, through which both child and socializing agent are affected (Gelman, 2009). People can show great variation and flexibility in the formation of scripts and attitudes throughout their life (Bem, 1983). This variation is reflected in how the attitudes of parents towards hiring a male babysitter changed over the years following the breaking of a large-scale CSA case, as we found in Chapter 5. Future research into inequalities could therefore be further enriched by investigating which interpersonal scripts children form and use within interactions with their parents, teachers and other socialization agents, and how the interpersonal scripts of socializing agents change through interactions with children. And, especially within the context of Dutch schools, studying the effects of bias on pupil outcomes that go beyond educational achievement. For instance, the

transfer of bias, social exclusion (Alan et al., 2021), and the effects on their social functioning and emotional wellbeing (Wang et al., 2018).

Sample composition

In the preface of my dissertation, I explained that there were limitations in the composition of the samples in the context of the lack of LGBTQI+ people. Here, I will discuss some additional limitations of the sample composition.

Due to the exclusion criteria that were used during the recruitment process in the overarching longitudinal research project, the participants were mostly White, did not have severe physical or mental disabilities, and included only mixed-sex parent couples. Additionally, while this was not part of the inclusion criteria, most participants had a high socio-economic status (SES), as they were highly educated and had a high income. Overall, the samples were quite homogenous, and the participants came from privileged positions. This means that the results in most studies are only representative of a selective part of the Dutch population. This is problematic for several reasons. Firstly, previous research has shown that various socialization processes work differently for specific groups of people. For instance, a higher level of education is associated with a lower level of endorsement of prejudice (Gök et al., 2019; Peri, 2018), the effect of media on attitudes is stronger for people with low SES and people from ethnic minority groups (Callanan, 2012; B. J. Wilson et al., 2005), and the endorsement of ethnicity based bias varies between ethnic groups (Nosek et al., 2002; Tukachinsky et al., 2017). Secondly, the effects of bias and prejudice are often more severe for people who are part of marginalized groups (Fine & McClelland, 2006; Leaper & Brown, 2018). By excluding them, we miss exactly the people for whom the stakes of our research are the highest.

There are several reasons why our samples were not more representative. In part, this was because the samples were part of a research project that followed up on a longitudinal study on gender

socialization in young families. This meant that selection criteria were predetermined. In the context of SES, research has shown that people with low SES are more likely to not respond, opt-out or drop-out of research (Goodman & Gatward, 2008; Schmidt & Woll, 2017). It is possible that this was amplified by our passive recruitment strategy: we invited people through a formal letter that was sent to their home address. Using a combination of passive and active recruitment strategies, like face-to-face contact and direct conversations, could potentially have reached more low-SES individuals and families and increase the likelihood of participation (Carlson et al., 2014; Lam et al., 2016). Chances of willingness to participate are for instance further increased by word of mouth recruitment, referrals by community leaders, general practitioners, friends, and extended family members. However, adding other recruitment strategies was not feasible due to some of our inclusion criteria and due to privacy laws. As per our inclusion criteria, each family had to have at least two children, who were born in specific periods. The likelihood of finding families with these specifics through strategies like referrals or word of mouth recruitment is too small to be cost effective. We received addresses of families that satisfied the birthdates criterion from municipality records. In the past, research teams would sometimes visit addresses from these records in low-income neighborhoods for face-to-face recruitment. Because this could put undue pressure on people to participate, and because of renewed privacy laws, our only option was contacting families through mail. Using strict age criteria for children can be an important requirement to investigate detailed developmental patterns, and to compare these patterns between families and social groups. However, in future research projects, it could be beneficial to use less restrictive inclusion criteria, which gives more possibilities for various recruitment strategies. This could additionally be helpful in diversifying future samples in other areas, like ethnic and sexual diversity.

In the study on school reports in Chapter 4, I recruited an extra sample of ethnic minority pupils with the aim of increasing representation. As this was during the height of the COVID-19 pandemic, options for active recruitment were limited. It was mostly impossible to enter schools, community centers, youth clubs, religious organizations, and other places where participants could be recruited. Moreover, the pandemic made it difficult for potential participants to access report cards, as these are usually held at schools and only distributed two to three times a year. Ultimately, this resulted in a small subsample of ethnic minority pupils. While this subsample was big enough to run the analyses of variance in Chapter 4, it was not big enough to differentiate between various ethnic groups. It is quite possible that there are large differences between various ethnic minority groups (Crawford et al., 2019), which may now have gone unnoticed. In the future, collecting report cards texts may be more fruitful when done through school boards, instead of through families. By recruiting via school boards, report cards written by a large number of teachers can be examined, as we did in our study, but with a far larger numbers of report cards per teacher. This could potentially result in a higher number of school reports belonging to pupils from various ethnic groups. Additionally, future studies could employ computer-automated coding processes and natural language processing (Johnson-Grey et al., 2020; Rojek et al., 2019), making the analysis of large amounts of data more feasible.

Societal relevance: Practical and policy implications

Within specific social contexts, we found bias in the messages and attitudes of socialization agents. This is in concordance with studies that have found that bias changes over time and is still present, albeit in different and more subtle forms (Dovidio et al., 2016; Endendijk et al., 2020; Jones et al., 2016; Leaper & Brown, 2018; Van Schie, 2018). This begs the question what can be done to combat bias, and although this question was not the focus of my dissertation, it is an essential next step that can

be informed by our findings. Previous research gives several directions for fruitful interventions that have been successful in decreasing bias and prejudice. I will provide examples of various kinds of interventions below.

First, as bias expressed through various forms of microaggressions can result in serious adverse effects and increasing social inequalities (Dovidio et al., 2016; Jones et al., 2016; Rubie-Davies, 2015), paying attention to subtle forms of discrimination is worthwhile. As discussed in Chapters 3 and 4, bias can quite easily and unintentionally be expressed through language. It is vital to take inclusive language seriously (Murray, 2018). People, especially important socializing agents like teachers and policy-makers, can follow various courses and guidelines in order to improve their inclusive language skills. One example of such a guideline was published in the spring of 2023 by a Dutch national organization led by secondary school pupils which advocates for the rights of pupils (Landelijk Aktie Komitee Scholieren, 2023). However, the guideline published by this organization was met with hostility and mockery by politicians and in the media (Van der Aa, 2023). If inclusive language would be made inherent and integral in various ecological systems like schools and government communications, this could lead to more inclusive language in other social systems too. Social norms are a strong motivator to change behavior (Felten & Broekroelofs, 2022). Therefore, when people become used to reading inclusive language in many place, they will become more inclined to use similar language themselves. Additionally, publications in (for instance) parenting magazines, newspapers, and on online parenting platforms could inform parents specifically on the importance of inclusive and unbiased language, and offer valuable tips on how to use such language.

Second, direct instruction and group discussions about bias and prejudice can help in reducing them (Kågesten et al., 2016; Leaper & Brown, 2018). For instance, talking about ethnicity can reduce ethnicity based bias (Sue, 2013). Acknowledging and addressing subtle negative messages can aid in combatting bias as well (Byers et al., 2020). In a study

where a primary school teacher taught her pupils about transgender and gender-nonconforming people, pupils showed more inclusive attitudes towards gender expression and identity, as well as the ability to question existing social systems that can be restrictive (Ryan et al., 2013). Discussing gendered stereotypes about specific school subjects, like stereotypes about girls being less fit for science, technology, engineering, and mathematics (STEM) than boys, has also shown reduce bias, while also increasing girls' interest in those subjects and boys' overall school engagement (Leaper & Brown, 2018). Based on our findings in Chapter 3, where we found that some parents convey biased gender messages about sexual behavior, it may be advantageous to aim interventions at parents, suggesting how they can communicate positive aspects of sexuality to sons and daughters alike, including talk about age-appropriate sexual behaviors.

Third, facilitating intergroup contact can decrease intergroup bias, by increasing knowledge about and increasing empathy towards other groups, and by reducing fear about intergroup contact (Pettigrew & Tropp, 2008). The effects of intergroup contact are further enhanced when the groups share similar status and common goals, when the groups cooperate rather than compete, and when the contact is supported by authorities, laws, and policies (Pettigrew et al., 2011). While prejudice reduction through intergroup contact has mainly been investigated in the context of differing ethnic groups, research has shown that it can also reduce gender prejudice (Martin et al., 2017). The intergroup contact can even be indirect, for instance by watching television programs including characters from specific (marginalized) groups, like gay and bisexual people (Schiappa et al., 2006). Additionally, intergroup contact has been found to moderate linguistic bias (Prati et al., 2015).

Fourth, interventions aimed at engaging in (counter-stereotypical) activities and contact with role models that defy stereotypes have been successful in reducing prejudice and stereotypical attitudes (Kågesten et al., 2016; Leaper & Brown, 2018). For example, children who received guest

lectures by female STEM professionals showed higher gender awareness and decreased gender stereotypical ideas about STEM (Booij et al., 2013). As only very few men work in early childhood education, which appears to increase negative attitudes towards the men who do (Rentzou, 2011), interventions that show positive male role models may both decrease negative stereotypes, and increase boys' aspirations to work in this type of profession. These types of role models may offer a counterweight to the effects of mass media, as exemplified in Chapter 5.

While the examples above give directions for interventions, some studies indicate that finding (more) successful interventions is still needed. For example, a meta-review has shown that interventions using counter stereotypical role models are often not effective, and it is unclear why some interventions work while others do not (Olsson & Martiny, 2018). In a study focusing on improving the attitudes of secondary school students towards the LGBTIQ+ community through a peer educator intervention, there were even negative effects found in some student populations (Kroneman et al., 2018).

Ideally, many socialization agents and ecological systems are involved in socialization against bias. Solving societal inequalities should be a shared endeavor that does not rest on one type of agent or context (Good et al., 2018). Macro systems can be involved in decreasing bias, for instance through governmental and educational policies that stimulate interventions aimed at various socializing agents. Successful interventions include programs aimed at parents, but also at other adults that are part of other microsystems, like teachers and sport coaches (Kågesten et al., 2016). The government can stimulate makers of (educational) television programs and social media influencers to promote counter stereotypical attitudes and equality. Policies can also be aimed at making anti-prejudice programs a mandatory part of teacher education, so all teachers are stimulated to become socializing agents of inclusion and social justice (Nieto, 2000; Pantić & Florian, 2015). For teachers specifically, informed by our findings in Chapter 4, new

interventions could be developed aimed at reducing bias in written evaluations.

Other socializing agents that are likely involved in the formation of bias and prejudice include religious institutions, youth clubs and sports clubs (Kågesten et al., 2016). Better understanding of the roles of these agents is needed, as these agents can also play an important part in reducing bias. However, it is impossible to force each socialization agent to implement anti-bias interventions, and it is undesirable and unethical in a free society. Additionally, not every child is involved in a club or religious institution, and not every child watches the same television programs or social media influencers. As each child attends school, schools have been called the great equalizer, as they have the potential to create social justice and unity (Grant & Agosto, 2008). Therefore, schools and teacher education may be the best places to test and use various types of interventions, with school policies aimed at equality, mandated by the government, informed by research (Fine & McClelland, 2006; Grant & Agosto, 2008; Nieto, 2000).

Conclusions

In the past, research findings have not been conclusive on the existence and transmission of bias in various specific contexts. Dutch people indicate that they find equality important, but that they feel like emancipation has gone far enough (Kanne & van der Schelde, 2022). This attitude combined with unsureness about to what extent inequality and bias exist, may lead to inaction while action is still warranted. Therefore, I aimed to document instances of bias and provide insight into mechanisms of bias in socialization in various social contexts.

In my dissertation, I have presented theory-based evidence for bias in specific social settings in the messages of three different kinds of socialization agents: parents, teachers, and the media. Messages that were sent by parents and teachers contained various indicators of bias, based on gender and ethnicity. Biased media messages affected parental

attitudes over a prolonged period of time. To quote the title of an article by Lynn Liben (2016) on the perpetuated existence of gender inequality: "We've come a long way baby, but we're not there yet." The bias we found in various contexts, underlines that research into the characteristics of bias, socializing processes, and effective interventions remains important. The findings in my dissertation indicate that previous inconsistencies in research findings may be clarified firstly by focusing on participants for whom the measured concepts resonate with their everyday reality and the social contexts they are part of. Secondly, using data that is as authentic as possible, like responses within parent-child dyads and actual school reports, may have higher validity for bias research than, for instance, responses to Likert-type scales.

For the future, investigating authentic data, within multiple ecological systems among diverse samples forms a worthwhile recommendation. As bias and social learning are affected by time and era, this has to be done consistently and continuously. When researchers and educators find bias that leads to social inequalities, this appeals to a moral duty to intervene. Thus, hopefully, the findings in my dissertation do exactly that, and as such make a contribution to creating a fairer society.

References

A

Afifi, T. D., Joseph, A., & Aldeis, D. (2008). Why can't we just talk about it? An observational study of parents' and adolescents' conversations about sex. *Journal of Adolescent Research*, *23*(6), 689–721. https://doi.org/10.1177/0743558408323841

Alan, S., Duysak, E., Kubilay, E., & Mumcu, I. (2021). Social exclusion and ethnic segregation in schools: The role of teacher's ethnic prejudice. *The Review of Economics and Statistics*, 1–45. https://doi.org/10.1162/rest_a_01111

Albers, C. (2015). *Comment on "Why you should use omega^2 instead of eta^2."* http://blog.casperalbers.nl/science/statistics/comment-on-why-you-should-use-omega^2-instead-of-eta^2/

Albers, C., & Lakens, D. (2018). When power analyses based on pilot data are biased: Inaccurate effect size estimators and follow-up bias. *Journal of Experimental Social Psychology*, *74*, 187–195. https://doi.org/10.1016/j.jesp.2017.09.004

Algina, J., Keselman, H. J., & Penfield, R. D. (2006). Confidence interval coverage for Cohen's effect size statistic. *Educational and Psychological Measurement*, *66*(6), 945–960. https://doi.org/10.1177/0013164406288161

Alhija, F. N.-A., & Levy, A. (2009). Effect size reporting practices in published articles. *Educational and Psychological Measurement*, *69*(2), 245–265. https://doi.org/10.1177/0013164408315266

Allan, A. J. (2009). The importance of being a 'lady': Hyper-femininity and heterosexuality in the private, single-sex primary school. *Gender and Education*, *21*(2), 145–158. https://doi.org/10.1080/09540250802213172

American Psychological Association. (1994). *Publication manual of the American Psychological Association* (4th ed.). American Psychological Association.

American Psychological Association. (2001). *Publication Manual of the American Psychological Association* (5th ed.). American Psychological Association.

American Psychological Association. (2010). *Publication Manual of the American Psychological Association* (6th ed.). American Psychological Association.

American Psychological Association. (2020). *Publication Manual of the American Psychological Association* (7th ed.). American Psychological Association. https://doi.org/10.1037/0000165-000

Anvari, F., & Lakens, D. (2021). Using anchor-based methods to determine the smallest effect size of interest. *Journal of Experimental Social Psychology*, *96*, 104159. https://doi.org/10.1016/j.jesp.2021.104159

Arendt, F. (2019). Suicide on Instagram—Content analysis of a German suicide-related hashtag. *Crisis: The Journal of Crisis Intervention and Suicide Prevention*, *40*(1), 36–41. https://doi.org/10.1027/0227-5910/a000529

Arendt, F., & Northup, T. (2015). Effects of long-term exposure to news stereotypes on implicit and explicit attitudes. *International Journal of Communication*, *9*, 2370–2390.

Armstrong, E. A., Hamilton, L. T., Armstrong, E. M., & Seeley, J. L. (2014). "Good girls": Gender, social class, and slut discourse on campus. *Social Psychology Quarterly*, *77*(2), 100–122. https://doi.org/10.1177/0190272514521220

Ashcraft, A. M., & Murray, P. J. (2017). Talking to parents about adolescent sexuality. *Pediatric Clinics of North America*, *64*(2), 305. https://doi.org/10.1016/J.PCL.2016.11.002

Astle, S. M., & Anders, K. M. (2022). The relationship between topic-specific quality of parent–child sexual communication and measures of sexual self-concept and sexual subjectivity. *The Journal of Sex Research*, 1–13. https://doi.org/10.1080/00224499.2022.2081312

Astle, S. M., Anders, K., Shigeto, A., & Rodriguez, K. (2022). College women's memorable sexual messages from mothers, fathers, friends/peers, and online media: A mixed-methods latent class analysis. *Emerging Adulthood*, *11*(1), 1–15. https://doi.org/10.1177/21676968221098453

ATLAS.ti (Version 9). (2009). [Computer software]. ATLAS.ti Scientific Software Development GmbH. https://atlasti.com/

Aubrey, J. S., Yan, K., Terán, L., & Roberts, L. (2020). The heterosexual script on tween, teen, and young-adult television programs: A content analytic update and extension. *The Journal of Sex Research*, *57*(9), 1134–1145. https://doi.org/10.1080/00224499.2019.1699895

Axinn, W. G., Young-DeMarco, L., & Ro, M. C. (2011). Gender double standards in parenting attitudes. *Social Science Research*, *40*(2), 417–432. https://doi.org/10.1016/J.SSRESEARCH.2010.08.010

B

Baarsen, D. (n.d.). *Het grote verschil* [The big difference]. Baarsen. Retrieved January 9, 2023, from http://baarsen.com/nederlands.html

Bandura, A. (1977). *Social learning theory*. Englewood Cliffs, N.J. : Prentice Hall.

Bandura, A. (1986). *Social foundations of thought and action: A social cognitive theory*. Prentice-Hall.

Bandura, A., & Walters, R. H. (1963). *Social learning and personality development*. Holt Rinehart and Winston.

Barry, A. E., Szucs, L. E., Reyes, J. V., Ji, Q., Wilson, K. L., & Thompson, B. (2016). Failure to report effect sizes: The handling of quantitative results in published health education and behavior research. *Health Education and Behavior*, *43*(5), 518–527. https://doi.org/10.1177/1090198116669521

Bašaragin, M., & Savic, S. (2019). Gender perspective on teacher-pupil classroom interaction: Feedback and evaluation. *Slovenščina 2.0*, *7*(2), 172–196. https://doi.org/10.4312/SLO2.0.2019.2.172-196

Bassi, M., Blumberg, R. L., & Mateo Díaz, M. (2016). *Under the "cloak of invisibility": Gender bias in teaching practices and learning outcomes* (IDB Working Paper Series). Inter-American Development Bank. https://publications.iadb.org/en/under-cloak-invisibility-gender-bias-teaching-practices-and-learning-outcomes

Bay-Cheng, L. Y. (2015). The agency line: A neoliberal metric for appraising young women's sexuality. *Sex Roles*, *73*(7), 279–291. https://doi.org/10.1007/s11199-015-0452-6

Bem, S. L. (1983). Gender schema theory and its implications for child development: Raising gender- aschematic children in a gender-schematic society. *Journal of Women in Culture and Society*, *8*, 598–616. https://doi.org/10.1086/493998

Beukeboom, C. J. (2014). Mechanisms of linguistic bias: How words reflect and maintain stereotypic expectancies. In J. Forgas & L. Vincze (Eds.), *Social Cognition and Communication* (pp. 313–330). Psychology Press.

Biernat, M., Tocci, M. J., & Williams, J. C. (2012). The language of performance evaluations: Gender-based shifts in content and consistency of judgment. *Social Psychological and Personality Science*, *3*(2), 186–192. https://doi.org/10.1177/1948550611415693

Bigler, R. S., & Liben, L. S. (2006). A developmental intergroup theory of social stereotypes and prejudice. *Advances in Child Development and Behavior*, *34*, 39–89. https://doi.org/10.1016/S0065-2407(06)80004-2

Bigler, R. S., & Patterson, M. M. (2017). Social stereotyping and prejudice in children. In *The Wiley Handbook of Group Processes in Children and Adolescents*

(pp. 184–202). John Wiley & Sons, Ltd. https://doi.org/10.1002/9781118773123.ch9

Boe, J. L., & Woods, R. J. (2018). Parents' influence on infants' gender-typed toy preferences. *Sex Roles*, *79*(5), 358–373. https://doi.org/10.1007/s11199-017-0858-4

Boelhouwer, M., Brouwer, J., Scholte Lubberink, C., & Van Geert, P. (2014). *Seksuele ontwikkeling tussen 9 en 22 jaar: Meer dan van zoen tot het doen* [Sexual development between 9 and 22 years: More than going from kiss to doing it]. Vakgroep Ontwikkelingspsychologie, Faculteit Gedrags- en Maatschappijwetenschappen, Rijksuniversiteit Groningen. https://www.fwos.nl/wp-content/uploads/2015/01/30-definitieve-eindrapport-Van-Geert-dec-2014.pdf

Boik, R. J. (1981). A priori tests in repeated measures designs: Effects of nonsphericity. *Psychometrika*, *46*(3), 241–255. https://doi.org/10.1007/BF02293733

Bolles, R., & Messick, S. (1958). Statistical utility in experimental inference. *Psychological Reports*, *4*(3), 223–227. https://doi.org/10.2466/PR0.4.3.223-227

Booij, C., Jansen, E., & Schaik, E. van. (2013). Long term, interrelated interventions to increase women's participation in STEM in the Netherlands. *International Journal of Gender, Science and Technology*, *5*(3), 305–316.

Bordini, S. G., & Sperb, T. M. (2013). Sexual double standard: A review of the literature between 2001 and 2010. *Sexuality & Culture*, *17*(4), 686–704. https://doi.org/10.1007/S12119-012-9163-0

Bosson, J. K., Wilkerson, M., Kosakowska-Berezecka, N., Jurek, P., & Olech, M. (2022). Harder won and easier lost? Testing the double standard in gender rules in 62 countries. *Sex Roles*, *87*(1), 1–19. https://doi.org/10.1007/s11199-022-01297-y

Boterman, W., Musterd, S., Pacchi, C., & Ranci, C. (2019). School segregation in contemporary cities: Socio-spatial dynamics, institutional context and urban outcomes. *Urban Studies*, *56*(15), 3055–3073. https://doi.org/10.1177/0042098019868377

Breuer, J., Kowert, R., Festl, R., & Quandt, T. (2015). Sexist games=sexist gamers? A longitudinal study on the relationship between video game use and sexist

attitudes. *Cyberpsychology, Behavior, and Social Networking*, *18*(4), 197–202. https://doi.org/10.1089/cyber.2014.0492

Bron, J., Loenen, S., Haverkamp, M., & Van Vliet, E. (2015). *Seksualiteit en seksuele diversiteit in de kerndoelen: Een leerplanvoorstel en voorbeeldlesmateriaal* [Sexuality and sexual diversity in the core goals: A proposed curriculum and examples of class materials]. SLO. https://www.slo.nl/@4273/seksualiteit/

Bronfenbrenner, U. (1977). Toward an experimental ecology of human development. *American Psychologist*, *32*(7), 513–531. https://doi.org/10.1037/0003-066X.32.7.513

Bronfenbrenner, U. (1979). *The ecology of human development: Experiments by nature and design*. Harvard University Press.

Brown, C. S., Mistry, R. S., & Yip, T. (2019). Moving from the margins to the mainstream: Equity and justice as key considerations for developmental science. *Child Development Perspectives*, *13*(4), 235–240. https://doi.org/10.1111/cdep.12340

Brown, C. S., & Stone, E. A. (2018). Environmental and social contributions to children's gender-typed toy play: The role of family, peers, and media. In E. S. Weisgram & L. M. Dinella (Eds.), *Gender typing of children's toys: How early play experiences impact development* (pp. 121–140). American Psychological Association. https://doi.org/10.1037/0000077-013

Bueno-Fernandes, A. C., & Campos-Pellanda, E. (2022). Gender stereotypes in TikTok and Instagram: A reverse engineering experiment for understanding the mechanisms of social network algorithms. *Universitas-XXI, Revista de Ciencias Sociales y Humanas*, *37*, 247–270. https://doi.org/10.17163/uni.n37.2022.10

Burge, P. L. (1981). Parental child-rearing sex-role attitudes related to social issue sex-role attitudes and selected demographic variables. *Home Economics Research Journal*, *9*(3), 193–199. http://doi.org/10.1177/1077727X810090030 2

Bussey, K., & Bandura, A. (1999). Social cognitive theory of gender development and differentiation. *Psychological Review*, *106*, 676–713. https://doi.org/10.1037/0033-295X.106.4.676

Byers, D. S., McInroy, L. B., Craig, S. L., Slates, S., & Kattari, S. K. (2020). Naming and addressing homophobic and transphobic microaggressions in social

work classrooms. *Journal of Social Work Education*, *56*(3), 484–495. https://doi.org/10.1080/10437797.2019.1656688

C

Callanan, V. J. (2012). Media consumption, perceptions of crime risk and fear of crime: Examining race/ethnic differences. *Sociological Perspectives*, *55*(1), 93–115. https://doi.org/10.1525/sop.2012.55.1.93

Cardinal, R. N., & Aitken, M. R. F. (2005). *ANOVA for the behavioral sciences researchers* (1st ed.). Psychology Press. https://doi.org/10.4324/9780203763933

Carlson, R. G., Fripp, J., Munyon, M. D., Daire, A., Johnson, J. M., & DeLorenzi, L. (2014). Examining Passive and Active Recruitment Methods for Low-Income Couples in Relationship Education. *Marriage & Family Review*, *50*(1), 76–91. https://doi.org/10.1080/01494929.2013.851055

Caron, S. L., Davis, C. M., Halteman, W. A., & Stickle, M. (1993). Predictors of condom-related behaviors among first-year college students. *Journal of Sex Research*, *30*(3), 252–259. https://doi.org/10.1080/00224499309551709

Carroll, R. M., & Nordholm, L. A. (1975). Sampling characteristics of Kelley's epsilon squared and Hays' omega squared. *Educational and Psychological Measurement*, *35*(3), 541–554. http://doi.org/10.1177/001316447503500304

Childs, T. M., & Wooten, N. R. (2023). Teacher bias matters: An integrative review of correlates, mechanisms, and consequences. *Race Ethnicity and Education*, *26*(3), 368–397. https://doi.org/10.1080/13613324.2022.2122425

Christina Aguilera, & Lil' Kim. (2002). Can't hold us down. In *Stripped*. RCA. https://genius.com/Christina-aguilera-cant-hold-us-down-lyrics

Cimpian, A. (2010). The impact of generic language about ability on children's achievement motivation. *Developmental Psychology*, *46*, 1333–1340. https://doi.org/10.1037/a0019665

Cohen, J. (1973). Eta-squared and partial eta-squared in fixed factor ANOVA designs. *Educational and Psychological Measurement*, *33*, 107–112. https://doi.org/10.1177/001316447303300111

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Lawrence Erlbaum Associates.

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist*, *49*(12), 997–1003. https://doi.org/10.1037/0003-066X.49.12.997

Connell, C., & Elliott, S. (2009). Beyond the birds and the bees: Learning inequality through sexuality education. *American Journal of Sexuality Education*, *4*, 83–102. https://doi.org/10.1080/15546120903001332

Connell, R. W. (2005). *Masculinities* (2nd ed.). Polity.

Coyne, S. M., Linder, J. R., Booth, M., Keenan-Kroff, S., Shawcroft, J. E., & Yang, C. (2021). Princess power: Longitudinal associations between engagement with princess culture in preschool and gender stereotypical behavior, body esteem, and hegemonic masculinity in early adolescence. *Child Development*, *92*(6), 2413–2430. https://doi.org/10.1111/cdev.13633

Coyne, S. M., Linder, J. R., Rasmussen, E. E., Nelson, D. A., & Birkbeck, V. (2016). Pretty as a princess: Longitudinal effects of engagement with disney princesses on gender stereotypes, body esteem, and prosocial behavior in children. *Child Development*, *87*(6), 1909–1925. https://doi.org/10.1111/cdev.12569

Coyne, S. M., Linder, J. R., Rasmussen, E. E., Nelson, D. A., & Collier, K. M. (2014). It's a bird! It's a plane! It's a gender stereotype!: Longitudinal associations between superhero viewing and gender stereotyped play. *Sex Roles*, *70*(9), 416–430. https://doi.org/10.1007/s11199-014-0374-8

Crawford, C. E., Demack, S., Gillborn, D., & Warmington, P. (2019). Quants and crits: Using numbers for social justice. In J. T. DeCuir-Gunby, T. K. Chapman, & P. A. Schutz (Eds.), *Understanding critical race research methods and methodologies* (pp. 125–137). Routledge.

Cromer, L. D., & Goldsmith, R. E. (2010a). Socio-cultural issues and child sexual abuse: Child sexual abuse myths: Attitudes, beliefs, and individual differences. *Journal of Child Sexual Abuse*, *19*(6), 618–647. https://doi.org/10.1080/10538712.2010.522493

Cromer, L. D., & Goldsmith, R. E. (2010b). Socio-cultural issues and child sexual abuse: Child sexual abuse myths: Attitudes, beliefs, and individual differences. *Journal of Child Sexual Abuse*, *19*(6), 618–647. https://doi.org/10.1080/10538712.2010.522493

Cumming, G., & Finch, S. (2001). A primer on the understanding, use, and calculation of confidence intervals that are based on central and noncentral distributions. *Educational and Psychological Measurement*, *61*(4), 532–574. https://doi.org/10.1177/0013164401614002

Currier, D. M. (2013). Strategic ambiguity: Protecting emphasized femininity and hegemonic masculinity in the hookup culture. *Gender & Society*, *27*(5), 704–727. https://doi.org/10.1177/0891243213493960

Dajches, L., Terán, L., Yan, K., & Aubrey, J. S. (2021). Not another teen show: Exploring the impact of sexual scripts in sexually oriented teenage television on adolescent girls' romantic relationship and sexual expectations. *Journal of Broadcasting & Electronic Media*, *65*(4), 575–594. https://doi.org/10.1080/08838151.2021.1981903

D

Darwin Holmes, A. G. (2020). Researcher positionality—A consideration of its influence and place in qualitative research—A new researcher guide. *Shanlax International Journal of Education*, *8*(4), 1–10. https://doi.org/10.34293/education.v8i4.3232

Das, S. (2022, August 6). How TikTok bombards young men with misogynistic videos. *The Observer*. https://www.theguardian.com/technology/2022/aug/06/revealed-how-tiktok-bombards-young-men-with-misogynistic-videos-andrew-tate

Daub, M. (2022, November 15). *Waarom de voltijdsbonus niet volstaat* [Why the fulltime bonus is not enough]. Het Potentieel Pakken. https://www.hetpotentieelpakken.nl/waarom-de-voltijdsbonus-niet-volstaat/

De Boer, H., Bosker, R. J., & van der Werf, M. P. C. (2010). Sustainability of teacher expectation bias effects on long-term student performance. *Journal of Educational Psychology*, *102*(1), 168–179. https://doi.org/10.1037/a0017289

De Graaf, H., Kruijer, H., Van Acker, J., & Meijer, S. (2012). *Seks onder je 25e: Seksuele gezondheid van jongeren in Nederland anno 2012* [Sex before your 25th: Sexual health of young Dutch people anno 2012]. Rutgers WPF/Soa aids Nederland.

De Graaf, H., Meijer, S., Poelman, J., & Vanwesenbeeck, I. (2005). *Seks onder je 25e: Seksuele gezondheid van jongeren in Nederland anno 2005* [Sex before your 25th: Sexual health of young Dutch people anno 2005]. Rutgers Nisso Groep/Soa Aids Nederland.

De Graaf, H., ter Schure, J., & van Liere, G. A. F. S. (2022). How old are young people when they start having sex? Unravelling the applicability of Cox

proportional hazards regression. *Journal of Public Health*, *30*(8), 1873–1880. https://doi.org/10.1007/s10389-021-01619-6

De Graaf, H., Van den Borne, M., Nikkelen, S., Twisk, D., & Meijer, S. (2017). *Seks onder je 25e: Seksuele gezondheid van jongeren in Nederland anno 2017* [Sex before your 25th: Sexual health of young Dutch people anno 2017]. Ebubron.

De Meyer, S., Kågesten, A., Mmari, K., McEachran, J., Chilet-Rosell, E., Kabiru, C. W. , Maina, B., Jerves, E. M., Currie, C., & Michielsen, K. (2017). "Boys should have the courage to ask a girl out": Gender norms in early adolescent romantic relationships. *Journal of Adolescent Health*, *61*(4, Supplement), S42–S47. https://doi.org/10.1016/j.jadohealth.2017.03.007

Degner, J., & Dalege, J. (2013). The apple does not fall far from the tree, or does it? A meta-analysis of parent–child similarity in intergroup attitudes. *Psychological Bulletin*, *139*, 1270–1304. https://doi.org/10.1037/a0031436

Denessen, E., Keller, A., van den Bergh, L., & van den Broek, P. (2020). Do teachers treat their students differently? An observational study on teacher-student interactions as a function of teacher expectations and student achievement. *Education Research International*, 1–8. https://doi.org/10.1155/2020/2471956

Dodd, D. H., & Schultz, R. F. (1973). Computational procedures for estimating magnitude. *Psychological Bulletin*, *79*(6), 391–395. https://doi.org/10.1037/h0034347

Dovidio, J. F., Gaertner, S. L., Ufkes, E. G., Saguy, T., & Pearson, A. R. (2016). Included but invisible? Subtle bias, common identity, and the darker side of "we": subtle bias, common identity, and the darker side of "we." *Social Issues and Policy Review*, *10*(1), 6–46. https://doi.org/10.1111/sipr.12017

Dutch Central Bureau of Statistics. (2022a). *Emancipatiemonitor* [Emancipation monitor]. Dutch Central Bureau of Statistics. https://www.cbs.nl/-/media/_pdf/2022/48/em2022_web.pdf

Dutch Central Bureau of Statistics. (2016, November 21). *Wie zijn de derde generatie?* https://www.cbs.nl/nl-nl/nieuws/2016/47/wie-zijn-de-derde-generatie-

Dutch Central Bureau of Statistics. (2022b). *Werkzame beroepsbevolking; beroep, 2022 3e kwartaal, 0131 Leidsters kinderopvang en onderwijsassistenten*. https://opendata.cbs.nl/?dl=7276C#/CBS/nl/dataset/85276NED/table?ts=1673448021532

Dutch Central Bureau of Statistics. (2022c, January 11). *Jeugdbescherming en de toeslagenaffaire*. Centraal Bureau voor de Statistiek. https://www.cbs.nl/nl-nl/longread/aanvullende-statistische-diensten/2022/jeugdbescherming-en-de-toeslagenaffaire

Dutch Central Bureau of Statistics. (2022d, November 14). *Man-vrouwloonverschil weer iets verder afgenomen*. Centraal Bureau voor de Statistiek. https://www.cbs.nl/nl-nl/nieuws/2022/46/man-vrouwloonverschil-weer-iets-verder-afgenomen

Duveen, G. (1996). The development of social representations of gender. *Japanese Journal of Experimental Social Psychology Jikken-Shakai-Shinrigaku-Kenkyu*, *35*(3), 1–7. https://doi.org/10.2130/jjesp.35.256

Dwyer, J. H. (1974). Analysis of variance and the magnitude of effects: A general approach. *Psychological Bulletin*, *81*(10), 731–737. https://doi.org/10.1037/h0036937

E

Eagly, A. H., & Wood, W. (1999). The origins of sex differences in human behavior: Evolved dispositions versus social roles. *American Psychologist*, *54*(6), 408–423. https://doi.org/10.1037/0003-066X.54.6.408

Eagly, A. H., Wood, W., & Diekman, A. B. (2000). Social role theory of sex differences and similarities: A current appraisal. In T. Eckes & H. M. Trautner (Eds.), *The developmental social psychology of gender* (pp. 123–174). Erlbaum.

Easteal, P., Holland, K., & Judd, K. (2015). Enduring themes and silences in media portrayals of violence against women. *Women's Studies International Forum*, *48*, 103–113. https://doi.org/10.1016/j.wsif.2014.10.015

Eidevald, C., Bergström, H., & Broström, A. W. (2018). Maneuvering suspicions of being a potential pedophile: Experiences of male ECEC-workers in sweden. *European Early Childhood Education Research Journal*, *26*(3), 407–417. https://doi.org/10.1080/1350293X.2018.1463907

Elliott, S. (2010). Parents' constructions of teen sexuality: Sex panics, contradictory discourses, and social inequality. *Symbolic Interaction*, *33*(2), 191–212. https://doi.org/10.1525/si.2010.33.2.191

Ellis, P. E. (2010). *The essential guide to effect sizes: Statistical power, meta-analysis, and the interpretation of research results*. Cambridge University Press. https://doi.org/10.1017/CBO9780511761676

Emmerink, P. M. J., van den Eijnden, R. J. J. M., ter Bogt, T. F. M., & Vanwesenbeeck, I. (2017). A scale for the assessment of sexual standards among youth: Psychometric properties. *Archives of Sexual Behavior*, *46*(6), 1699–1709. https://doi.org/10.1007/s10508-017-1001-x

Emmerink, P. M. J., Vanwezenbeeck, I., van den Eijnden, R. J. J. M., van Lankveld, J. J. D. M., & Ter Bogt, T. F. M. (2018). Een verkenning van de seksuele dubbele standaard bij adolescenten en jongvolwassenen in Nederland [An exploration of the sexual double standard among adolescents and young adults in the Netherlands]. *Tijdschrift Voor Seksuologie*, *42*(1), 8–13.

Endendijk, J. J., Deković, M., Vossen, H., van Baar, A. L., & Reitz, E. (2022). Sexual double standards: Contributions of sexual socialization by parents, peers, and the media. *Archives of Sexual Behavior*, *51*(3), 1721–1740. https://doi.org/10.1007/S10508-021-02088-4/TABLES/10

Endendijk, J. J., van Baar, A. L., & Deković, M. (2020). He is a stud, she is a slut! A meta-analysis on the continued existence of sexual double standards. *Personality and Social Psychology Review*, *24*(2), 163–190. https://doi.org/10.1177/1088868319891310

Evans, R., Widman, L., Kamke, K., & Stewart, J. L. (2020). Gender differences in parents' communication with their adolescent children about sexual risk and sex-positive topics. *The Journal of Sex Research*, *57*(2), 177–188. https://doi.org/10.1080/00224499.2019.1661345

F

Fahmy, M., & Ibrahim, F. (2022). How portrayal of child sexual abuse in the news media increases fear of crime among guardians: The case of greater Malé area, Maldives. *International Journal of Social Research & Innovation*, *6*(1), Article 1. https://doi.org/10.55712/ijsri.v6i1.51

Farris, S. R., & de Jong, S. (2014). Discontinuous intersections: Second-generation immigrant girls in transition from school to work. *Ethnic and Racial Studies*, *37*(9), 1505–1525. https://doi.org/10.1080/01419870.2013.774033

Farvid, P. (2010). The benefits of ambiguity: Methodological insights from researching "heterosexual casual sex." *Feminism & Psychology*, *20*(2), 232–237. https://doi.org/10.1177/0959353509359136

Farvid, P., & Braun, V. (2014). The "Sassy Woman" and the "Performing Man": Heterosexual casual sex advice and the (re)constitution of gendered

subjectivities. *Feminist Media Studies*, *14*(1), 118–134. https://doi.org/10.1080/14680777.2012.724027

Farvid, P., Braun, V., & Rowney, C. (2017). 'No girl wants to be called a slut!': Women, heterosexual casual sex and the sexual double standard. *Journal of Gender Studies*, *26*(5), 544–560. https://doi.org/10.1080/09589236.2016.1150818

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavioral Research Methods*, *39*(2), 175–191.

Fausto-Sterling, A. (2000). *Sexing the body: Gender politics and the construction of sexuality* (1st ed.). Basic Books.

Fausto-Sterling, A. (2012). The dynamic development of gender variability. *Journal of Homosexuality*, *59*(3), 398–421. https://doi.org/10.1080/00918369.2012.653310

Fazackerley, A. (2023, January 7). 'Vulnerable boys are drawn in': Schools fear spread of Andrew Tate's misogyny. *The Observer*. https://www.theguardian.com/society/2023/jan/07/andrew-tate-misogyny-schools-vulnerable-boys

Felten, H., & Broekroelofs, R. (2022). *Wat werkt bij het verminderen van discriminatie*. Kennisplatform Inclusief Samenleven. https://www.kis.nl/publicatie/wat-werkt-bij-het-verminderen-van-discriminatie

Fidler, F., & Thompson, B. (2001). Computing correct confidence intervals for ANOVA fixed-and random-effects effect sizes. *Educational and Psychological Measurement*, *61*(4), 575–604. https://doi.org/10.1177/0013164401614003

Field, A. (2017). *Discovering statistics using IBM SPSS statistics* (5th ed.). SAGE Edge.

Finch, W. H., & French, B. F. (2012). A comparison of methods for estimating confidence intervals for omega-squared effect size. *Educational and Psychological Measurement*, *72*(1), 68–77. https://doi.org/10.1177/0013164411406533

Fine, M. (1988). Sexuality, schooling, and adolescent females: The missing discourse of desire. *Harvard Educational Review*, *58*(1), 29–53. https://doi.org/10.17763/haer.58.1.u0468k1v2n2n8242

Fine, M., & McClelland, S. I. (2006). Sexuality education and desire: Still missing after all these years. *Educational Review*, *76*, 297–338. https://doi.org/10.17763/haer.76.3.w5042g23122n6703

Finley, J. R., & Benjamin, A. S. (2012). Adaptive and qualitative changes in encoding strategy with experience: Evidence from the test-expectancy paradigm. *Journal of Experimental Psychology: Learning Memory and Cognition*, *38*(3), 632–652. https://doi.org/10.1037/A0026215

Finley, J. R., Benjamin, A. S., Hays, M. J., Bjork, R. A., & Kornell, N. (2011). Benefits of accumulating versus diminishing cues in recall. *Journal of Memory and Language*, *64*(4), 289–298. https://doi.org/10.1016/J.JML.2011.01.006

Finley, J. R., & Naaz, F. (2023). Strategic use of internal and external memory in everyday life: Episodic, semantic, procedural, and prospective purposes. *Memory*, *31*(1), 108–126. https://doi.org/10.1080/09658211.2022.2126858

Finley, J. R., Roediger, H. L., Hughes, A. D., Wahlheim, C. N., & Jacoby, L. L. (2015). Simultaneous versus sequential presentation in testing recognition memory for faces. *American Journal of Psychology*, *128*(2), 173–195. https://doi.org/10.5406/AMERJPSYC.128.2.0173

Finley, J. R., Sungkhasettee, V. W., Roediger, H. L., & Balota, D. A. (2017). Relative contributions of semantic and phonological associates to over-additive false recall in hybrid DRM lists. *Journal of Memory and Language*, *93*, 154–168. https://doi.org/10.1016/J.JML.2016.07.006

Fisher, R. A. (1925). *Statistical methods for research workers* (1st ed.). Oliver and Boyd.

Fisher, R. A. (1928). The general sampling distribution of the multiple correlation coefficient. *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, *121*(788), 654–673. https://doi.org/10.1098/RSPA.1928.0224

Fjær, E. G., Pedersen, W., & Sandberg, S. (2015). "I'm not one of those girls": Boundary-work and the sexual double standard in a liberal hookup context. *Gender & Society*, *29*(6), 960–981. https://doi.org/10.1177/0891243215602107

Flores, D., & Barroso, J. (2017). 21st century parent–child sex communication in the United States: A process review. *Journal of Sex Research*, *54*(4–5), 532–548. https://doi.org/10.1080/00224499.2016.1267693

Foster, J., & Baker, J. (2022). Muscles, makeup, and femboys: Analyzing tiktok's "radical" masculinities. *Social Media + Society*, *8*(3), 1–14. https://doi.org/10.1177/20563051221126040

Frawley, T. (2005). Childhood education gender bias in the classroom: Current controversies and implications for teachers. *Childhood Education*, *81*(4), 221–227. https://doi.org/10.1080/00094056.2005.10522277

Freeman, N. (2007). Preschoolers' perceptions of gender appropriate toys and their parents' beliefs about genderized behaviors: Miscommunication, mixed messages, or hidden truths? *Early Childhood Education Journal*, *34*(5), 357–366. https://doi.org/10.1007/s10643-006-0123-x

Fritz, C. O., Morris, P. E., & Richler, J. J. (2012). Effect size estimates: Current use, calculations, and interpretation. *Journal of Experimental Psychology: General*, *141*(1), 2–18. https://doi.org/10.1037/a0024338

Funnekotter, B. (2021, October 1). The Lancet in problemen om lichamen met vagina's. *NRC*. https://www.nrc.nl/nieuws/2021/10/01/the-lancet-in-problemen-om-lichamen-met-vaginas-a4060361

G

Gaebelein, J. W., & Soderquist, D. R. (1978). The utility of within-subjects variables: Estimates of strength. *Educational and Psychological Measurement*, *38*(2), 351–360. https://doi.org/10.1177/001316447803800216

Gaito, J., & Firth, J. (1973). Procedures for estimating magnitude of effects. *The Journal of Psychology*, *83*(1), 151–161. https://doi.org/10.1080/00223980.1973.9915603

Gelman, S. A. (2009). Learning from others: Children's construction of concepts. *Annual Review of Psychology*, *60*(1), 115–140. https://doi.org/10.1146/annurev.psych.59.103006.093659

Gentrup, S., Lorenz, G., Kristen, C., & Kogan, I. (2020). Self-fulfilling prophecies in the classroom: Teacher expectations, teacher feedback and student achievement. *Learning and Instruction*, *66*. https://doi.org/10.1016/J.LEARNINSTRUC.2019.101296

Gerritsen, R. (2022, December 5). *Clitoris niet volledig afgebeeld in alle biologieboeken: "Vergroot taboe op vrouwelijke seksualiteit"* [Clitoris not fully pictured in all biology books: "Increases taboo on female sexuality"]. https://www.at5.nl/artikelen/217871/clitoris-niet-volledig-afgebeeld-in-alle-biologieboeken-vergroot-taboe-op-vrouwelijke-seksualiteit

Geven, S., Batruch, A., & Van de Werfhorst, H. G. (2018). *Inequality in teacher judgements, expectations and track recommendations: A review study* (pp. 1–78).

Amsterdam Institute for Social Science Research. https://zoek.officielebekendmakingen.nl/blg-864911

Gillborn, D., Warmington, P., & Demack, S. (2018). QuantCrit: Education, policy, 'Big Data' and principles for a critical race theory of statistics. *Race Ethnicity and Education*, *21*(2), 158–179. https://doi.org/10.1080/13613324.2017.1377417

Glass, G. V., & Hakstian, A. R. (1969). Measures of association in comparative experiments: Their development and interpretation. *American Educational Research Journal*, *6*(3), 403–414. https://doi.org/10.2307/1161859

Glock, S., & Klapproth, F. (2017). Bad boys, good girls? Implicit and explicit attitudes toward ethnic minority students among elementary and secondary school teachers. *Studies in Educational Evaluation*, *53*, 77–86. https://doi.org/10.1016/j.stueduc.2017.04.002

Glock, S., & Kleen, H. (2017). Gender and student misbehavior: Evidence from implicit and explicit measures. *Teaching and Teacher Education*, *67*, 93–103. https://doi.org/10.1016/J.TATE.2017.05.015

Goblet, M., Glowacz, F., Molina, V., & Tchounwou, P. B. (2021). Slut shaming in adolescence: A violence against girls and its impact on their health. *International Journal of Environmental Research and Public Health*, *18*(12), 1–15. https://doi.org/10.3390/IJERPH18126657

Gök, E., Aydın, B., & Weidman, J. C. (2019). The impact of higher education on unemployed Turkish people's attitudes toward gender: A multilevel analysis. *International Journal of Educational Development*, *66*, 155–163. https://doi.org/10.1016/j.ijedudev.2018.10.004

Good, T. L., & Brophy, J. E. (1997). *Looking in classrooms*. New York : Longman. http://archive.org/details/lookinginclassr000good

Good, T. L., Sterzinger, N., & Lavigne, A. (2018). Expectation effects: Pygmalion and the initial 20 years of research. *Educational Research and Evaluation*, *24*(3–5), 99–123. https://doi.org/10.1080/13803611.2018.1548817

Goodman, A., & Gatward, R. (2008). Who are we missing? Area deprivation and survey participation. *European Journal of Epidemiology*, *23*(6), 379–387. https://doi.org/10.1007/s10654-008-9248-0

Goos, P. (2010). Split-plot factorial design. In N. J. Salkind (Ed.), *Encyclopedia of Research Design* (pp. 1413–1418). SAGE Publications Inc. https://doi.org/doi.org/10.4135/9781412961288

Grant, C. A., & Agosto, V. (2008). Teacher capacity and social justice in teacher education. In *Handbook of research on teacher education* (3rd ed., pp. 175–200). Routledge.

Greenwald, A. G., Banaji, M. R., & Nosek, B. A. (2015). Statistically small effects of the Implicit Association Test can have societally large effects. *Journal of Personality and Social Psychology*, *108*, 553–561. https://doi.org/10.1037/pspa0000016

Grosholz, J., & Kubrin, C. (2007). Crime in the news: How crimes, offenders and victims are portrayed in the media. *Journal of Criminal Justice and Popular Culture*, *14*(1), 59–83.

Guskey, T. R. (2019). Grades versus comments: Research on student feedback. *Phi Delta Kappan*, *101*(3), 42–47. https://doi.org/10.1177/0031721719885920

H

Hallers-Haalboom, E. T., Groeneveld, M. G., van Berkel, S. R., Endendijk, J. J., van der Pol, L. D., Linting, M., Bakermans-Kranenburg, M. J., & Mesman, J. (2017). Mothers' and fathers' sensitivity with their two children: A longitudinal study from infancy to early childhood. *Developmental Psychology*, *53*(5), 860–872. https://doi.org/10.1037/dev0000293

Halman, L., Reeskens, T., Sieben, I., & van Zundert, M. (2022). *Atlas of European Values: Change and continuity in turbulent times* (Vol. 1). Open Press TiU. https://doi.org/10.26116/6p8v-tt12

Hamby, S. (2018, May 22). *Know thyself: How to write a reflexivity statement*. Psychology Today. https://www.psychologytoday.com/us/blog/the-web-violence/201805/know-thyself-how-write-reflexivity-statement

Hattie, J., & Peddie, R. (2003). School reports: "Praising with faint damns." *NZCER*, *29*(3), 4–9. https://doi.org/10.18296/set.0710

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, *77*(1), 81–112. https://doi.org/10.3102/003465430298487

Hays, W. L. (1963). *Statistics* (1st ed.). Holt, Rinehart and Winston.

Hendrickx, M., Mothupi, M. C., Cooper, D., De Meyer, S., Knight, L., Michielsen, K., & Tabana, H. (2022). Same but different? Comparing attitudes regarding gender, gender diversity, and sexual diversity among early adolescents in

South Africa and Belgium. *Journal of Adolescent Health*, *71*(4), 446–454. https://doi.org/10.1016/j.jadohealth.2022.05.006

Hensums, M., Overbeek, G., & Jorgensen, T. D. (2022). Not one sexual double standard but two? Adolescents' attitudes about appropriate sexual behavior. *Youth & Society*, *54*(1), 23–42. https://doi.org/10.1177/0044118X20957924

Hickey, G., & Kipping, C. (1996). A multi-stage approach to the coding of data from open-ended questions. *Nurse Researcher*, *4*(1), 81–91. https://doi.org/10.7748/NR.4.1.81.S9

Hird, M. J., & Jackson, S. (2001). Where 'angels' and 'wusses' fear to tread: Sexual coercion in adolescent dating relationships. *Journal of Sociology*, *37*(1), 27–43. https://doi.org/10.1177/144078301128756184

Hofhuis, S., & Pas, N. (2021, October 24). 'Woke' bedreigt de academische vrijheid ["Woke" threatens academic freedom]. *NRC*. https://www.nrc.nl/nieuws/2021/10/24/woke-bedreigt-de-academische-vrijheid-a4062889

Hollingsworth, H., Heard, J., & Weldon, P. (2019). *Communicating student learning progress: A review of student reporting in Australia*. Australian Council for Educational Research. https://research.acer.edu.au/cgi/viewcontent.cgi?article=1035&context=ar_misc

Holman, A., & Koenig Kellas, J. (2018). "Say something instead of nothing": Adolescents' perceptions of memorable conversations about sex-related topics with their parents. *Communication Monographs*, *85*(3), 357–379. https://doi.org/10.1080/03637751.2018.1426870

Hondius, D. (2009). Paradoxes of cultural recognition: Perspectives from Northern Europe. In S. Alghasi, T. Hylland Eriksen, & H. Ghorashi (Eds.), *Race and the Dutch: On the uneasiness surrounding racial issues in the Netherlands* (pp. 39–57). Routledge. https://doi.org/10.4324/9781315599380-8

Hong, J. S., & Garbarino, J. (2012). Risk and protective factors for homophobic bullying in schools: An application of the social–ecological framework. *Educational Psychology Review*, *24*(2), 271–285. https://doi.org/10.1007/s10648-012-9194-y

Hove, T., Paek, H.-J., Isaacson, T., & Cole, R. T. (2013). Newspaper portrayals of child abuse: Frequency of coverage and frames of the issue. *Mass Communication and Society*, *16*(1), 89–108. https://doi.org/10.1080/15205436.2011.632105

Huberty, C. J. (2002). A history of effect size indices. *Educational and Psychological Measurement*, *62*(2), 227–240. https://doi.org/10.1177/0013164402062002002

Huijnk, W., & van Beusekom, G. (2021). *Wat maakt het verschil?* [What makes a difference?]. Sociaal Cultureel Planbureau. https://www.scp.nl/publicaties/publicaties/2021/12/10/wat-maakt-het-verschil

Hyde, J. S. (2001). Reporting effect size: The role of editors, textbook authors, and publication manuals. *Educational and Psychological Measurement*, *61*(2), 225–228. https://doi.org/10.1177/00131640121971202

Hyde, J. S., Bigler, R. S., Joel, D., Tate, C. C., & van Anders, S. M. (2019). The future of sex and gender in psychology: Five challenges to the gender binary. *American Psychologist*, *74*(2), 171–193. https://doi.org/10.1037/amp0000307

Hyland, K. (2013). Student perceptions of hidden messages in teacher written feedback. *Studies in Educational Evaluation*, *39*(3), 180–187. https://doi.org/10.1016/j.stueduc.2013.06.003

Hyland, K., & Hyland, F. (2006). Feedback on second language students' writing. *Language Teaching*, *39*(2), 83–101. https://doi.org/10.1017/S0261444806003399

I

İnan-Kaya, G., & Rubie-Davies, C. M. (2021). Teacher classroom interactions and behaviours: Indications of bias. *Learning and Instruction*. https://doi.org/10.1016/J.LEARNINSTRUC.2021.101516

Inspectorate of Education. (2014). *De kwaliteit van het basisschooladvies*. Inspectorate of Education. https://www.onderwijsinspectie.nl/documenten/publicaties/2014/10/14/de-kwaliteit-van-het-basisschooladvies

Instituut voor Media Auditing. (2011). *HOI Kernrapport 1e kwartaal 2011* (pp. 1–57). Instituut voor Media Auditing.

Intravia, J., Wolff, K. T., Paez, R., & Gibbs, B. R. (2017). Investigating the relationship between social media consumption and fear of crime: A partial analysis of mostly young adults. *Computers in Human Behavior*, *77*(C), 158–168. https://doi.org/10.1016/j.chb.2017.08.047

J

Jellinek, M. S. (2010, May 1). *Don't let ADHD crush children's self-esteem*. MDedge Psychiatry. https://www.mdedge.com/psychiatry/article/23971/pediatrics/dont-let-adhd-crush-childrens-self-esteem

Jensen, S. Q. (2011). Othering, identity formation and agency. *Qualitative Studies*, *2*(2), Article 2. https://doi.org/10.7146/qs.v2i2.5510

Johnson-Grey, K. M., Boghrati, R., Wakslak, C. J., & Dehghani, M. (2020). Measuring abstract mind-sets through syntax: Automating the linguistic category model. *Social Psychological and Personality Science*, *11*(2), 217–225. https://doi.org/10.1177/1948550619848004

Jones, K. P., Peddie, C. I., Gilrane, V. L., King, E. B., & Gray, A. L. (2016). Not so subtle: A meta-analytic investigation of the correlates of subtle and overt discrimination. *Journal of Management*, *42*(6), 1588–1613. https://doi.org/10.1177/0149206313506466

Joshi, S. P., Peter, J., & Valkenburg, P. M. (2011). Scripts of sexual desire and danger in US and Dutch teen girl magazines: A cross-national content analysis. *Sex Roles*, *64*, 463–474. https://doi.org/10.1007/s11199-011-9941-4

Jussim, L. (2009). Teacher expectations and self-fulfilling prophecies. In K. Wentzel & A. Wigfield (Eds.), *The handbook of motivation at school* (pp. 349–380). Routledge.

K

Kågesten, A., Gibbs, S., Blum, R. W., Moreau, C., Chandra-Mouli, V., Herbert, A., & Amin, A. (2016). Understanding factors that shape gender attitudes in early adolescence globally: A mixed-methods systematic review. *PLOS ONE*, *11*(6), e0157805. https://doi.org/10.1371/JOURNAL.PONE.0157805

Kanne, P., & van der Schelde, A. (2022). *Hoe geëmancipeerd is de Nederlander anno 2022?* [How emancipated are the Dutch in 2022?]. I&O Research. https://www.ioresearch.nl/actueel/we-vinden-onszelf-erg-geemancipeerd-maar-gedragen-ons-traditioneel/

Kelley, K. (2005). The effects of nonnormal distributions on confidence intervals around the standardized mean difference: Bootstrap and parametric confidence intervals. *Educational and Psychological Measurement*, *65*(1), 51–69. https://doi.org/10.1177/0013164404264850

Kelley, K. (2007). Confidence intervals for standardized effect sizes: Theory, application, and implementation. *Journal of Statistical Software*, *20*(8), 1–24. https://doi.org/10.18637/JSS.V020.I08

Kelley, K. (2022). *MBESS: The MBESS R Package* (4.9.2) [Computer software]. https://CRAN.R-project.org/package=MBESS

Kelley, T. L. (1935). An unbiased correlation ratio measure. *Proceedings of the National Academy of Sciences of the United States of America*, *21*(9), 554–559. https://doi.org/10.1073/PNAS.21.9.554

Kennedy, M. (2020). 'If the rise of the TikTok dance and e-girl aesthetic has taught us anything, it's that teenage girls rule the internet right now': TikTok celebrity, girls and the Coronavirus crisis. *European Journal of Cultural Studies*, *23*(6), 1069–1076. https://doi.org/10.1177/1367549420945341

Keppel, G. (1991). *Design and analysis: A researcher's handbook* (3rd ed.). Prentice-Hall.

Keppel, G., & Wickens, T. D. (2004). *Design and analysis: A researcher's handbook* (4th ed.). Pearson Prentice Hall.

Keren, G., & Lewis, C. (1979). Partial omega squared for ANOVA designs. *Educational and Psychological Measurement*, *39*(1), 119–128. https://doi.org/10.1177/001316447903900116

Keselman, H. J. (1975). A Monte Carlo investigation of three estimates of treatment magnitude: Epsilon squared, eta squared, and omega squared. *Canadian Psychological Review/Psychologie Canadienne*, *16*(1), 44–48. https://doi.org/10.1037/H0081789

Kettrey, H. H. (2016). What's gender got to do with it? Sexual double standards and power in heterosexual college hookups. *The Journal of Sex Research*, *53*(7), 754–765. https://doi.org/10.1080/00224499.2016.1145181

Kim, J. L., Lynn Sorsoli, C., Collins, K., Zylbergold, B. A., Schooler, D., & Tolman, D. L. (2007). From sex to sexuality: Exposing the heterosexual script on primetime network television. *Journal of Sex Research*, *44*(2), 145–157. https://doi.org/10.1080/00224490701263660

Kirk, R. E. (1996). Practical significance: A concept whose time has come. *Educational and Psychological Measurement*, *56*(9), 746–759. https://doi.org/10.1177/0013164496056005002

Kirk, R. E. (2012). Research strategies and the control of nuisance variables. In *Experimental Design: Procedures for the Behavioral Sciences*. https://doi.org/10.4135/9781483384733.n1

Kleen, H., & Glock, S. (2018). The roles of teacher and student gender in German teachers' attitudes toward ethnic minority students. *Studies in Educational Evaluation*, *59*, 102–111. https://doi.org/10.1016/J.STUEDUC.2018.04.002

Koch, H. (2012, November 17). Groot onderzoek naar slachtoffers M. [Large investigation into the victims of M.]. *Trouw*. https://www.trouw.nl/nieuws/groot-onderzoek-naar-slachtoffers-m~b2fec2cc/

Koenig, A. M., & Eagly, A. H. (2014). Evidence for the social role theory of stereotype content: Observations of groups' roles shape stereotypes. *Journal of Personality and Social Psychology*, *107*(3), 371–392. https://doi.org/10.1037/a0037215

Kohlberg, L. (1966). A cognitive developmental analysis of children's sex role concepts and attitudes. In E. E. Maccoby (Ed.), *The development of sex differences* (pp. 82–172). Stanford University Press.

Kraska, M. (2010). Mixed model design. In N. J. Salkind (Ed.), *Encyclopedia of Research Design* (pp. 818–823). SAGE Publications Inc. https://doi.org/10.4135/9781412961288

Kreager, D. A., & Staff, J. (2009). The sexual double standard and adolescent peer acceptance. *Social Psychology Quarterly*, *72*(2), 143–164.

Kreager, D. A., Staff, J., Gauthier, R., Lefkowitz, E. S., & Feinberg, M. E. (2016). The double standard at sexual debut: Gender, sexual behavior and adolescent peer acceptance. *Sex Roles*, *75*(7–8), 377–392. https://doi.org/10.1007/S11199-016-0618-X/TABLES/4

Krippendorff, K. (2012). *Content Analysis: An Introduction to Its Methodology*. Sage publications.

Kroneman, M., Admiraal, W., & Ketelaars, M. (2018). A peer–educator intervention: Attitudes towards LGB in prevocational secondary education in the Netherlands. *Journal of LGBT Youth*, *16*(1), 62–82. https://doi.org/10.1080/19361653.2018.1531101

Kuhle, B. X., Melzer, D. K., Cooper, C. A., Merkle, A. J., Pepe, N. A., Ribanovic, A., Verdesco, A. L., & Wettstein, T. L. (2015). The "birds and the bees" differ for boys and girls: Sex differences in the nature of sex talks. *Evolutionary Behavioral Sciences*, *9*(2), 107–115. https://doi.org/10.1037/ebs0000012

L

Laan, E. T. M., Klein, V., Werner, M. A., van Lunsen, R. H. W., & Janssen, E. (2021). In pursuit of pleasure: A biopsychosocial perspective on sexual pleasure and gender. *International Journal of Sexual Health*, *33*(4), 516–536. https://doi.org/10.1080/19317611.2021.1965689

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and ANOVAs. *Frontiers in Psychology*, *4*, 1–12. https://doi.org/10.3389/fpsyg.2013.00863

Lakens, D. (2015). *The 20% Statistician: Why you should use omega-squared instead of eta-squared.* http://daniellakens.blogspot.com/2015/06/why-you-should-use-omega-squared.html

Lakens, D. [@lakens]. (2022, May 12). *In most fields, people have not even started to consider the extremely important question which effect sizes matter. So your contribution will often be the first step on this challenging question in your specific research area.* [Tweet]. Twitter. https://twitter.com/lakens/status/1524799949015326722

Lam, E., Partridge, S. R., & Allman-Farinelli, M. (2016). Strategies for successful recruitment of young adults to healthy lifestyle programmes for the prevention of weight gain: A systematic review. *Obesity Reviews*, *17*(2), 178–200. https://doi.org/10.1111/obr.12350

Landelijk Aktie Komitee Scholieren. (2023). LAKS Taalgids 2023 [LAKS (National Action Committee School Students) Language guide]. *Landelijk Aktie Komitee Scholieren*. https://www.laks.nl/projecten/laks-taalgids/

Leaper, C. (2011). More similarities than differences in contemporary theories of social development?: A plea for theory bridging. *Advances in Child Development and Behavior*, *40*, 337–378. https://doi.org/10.1016/b978-0-12-386491-8.00009-8

Leaper, C. (2015). Gender and social-cognitive development. In L. S. Liben & U. Muller (Eds.), *Handbook of child psychology and developmental science* (7th ed. , Vol. 2, pp. 806–853). Wiley.

Leaper, C., & Bigler, R. S. (2018). Societal causes and consequences of gender typing of children's toys. In E. S. Weisgram & L. M. Dinella (Eds.), *Gender typing of children's toys: How early play experiences impact development* (pp. 287–308).

American Psychological Association. https://doi.org/10.1037/0000077-013

Leaper, C., & Brown, C. S. (2018). Sexism in childhood and adolescence: Recent trends and advances in research. *Child Development Perspectives*, *12*(1), 10–15. https://doi.org/10.1111/cdep.12247

Lefkowitz, E. S., Shearer, C. L., Gillen, M. M., & Espinosa-Hernandez, G. (2014). How gendered attitudes relate to women's and men's sexual behaviors and beliefs. *Sexuality & Culture*, *18*(4), 833–846. https://doi.org/10.1007/S12119-014-9225-6

Levine, T. R., & Hullett, C. R. (2002). Eta squared, partial eta squared, and misreporting of effect size in communication research. *Human Communication Research*, *28*(4), 612–625. https://doi.org/10.1111/J.1468-2958.2002.TB00828.X

Levine, T. R., & Hullett, C. R. (2010). Partial eta-squared. In N. J. Salkind (Ed.), *Encyclopedia of Research Design* (pp. 1006–1008). SAGE Publications Inc. https://doi.org/10.4135/9781412961288

Levy, S. R., & Hughes, J. M. (2009). Development of racial and ethnic prejudice among children. In T. D. Nelson (Ed.), *Handbook of Prejudice, Stereotyping, and Discrimination* (pp. 23–42). Psychology Press.

Liben, L. S. (2016). We've come a long way, baby (but we're not there yet): Gender past, present, and future. *Child Development*, *87*(1), 5–28. https://doi.org/10.1111/cdev.12490

Liben, L. S. (2017). Gender development: A constructivist-ecological perspective. In E. Turiel, N. Budwig, & P. D. Zelazo (Eds.), *New Perspectives on Human Development* (pp. 145–164). Cambridge University Press. https://doi.org/10.1017/CBO9781316282755.010

Liben, L. S., & Bigler, R. S. (2002). The developmental course of gender differentiation: Conceptualizing, measuring, and evaluating constructs and pathways. *Monographs of the Society for Research in Child Development*, *67*(2), 1–187. https://doi.org/10.1111/1540-5834.t01-1-00187

Liu, X. S. (2022). Bias correction for eta squared in one-way ANOVA. *European Journal of Research Methods for the Behavioral and Social Sciences*, *18*(1), 44–57. https://doi.org/10.5964/METH.7745

Loftus, G. R., & Masson, M. E. J. (1994). Using confidence intervals in within-subject designs. *Psychonomic Bulletin & Review*, *1*(4), 476–490. https://doi.org/10.3758/BF03210951

Lowrie, M. (2018, March 21). Duclos on defensive as opposition mocks gender-neutral language policy. *The Globe and Mail*. https://www.theglobeandmail.com/canada/article-duclos-on-defensive-as-opposition-mocks-gender-neutral-language-policy/

Lunney, G. H. (1970). Using analysis of variance with a dichotomous dependent variable: An empirical study. *Journal of Educational Measurement*, *7*(4), 263–269. https://doi.org/10.1111/j.1745-3984.1970.tb00727.x

M

Maass, A., Karasawa, M., Politi, F., & Suga, S. (2006). Do verbs and adjectives play different roles in different cultures? A cross-linguistic analysis of person representation. *Journal of Personality and Social Psychology*, *90*, 734–750. https://doi.org/10.1037/0022-3514.90.5.734

Maass, A., Salvi, D., Arcuri, L., & Semin, G. (1989). Language use in intergroup contexts: The linguistic intergroup bias. *Journal of Personality and Social Psychology*, *57*(6), 981–993. https://doi.org/10.1037/0022-3514.57.6.981

Marks, M. J., & Fraley, R. C. (2007). The impact of social interaction on the sexual double standard. *Social Influence*, *2*, 29–54. https://doi.org/10.1080/15534510601154413

Marks, M. J., Young, T. M., & Zaikman, Y. (2019). The sexual double standard in the real world. *Social Psychology*, *50*(2), 67–79. https://doi.org/10.1027/1864-9335/a000362

Martin, C. L., Fabes, R. A., Hanish, L. D., Gaertner, B., Miller, C. F., Foster, S., & Updegraff, K. A. (2017). Using an intergroup contact approach to improve gender relationships. In *The Wiley Handbook of Group Processes in Children and Adolescents* (pp. 435–454). John Wiley & Sons, Ltd. https://doi.org/10.1002/9781118773123.ch21

Martinez, M. A., Gutierrez, B. C., Halim, M. L. D., & Leaper, C. (2021). Gender and ethnic variation in emerging adults' recalled dating socialization in relation to current romantic attitudes and relationship experiences.

Sexuality & Culture, *25*(6), 2208–2230. https://doi.org/10.1007/s12119-021-09873-2

Martínez-San Miguel, Y., & Tobias, S. (Eds.). (2016). *Trans studies: The challenge to hetero/homo normativities*. Rutgers University Press.

Mason, B. A., Gunersel, A. B., & Ney, E. A. (2014). Cultural and ethnic bias in teacher ratings of behavior: A criterion-focused review. *Psychology in the Schools*, *51*(10), 1017–1030. https://doi.org/10.1002/PITS.21800

Mathew, M. (2020, June 20). *What's New in SPSS Statistics 27*. IBM Data Science Community. https://community.ibm.com/community/user/datascience/blogs/mathew-mathew1/2020/06/10/whats-new-in-spss-statistics-27#proceduralEnhancements

Maxwell, S. E., Camp, C. J., & Arvey, R. D. (1981). Measures of strength of association: A comparative examination. *Journal of Applied Psychology*, *66*(5), 525–534. https://doi.org/10.1037/0021-9010.66.5.525

Maxwell, S. E., Delaney, H. D., & Kelley, K. (2018). *Designing experiments and analyzing data: A model comparison perspective* (3rd ed.). Routledge. https://doi.org/10.2307/2348301

Maykut, P., & Morehouse, R. (1994). *Beginning qualitative research: A philosophic and practical guide*. The Falmer Press. https://doi.org/10.4324/9780203485781

McMillan, J. H. (2019). Surveys of teachers' grading practices and perceptions. In T. R. Guskey & S. Brookhart (Eds.), *What we know about grading; what works, what doesn't, and what's next* (pp. 84–111). ASCD.

Memon, S., Umrani, S., & Pathan, H. (2017). Application of constant comparison method in social sciences: A useful technique to analyze interviews. *Grassroots*, *51*(1), 152–165.

Menegatti, M., Crocetti, E., & Rubini, M. (2017). Do gender and ethnicity make the difference? Linguistic evaluation bias in primary school. *Journal of Language and Social Psychology*, *36*(4), 415–437. https://doi.org/10.1177/0261927X17694980

Menegatti, M., & Rubini, M. (2017). Gender bias and sexism in language. *Oxford Research Encyclopedia of Communication*. https://doi.org/10.1093/ACREFORE/9780190228613.013.470

Mertens, S., D'Haenens, L., & De Cock, R. (2019). Online news consumption and public sentiment toward refugees: Is there a filter bubble at play? Belgium, France, the Netherlands, and Sweden: A comparison. In L. D'Haenens, W.

Joris, & F. Heinderyckx (Eds.), *Images of immigrants and refugees in western europe* (pp. 141–158). Leuven University Press.

Miller, K. R. (2012). *Influence of parental communication of sexual messages on late adolescent sexual assertiveness and sexual experience and the influence of adolescent adherence to the sexual double standard: An exploratory study* [Dissertation]. http://rave.ohiolink.edu/etdc/view?acc_num=osu1337093690

Minadeo, M., & Pope, L. (2022). Weight-normative messaging predominates on TikTok—A qualitative content analysis. *PLOS ONE*, *17*(11), e0267997. https://doi.org/10.1371/journal.pone.0267997

Ministerie van Onderwijs, Cultuur en Wetenschap. (2014). Wet van 11 december 2013 tot wijziging van onder meer de Wet op het primair onderwijs, de Wet op de expertisecentra en de Wet College voor examens in verband met de invoering van een centrale eindtoets, de invoering van een leerling- en onderwijsvolgsyste [Law of 11 December 2013 to change the Law on primary eduction, the Law on expert centres and the Law for examinations with regards to the introduction of the school leaver's test, the introduction of a pupil and education monitoring system and the introduction of funding regulations for minimum learning outcomes for schools for special primary education and schools for special education and special and secondary special education (school leaver's test and pupil and education tracking system for primary education)]. *Staatsblad van Het Koninkrijk Der Nederlanden*, *2014*(13). https://zoek.officielebekendmakingen.nl/stb-2014-13.html

Möller, J., van de Velde, R. N., Merten, L., & Puschmann, C. (2020). Explaining online news engagement based on browsing behavior: Creatures of habit? *Social Science Computer Review*, *38*(5), 616–632. https://doi.org/10.1177/0894439319828012

Morgenroth, T., & Ryan, M. K. (2018, March 28). *Gender in a social psychology context*. Oxford Research Encyclopedia of Psychology. https://doi.org/10.1093/acrefore/9780190236557.013.309

Muehlenhard, C. L., & McCoy, M. L. (1991). Double standard/double bind: The sexual double standard and women's communication about sex. *Psychology of Women Quarterly*, *15*, 447–461. https://doi.org/10.1111/j.1471-6402.1991.tb00420.x

Muehlenhard, C. L., & Quackenbush, D. M. (2013). Sexual double standard scale. In T. D. Fisher, C. M. Davis, & W. L. Yarber (Eds.), *Handbook of sexuality-related measures*. Routledge.

Munk, K. P., Larsen, L., Leander, E.-M. B., & Soerensen, K. (2013). Fear of child sex abuse: Consequences for childcare personnel in Denmark. *Nordic Psychology*, *65*(1), 19–32. https://doi.org/10.1080/19012276.2013.796081

Murray, A. H. (2018). Generation Snowflake? *RSA Journal*, *164*(4 (5576)), 44–47.

Musto, M. (2019). Brilliant or bad: The gendered social construction of exceptionalism in early adolescence. *American Sociological Review*, *84*(3), 369–393. https://doi.org/10.1177/0003122419837567

Mutz, D. C., & Goldman, S. K. (2010). Mass media. In J. F. Dovidio, M. Hewstone, P. Glick, & V. Esses, *The SAGE Handbook of Prejudice, Stereotyping and Discrimination* (pp. 241–258). SAGE Publications Ltd. https://doi.org/10.4135/9781446200919.n15

Myers, J. L., & Well, A. D. (2003). *Research design and statistical analysis* (2nd ed.). HarperCollins. https://doi.org/10.4324/9781410607034

Myers, J. L., & Well, A. D. (2010). *Research design and statistical analysis* (R. F. Lorch, Ed.; 3rd ed.). Routledge. https://doi.org/10.4324/9780203726631

Myhill, D., & Jones, S. (2006). "She doesn't shout at no girls": Pupils' perceptions of gender equity in the classroom. *Cambridge Journal of Education*, *36*(1), 99–113. https://doi.org/10.1080/03057640500491054

N

Natri, H., Abubakare, O., Beaud, F., Botha, M., Bottema-Beutel, K., Brown, L. X. Z., Burr, D. A., Cobbaert, L., Dabbs, C., Denome, D., Edwards, B., Flower, R. L., Gassner, D., Onaiwu, M. G., Grant, A., Hundt, A., Kapp, S., Tidd, J. L., Manalili, M. A. R., … Zisk, A. H. (2023). *Anti-ableist language is fully compatible with high-quality autism research: Response to Singer et al. (2022)*. PsyArXiv. https://doi.org/10.31234/osf.io/3vjnq

Nellis, A. M., & Savage, J. (2012). Does watching the news affect fear of terrorism? The importance of media exposure on terrorism fear. *Crime and Delinquency*, *58*(5), 748–768. https://doi.org/10.1177/0011128712452961

Nelson, A. A., & Brown, C. S. (2019). Too pretty for homework: Sexualized gender stereotypes predict academic attitudes for gender-typical early adolescent

girls. *The Journal of Early Adolescence*, *39*(4), 603–617. https://doi.org/10.1177/0272431618776132

Nelson, J. B. (2016). A robust function to return the cumulative density of non-central F distributions in Microsoft Office Excel. *Psicológica*, *37*(1), 61–83.

Ni, H., & Li, C. (2013). Cultural differences in teachers' narrative evaluations of students in non-academic areas: A study of school report cards of students from four ethnic backgrounds. *International Journal of Quantitative Research in Education*, *1*(2), 123. https://doi.org/10.1504/IJQRE.2013.056464

Nieto, S. (2000). Placing equity front and center: Some thoughts on transforming teacher education for a new century. *Journal of Teacher Education*, *51*(3), 180–187. https://doi.org/10.1177/0022487100051003004

Noorman, M. A. J., den Daas, C., & de Wit, J. B. F. (2022). How parents' ideals are offset by uncertainty and fears: A systematic review of the experiences of european parents regarding the sexual education of their children. *The Journal of Sex Research*. Advance online publication. https://doi.org/10.1080/00224499.2022.2064414

Norman, G. (2010). Likert scales, levels of measurement and the "laws" of statistics. *Advances in Health Science Education*, *15*, 625–632. https://doi.org/10.1007/s10459-010-9222-y

NOS. (2022, December 12). *Institutioneel racisme bij Buitenlandse Zaken, top biedt excuses aan*. https://nos.nl/artikel/2456127-institutioneel-racisme-bij-buitenlandse-zaken-top-biedt-excuses-aan

Nosek, B. A., Banaji, M. R., & Greenwald, A. G. (2002). Harvesting implicit group attitudes and beliefs from a demonstration web site. *Group Dynamics*, *6*(1), 101–115. https://doi.org/10.1037//1089-2699.6.1.101

O

OECD. (2018). *The Resilience of Students with an Immigrant Background*. OECD Publishing. https://doi.org/10.1787/9789264292093-en

Ogunyemi, D., Clare, C., Astudillo, Y. M., Marseille, M., Manu, E., & Kim, S. (2020). Microaggressions in the learning environment: A systematic review. *Journal of Diversity in Higher Education*, *13*(2), 97–119. https://doi.org/10.1037/dhe0000107

Okada, K. (2013). Is omega squared less biased? A comparison of three major effect size indices in one-way ANOVA. *Behaviormetrika*, *40*(2), 129–147. https://doi.org/10.2333/bhmk.40.129

Okada, K. (2017). Negative estimate of variance-accounted-for effect size: How often it is obtained, and what happens if it is treated as zero. *Behavior Research Methods*, *49*, 979–987.

Olejnik, S., & Algina, J. (2000). Measures of effect size for comparative studies: Applications, interpretations, and limitations. *Contemporary Educational Psychology*, *25*(3), 241–286. https://doi.org/10.1006/ceps.2000.1040

Olejnik, S., & Algina, J. (2003). Generalized eta and omega squared statistics: Measures of effect size for some common research designs. *Psychological Methods*, *8*(4), 434–447. https://doi.org/10.1037/1082-989X.8.4.434

Olsson, M., & Martiny, S. E. (2018). Does exposure to counterstereotypical role models influence girls' and women's gender stereotypes and career choices? A review of social psychological research. *Frontiers in Psychology*, *9*. https://www.frontiersin.org/articles/10.3389/fpsyg.2018.02264

Özdil, Z. (2014). "Racism is an American problem": Dutch exceptionalism and its politics of denial. *Frame*, *27*(2), 49–64.

P

Pantić, N., & Florian, L. (2015). Developing teachers as agents of inclusion and social justice. *Education Inquiry*, *6*(3), 27311. https://doi.org/10.3402/edui.v6.27311

Papp, L., Hagerman, C., Gnoleba, M., Erchull, M., Liss, M., Miles-McLean, H., & Robertson, C. (2015). Exploring perceptions of slut-shaming on facebook: Evidence for a reverse sexual double standard. *Gender Issues*, *32*(1), 57–76. https://doi.org/10.1007/s12147-014-9133-y

Pascoe, C. J. (2007). *Dude, you're a fag: Masculinity and sexuality in high school*. University of California Press.

Pearson, K. (1905). *On the general theory of skew correlation and non-linear regression*. Dulau and co.

Pearson, K. (1911). On a correction to be made to the correlation ratio eta. *Biometrika*, *8*(1/2), 254. https://doi.org/10.2307/2331454

Peng, J., Chen, L.-T., Chiang, H.-M., & Chiang, Y.-C. (2013). The impact of APA and AERA guidelines on effect size reporting. *Educational Psychology Review*, *25*(1). https://doi.org/10.1007/s10648-013-9218-2

Peri, P. (2018). Education and prejudice against immigrants. In L. Hagendoorn & S. Nekuee (Eds.), *Education and racism: A cross national inventory of positive effects of education on ethnic tolerance* (pp. 21–31). Routledge. https://doi.org/10.4324/9780429458101-2

Perry, C. L., Kelder, S. H., & Komro, K. A. (1993). The social world of adolescents: Family, peers, schools, and the community. In S. G. Millstein (Ed.), *Promoting the health of adolescents: New directions for the twenty-first century*. Oxford University Press.

Petersen, J. L., & Hyde, J. S. (2010). A meta-analytic review of research on gender differences in sexuality, 1993–2007. *Psychological Bulletin*, *136*(1), 21–38. https://doi.org/10.1037/a0017504

Pettigrew, T. F., & Tropp, L. R. (2008). How does intergroup contact reduce prejudice? Meta-analytic tests of three mediators. *European Journal of Social Psychology*, *38*(6), 922–934. https://doi.org/10.1002/ejsp.504

Pettigrew, T. F., Tropp, L. R., Wagner, U., & Christ, O. (2011). Recent advances in intergroup contact theory. *International Journal of Intercultural Relations*, *35*(3), 271–280. https://doi.org/10.1016/j.ijintrel.2011.03.001

Philips, D. C. (1995). The good, the bad, and the ugly: The many faces of constructivism. *Educational Researcher*, *24*(7), 5–12. https://doi.org/10.3102/0013189X024007005

Piaget, J. (1954). *The construction of reality in the child;* New York, Basic Books.

Popović, S. (2018a). Analysis of online child sexual abuse news comments: The role of media coverage in supporting attitudes about child sexual abuse and stereotypes against victims and perpetrators. *Child Maltreatment & Well-Being: Contemporary Issues, Research & Practice*, *2017*, 73–89.

Popović, S. (2018b). Child sexual abuse news: A systematic review of content analysis studies. *Journal of Child Sexual Abuse*, *27*(7), 752–777. https://doi.org/10.1080/10538712.2018.1486935

Prati, F., Menegatti, M., & Rubini, M. (2015). The beneficial role of multiple categorization and intergroup contact in reducing linguistic out-group derogation. *Journal of Language and Social Psychology*, *34*(5), 475–500. https://doi.org/10.1177/0261927X14567777

Prentice, D. A., & Miller, D. T. (1992). When small effects are impressive. *Psychological Bulletin*, *112*(1), 160–164. https://doi.org/10.1037/0033-2909.112.1.160

Prieto Curiel, R., Cresci, S., Muntean, C. I., & Bishop, S. R. (2020). Crime and its fear in social media. *Palgrave Communications*, *6*(1), 1–12. https://doi.org/10.1057/s41599-020-0430-7

Primbs, M. A., Pennington, C. R., Lakens, D., Silan, M. A. A., Lieck, D. S., Forscher, P. S., Buchanan, E. M., & Westwood, S. J. (2022). Are small effects the indispensable foundation for a cumulative psychological science? A reply to Götz et al., (2022). *Perspectives on Psychological Science*.

Q

Quinn, D. M. (2020). Experimental evidence on teachers' racial bias in student evaluation: The role of grading scales. *Educational Evaluation and Policy Analysis*, *42*(3), 375–392. https://doi.org/10.3102/0162373720932188

R

Reid, J. A., Elliott, S., & Webber, G. R. (2011). Double standard casual hookups to formal dates: Refining the boundaries of the sexual. *Gender & Society*, *25*(5), 545–568. https://doi.org/10.1177/0891243211418642

Reiss, I. L. (1956). The double standard in premarital sexual intercourse: A neglected concept. *Social Forces*, *34*(3), 224–230.

Reiss, I. L. (1964). The scaling of premarital sexual permissiveness. *Journal of Marriage and Family*, *26*(2), 188–198. https://doi.org/10.2307/349726

Rentzou, K. (2011). Greek parents' perceptions of male early childhood educators. *Early Years*, *31*(2), 135–147. https://doi.org/10.1080/09575146.2010.530247

Rios-Aguilar, C. (2014). The changing context of critical quantitative inquiry. *New Directions for Institutional Research*, *2013*(158), 95–107. https://doi.org/10.1002/ir.20048

Robson, C. (2011). *Real world research: A resource for users of social research methods in applied settings*. Chichester, West Sussex : Wiley.

Roeleveld, J., Driessen, G., Ledoux, G., & Cuppen, J. (2011). *Doelgroepleerlingen in het basisonderwijs* (857). Kohnstamm Instituut. https://pure.uva.nl/ws/files/1139844/116447_Doelgroepleerlingen_in_het_basisonderwijs.pdf

Rojek, A. E., Khanna, R., Yim, J. W. L., Gardner, R., Lisker, S., Hauer, K. E., Lucey, C. , & Sarkar, U. (2019). Differences in narrative language in evaluations of medical students by gender and under-represented minority status. *Journal of General Internal Medicine*, *34*(5), 684–691. https://doi.org/10.1007/s11606-019-04889-9

Rolfe, H. (2006). Where are the men? Gender segregation in the childcare and early years sector. *National Institute Economic Review*, *195*(1), 103–117. https://doi.org/10.1177/0027950106064038

Rose, A. (2022). "Dutch racism is not like anywhere else": Refusing color-blind myths in Black feminist otherwise spaces. *Gender & Society*, *36*(2), 239–263. https://doi.org/10.1177/08912432221075098

Rubie-Davies, C. M. (2015). *Becoming a high expectation teacher: Raising the bar*. Routledge Taylor & Francis Group.

Rubie-Davies, C. M., & Peterson, E. R. (2016). Relations between teachers' achievement, over- and underestimation, and students' beliefs for Māori and Pākehā students. *Contemporary Educational Psychology*, *47*, 72–83. https://doi.org/10.1016/J.CEDPSYCH.2016.01.001

Rubini, M., Moscatelli, S., Albarello, F., & Palmonari, A. (2007). Group power as a determinant of interdependence and intergroup discrimination. *European Journal of Social Psychology*, *37*(6), 1203–1221. https://doi.org/10.1002/EJSP.391

Rubini, M., Moscatelli, S., & Palmonari, A. (2007). Increasing group entitativity: Linguistic intergroup discrimination in the minimal group paradigm. *Group Processes and Intergroup Relations*, *10*(2), 280–296. https://doi.org/10.1177/1368430207075156

Runderkamp, Z., Pas, D. van der, Schotel, A. L., & Mügge, L. (2022). Space invaders and norm-politicians: How the media represent the intersectional identities of Members of Parliament. *European Political Science Review*, *14*(4), 566–582. https://doi.org/10.1017/S1755773922000339

Rutland, A., Killen, M., & Abrams, D. (2010). A new social-cognitive developmental perspective on prejudice: The interplay between morality and group identity. *Perspectives on Psychological Science*, *5*(3), 279–291. https://doi.org/10.1177/1745691610369468

Ryan, C. L., Patraw, J. M., & Bednar, M. (2013). Discussing princess boys and pregnant men: Teaching about gender diversity and transgender

experiences within an elementary school curriculum. *Journal of LGBT Youth*, *10*(1–2), 83–105. https://doi.org/10.1080/19361653.2012.718540

S

Sak, R., Rohrmann, T., Şahin Sak, İ. T., & Schauer, G. (2019). Parents' views on male ECEC workers: A cross-country comparison. *European Early Childhood Education Research Journal*, *27*(1), 68–80. https://doi.org/10.1080/1350293X.2018.1556535

Sanchez, D. T., Fetterolf, J. C., & Rudman, L. A. (2012). Eroticizing inequality in the United States: The consequences and determinants of traditional gender role adherence in intimate relationships. *The Journal of Sex Research*, *49*(2–3), 168–183. https://doi.org/10.1080/00224499.2011.653699

Schaafsma, D. (2022). A father's role in the sexual health education of their children: A scoping review. *Sex Education*, 1–16. https://doi.org/10.1080/14681811.2022.2084061

Schalet, A. (2000). Raging hormones, regulated love: Adolescent sexuality and the constitution of the modern individual in the united states and the Netherlands. *Body & Society*, *6*(1), 75–105.

Schalet, A. (2010). Sexual subjectivity revisited: The significance of relationships in Dutch and American girls' experiences of sexuality. *Gender & Society*, *24*(3), 304–329. https://doi.org/10.1177/0891243210368400

Schalet, A. (2011). *Not under my roof: Parents, teens, and the culture of sex*. The University of Chicago Press.

Schiappa, E., Hewes, D. E., & Gregg, P. B. (2006). Can one TV show make a difference? Will & Grace and the parasocial contact hypothesis. *Journal of Homosexuality*, *51*(4), 15–37. https://doi.org/10.1300/J082v51n04_02

Schmidt, S. C. E., & Woll, A. (2017). Longitudinal drop-out and weighting against its bias. *BMC Medical Research Methodology*, *17*(1), 1–11. https://doi.org/10.1186/s12874-017-0446-x

Schoel, C., Roessel, J., Jacobsen, A., & Stahlberg, D. (2014). Communication of stereotypes in the classroom: Biased language use of German and Turkish adolescents. *Journal of Applied Social Psychology*, *44*(4), 266–277. https://doi.org/10.1111/jasp.12213

Scull, T. M., Carl, A. E., Keefe, E. M., & Malik, C. V. (2021). Exploring parent-gender differences in parent and adolescent reports of the frequency, quality, and content of their sexual health communication. *The Journal of Sex Research*, 1–13. https://doi.org/10.1080/00224499.2021.1936439

Sechrest, L., & Yeaton, W. H. (1982). Magnitudes of experimental effects in social science research. *Evaluation Review*, *6*(5), 579–600. https://doi.org/10.1177/0193841X8200600501

Semin, G. R. (2000). Agenda 2000—communication: Language as an implementational device for cognition. *European Journal of Social Psychology*, *30*(5), 595–612. https://doi.org/10.1002/1099-0992(200009/10)30:5<595::AID-EJSP23>3.0.CO;2-A

Semin, G. R., & Fiedler, K. (1988). The cognitive functions of linguistic categories in describing persons: Social cognition and language. *Journal of Personality and Social Psychology*, *54*(4), 558–568. https://doi.org/10.1037/0022-3514.54.4.558

Shepherd, M. A. (2011). Effects of ethnicity and gender on teachers' evaluation of students' spoken responses. *Urban Education*, *46*(5), 1011–1028. https://doi.org/10.1177/0042085911400325

Shi, L., Roche, S. P., & McKenna, R. M. (2018). Media consumption and crime trend perceptions: A longitudinal analysis. *Deviant Behavior*. https://doi.org/10.1080/01639625.2018.1519129

Signorielli, N. (2009). Race and sex in prime time: A look at occupations and occupational prestige. *Mass Communication and Society*, *12*(3), 332–352. https://doi.org/10.1080/15205430802478693

Simon, W., & Gagnon, J. H. (1984). Sexual scripts. *Society*, *22*(1), 53–60. https://doi.org/10.1007/BF02701260

Simon, W., & Gagnon, J. H. (1986). Sexual scripts: Permanence and change. *Archives of Sexual Behavior*, *15*(2), 97–120. https://doi.org/10.1007/BF01542219

Simon, W., & Gagnon, J. H. (2003). Sexual scripts: Origins, influences and changes. *Qualitative Sociology*, *26*(4), 491–497. https://doi.org/10.1023/B:QUAS.0000005053.99846.e5

Singer, A., Lutz, A., Escher, J., & Halladay, A. (2023). A full semantic toolbox is essential for autism research and practice to thrive. *Autism Research*, *16*(3), 497–501. https://doi.org/10.1002/aur.2876

Slater, M. D. (2007). Reinforcing spirals: The mutual influence of media selectivity and media effects and their impact on individual behavior and social identity. *Communication Theory*, *17*(3), 281–303. https://doi.org/10.1111/j.1468-2885.2007.00296.x

Smit, K., Cense, M., & De Graaf, H. (2022). *Waar ligt de grens? Onderzoek naar opvattingen over seksueel (grensoverschrijdend) gedrag onder jongeren en hun ouders/verzorgers*. [Where is the line? Attitude research on sexual (unacceptable) behavior among adolescents and their parents/carers.] Rutgers Expertisecentrum seksualiteit.

Smith, G., Mysak, K., & Michael, S. (2008). Sexual double standards and sexually transmitted illnesses: Social rejection and stigmatization of women. *Sex Roles*, *58*(5), 391–401. https://doi.org/10.1007/s11199-007-9339-5

Smith, L. R., & Pegoraro, A. (2020). Media framing of Larry Nassar and the USA gymnastics child sex abuse scandal. *Journal of Child Sexual Abuse*, *29*(4), 373–392. https://doi.org/10.1080/10538712.2019.1703233

Smithson, M. (2001). Correct confidence intervals for various regression effect sizes and parameters: The importance of noncentral distributions in computing intervals. In *Educational and Psychological Measurement* (Vol. 61, Issue 4, pp. 605–632). https://doi.org/10.1177/00131640121971392

Solebello, N., & Elliott, S. (2011). "We want them to be as heterosexual as possible": Fathers talk about their teen children's sexuality. *Gender & Society*, *25*(3), 293–315. https://doi.org/10.1177/0891243211403926

Staiger, A. (2005). "Hoes can be hoed out, players can be played out, but pimp is for life"—The pimp phenomenon as strategy of identity formation. *Symbolic Interaction*, *28*(3), 407–428. https://doi.org/10.1525/si.2005.28.3.407

Standish, K. (2021). A coming wave: Suicide and gender after COVID-19. *Journal of Gender Studies*, *30*(1), 114–118. https://doi.org/10.1080/09589236.2020.1796608

Steiger, J. H. (2004). Beyond the F test: Effect size confidence intervals and tests of close fit in the analysis of variance and contrast analysis. *Psychological Methods*, *9*(2), 164–182. https://doi.org/10.1037/1082-989X.9.2.164

Steiger, J. H., & Fouladi, R. T. (1997). Noncentrality interval estimation and the evaluation of statistical models. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 221–257). Erlbaum.

Stevens, P. A. J., Clycq, N., Timmerman, C., & Van Houtte, M. (2011). Researching race/ethnicity and educational inequality in the Netherlands: A critical review of the research literature between 1980 and 2008. *British Educational Research Journal*, *37*(1), 5–43. https://doi.org/10.1080/01411920903342053

Stevens, P. A. J., Crul, M., Slootman, M. W., Clycq, N., & Timmerman, C. (2019). The Netherlands: From diversity celebration to a colorblind approach. In *The Palgrave Handbook of Race and Ethnic Inequalities in Education* (pp. 783–841). Springer International Publishing. https://doi.org/10.1007/978-3-319-94724-2_19

Steyer, I. (2014). Gender representations in children's media and their influence. *Campus-Wide Information Systems*, *31*(2/3), 171–180. https://doi.org/10.1108/CWIS-11-2013-0065

Sue, D. W. (2013). Race talk: The psychology of racial dialogues. *American Psychologist*, *68*, 663–672. https://doi.org/10.1037/a0033681

Sun, S., Pan, W., & Leigh Wang, L. (2006). A comprehensive review of effect size reporting and interpreting practices in academic journals in education and psychology purpose of NHST. *Association*, *102*(4), 989–1004. https://doi.org/10.1037/a0019507

T

Tabachnick, B. G., & Fidell, L. S. (2007a). *Experimental designs using ANOVA*. Duxbury.

Tabachnick, B. G., & Fidell, L. S. (2007b). *Using multivariate statistics*. Pearson Education.

Tamm, M. (2019). Fathers' parental leave-taking, childcare involvement and labor market participation. *Labour Economics*, *59*, 184–197. https://doi.org/10.1016/j.labeco.2019.04.007

Tanenbaum, L. (2000). *Slut! Growing up female with a bad reputation.* Harper Perennial.

Tenenbaum, H. R., & Ruck, M. D. (2007). Are teachers' expectations different for racial minority than for European American students? A meta-analysis. *Journal of Educational Psychology*, *99*(2), 253–273. https://doi.org/10.1037/0022-0663.99.2.253

The Economist. (2022, March 17). *The Economist's glass-ceiling index*. The Economist. https://www.economist.com/graphic-detail/glass-ceiling-index

Thompson, B. (1999a). If statistical significance tests are broken/misused, what practices should supplement or replace them? *Theory & Psychology*, *9*(2), 165–181. https://doi.org/10.1177/095935439992006

Thompson, B. (1999b). Statistical significance tests, effect size reporting and the vain pursuit of pseudo-objectivity. *Theory & Psychology*, *9*(2), 191–196. https://doi.org/10.1177/095935439992007

Thompson, B. (2007). Effect sizes, confidence intervals, and confidence intervals for effect sizes. *Psychology in the Schools*, *44*(5), 423–432. https://doi.org/10.1002/PITS.20234/

Timmermans, A. C., de Boer, H., Amsing, H. T. A., & Van der Werf, M. P. C. (2018). Track recommendation bias: Gender, migration background and SES bias over a 20-year period in the Dutch context. *British Educational Research Journal*, *44*(5), 847–874. https://doi.org/10.1002/berj.3470

Timmermans, A. C., de Boer, H., & Van der Werf, M. P. C. (2016). An investigation of the relationship between teachers' expectations and teachers' perceptions of student attributes. *Social Psychology of Education*, *19*(2), 217–240. https://doi.org/10.1007/s11218-015-9326-6

Tolman, D. L., Anderson, S. M., & Belmonte, K. (2015). Mobilizing metaphor: Considering complexities, contradictions, and contexts in adolescent girls' and young women's sexual agency. *Sex Roles*, *73*(7), 298–310. https://doi.org/10.1007/s11199-015-0510-0

Tolman, D. L., & McClelland, S. I. (2011). Normative sexuality development in adolescence: A decade in review, 2000-2009. *Journal of Research on Adolescence*, *21*(1), 242–255. https://doi.org/10.1111/j.1532-7795.2010.00726.x

Torney-Purta, J. V. (2006). *The development of political attitudes in children*. Routledge.

Tukachinsky, R., Mastro, D., & Yarchi, M. (2017). The effect of prime time television ethnic/racial stereotypes on Latino and Black Americans: A longitudinal national level study. *Journal of Broadcasting & Electronic Media*, *61*(3), 538–556. https://doi.org/10.1080/08838151.2017.1344669

Tukey, J. W. (1949). One degree of freedom for non-additivity. *Biometrics*, *5*(3), 232–242. https://doi.org/10.2307/3001938

Tuten, J. (2007). Report cards as a textual intersection: The perspectives of a teacher, parent, and student. *Teaching and Learning*, *22*(1), 5–21.

U

UWV. (2020). *Wie wil er meer werken, en wie minder?* [Who wants to work more, and who wants to work less?]. UWV. https://www.werk.nl/imagesdxa/wie-wil-er-meer-werken-en-wie-minder-maart2020_tcm95-425878.pdf

V

Van Baars, L. (2022, April 20). IEP en Cito worden deze week afgenomen: Meten eindtoetsen wel dezelfde reken- en taalkennis? [IEP and Cito are conducted this week: Do different school leavers's tests measure the same math and language skills?]. *Trouw*, 5.

Van Beusekom, G., & Kuyper, L. (2018). *LHBT-monitor 2018* [LGBT monitor 2018]. Sociaal Cultureel Planbureau. https://www.scp.nl/publicaties/monitors/2018/11/21/lhbt-monitor-2018

Van De Bongardt, D., Yu, R., Deković, M., & Meeus, W. (Eds.). (2017). *Romantic relationships and sexuality in adolescence and young adulthood: The role of parents peers and partners*. Routledge.

Van den Bergh, L., Denessen, E., Hornstra, L., Voeten, M., & Holland, R. W. (2010). The implicit prejudiced attitudes of teachers: Relations to teacher expectations and the ethnic achievement gap. *American Educational Research Journal*, *47*(2), 497–527. http://doi.org/10.3102/0002831209353594

Van der Aa, E. (2023, April 12). *Storm van kritiek op gids met ongewenste termen in het onderwijs: 'Taalpolitie' en 'wokegedram'* [Heaps of criticism in reaction to guide on undesirable terms in education: "language police" and "woke badgering"]. AD.nl. https://www.ad.nl/binnenland/storm-van-kritiek-op-gids-met-ongewenste-termen-in-het-onderwijs-taalpolitie-en-wokegedram~a274fbf2/

Van der Bunt, J., & Verheul, B. (2022, July 14). *Smartengeld voor gediscrimineerde gedupeerden in de Toeslagenaffaire* [Financial compensation for discriminated victims in the benefits scandal]. Nederlands Juristenblad. https://www.njb.nl/blogs/smartengeld-voor-gediscrimineerde-gedupeerden-in-de-toeslagenaffaire/

Van der Meer, T. G. L. A., Kroon, A. C., Verhoeven, P., & Jonkman, J. (2019). Mediatization and the disproportionate attention to negative news.

Journalism Studies, *20*(6), 783–803. https://doi.org/10.1080/1461670X.2018.1423632

Van Ewijk, R. (2011). Same work, lower grade? Student ethnicity and teachers' subjective assessments. *Economics of Education Review*, *30*(5), 1045–1058. https://doi.org/10.1016/j.econedurev.2011.05.008

Van Laer, K., & Janssens, M. (2011). Ethnic minority professionals' experiences with subtle discrimination in the workplace. *Human Relations*, *64*(9), 1203–1227. https://doi.org/10.1177/0018726711409263

Van Schie, G. (2018). Origins: A history of race-ethnic categorisation in the Dutch governmental data ontology (1899–2018). *TMG Journal for Media History*, *21*(2), 88. https://doi.org/10.18146/2213-7653.2018.367

Vaughan, G. M., & Corballis, M. C. (1969). Beyond tests of significance: Estimating strength of effects in selected ANOVA designs. *Psychological Bulletin*, *72*(3), 204–213. https://doi.org/10.1037/h0027878

Veldhuis, P., & Weeda, F. (2022, February 11). 'Wokeness' op de universiteit: Mag de docent nog beelden uit Aladdin gebruiken? ["Wokeness" at the university: Is the professor allowed to use images from Aladdin?]. *NRC*. https://www.nrc.nl/nieuws/2022/02/11/wokeness-op-de-universiteit-mag-de-docent-nog-beelden-uit-aladdin-gebruiken-a4088400

Vogt, W. P., & Johnson, R. B. (2015). *The SAGE dictionary of statistics & methodology: A nontechnical guide for the social sciences*. Sage.

Voyer, D., & Voyer, S. D. (2014). Gender differences in scholastic achievement: A meta-analysis. *Psychological Bulletin*, 1–31. https://doi.org/10.1037/a0036620

Vygotsky, L. S. (1978). *Mind in society: Development of higher psychological processes*. Harvard University Press.

W

Wakefield, M. A., Loken, B., & Hornik, R. C. (2010). Use of mass media campaigns to change health behaviour. *The Lancet*, *376*(9748), 1261–1271. https://doi.org/10.1016/S0140-6736(10)60809-4

Walsh, A., & Leaper, C. (2020). A content analysis of gender representations in preschool children's television. *Mass Communication and Society*, *23*(3), 331–355. https://doi.org/10.1080/15205436.2019.1664593

Wang, S., Rubie-Davies, C. M., & Meissel, K. (2018). A systematic review of the teacher expectation literature over the past 30 years. *Educational Research and Evaluation*, *24*(3–5), 124–179. https://doi.org/10.1080/13803611.2018.1548798

Watson, B., & Gallois, C. (2002). Patients' interactions with health providers: A linguistic category model approach. *Journal of Language and Social Psychology*, *21*(1), 32–52. https://doi.org/10.1177/0261927X02021001003

Weatherred, J. L. (2015). Child sexual abuse and the media: A literature review. *Journal of Child Sexual Abuse*, *24*(1), 16–34. https://doi.org/10.1080/10538712.2015.976302

Weatherred, J. L. (2017). Framing child sexual abuse: A longitudinal content analysis of newspaper and television coverage. *Journal of Child Sexual Abuse*, *26*(1), 3–22. https://doi.org/10.1080/10538712.2016.1257528

Weimann, G., & Masri, N. (2020). Research note: Spreading hate on TikTok. *Studies in Conflict & Terrorism*, 1–14. https://doi.org/10.1080/1057610X.2020.1780027

Weiner, M. F. (2015). Whitening a diverse Dutch classroom: White cultural discourses in an Amsterdam primary school. *Ethnic and Racial Studies*, *38*(2), 359–376. https://doi.org/10.1080/01419870.2014.894200

Weiner, M. F. (2016). Racialized classroom practices in a diverse Amsterdam primary school: The silencing, disparagement, and discipline of students of color. *Race Ethnicity and Education*, *19*(6), 1351–1367. https://doi.org/10.1080/13613324.2016.1195352

Weinstein, R. S., Gregory, A., & Strambler, M. J. (2004). Intractable self-fulfilling prophecies fifty years after brown v board of education. *American Psychologist*, *59*(6), 511–520. https://doi.org/10.1037/0003-066X.59.6.511

Weitzer, R., & Kubrin, C. E. (2004). *Breaking news: How local TV news and real-world conditions affect fear of crime*. *21*(1), 497–520. https://doi.org/10.1080/07418820400095881

Wekker, G. (2016). *White innocence: Paradoxes of colonialism and race*. Duke University Press.

Welten, L., & Abbas, T. (2022). *Islamophobia and securitisation*. Palgrave Macmillan. http://link.springer.com/book/10.1007/978-3-031-06205-6

Wesson, C. (2022). A mixed-methods exploration of young adults' views of the sexual double standard. *Sexuality & Culture*, *26*(4), 1314–1325. https://doi.org/10.1007/s12119-022-09944-y

Wiederman, M. W. (2005). The gendered nature of sexual scripts. *The Family Journal*, *13*(4), 496–502. https://doi.org/10.1177/1066480705278729

Wiederman, M. W. (2015). Sexual script theory: Past, present, and future. In J. DeLamater & R. F. Plante (Eds.), *Handbook of the Sociology of Sexualities* (pp. 7–22). Springer International Publishing. https://doi.org/10.1007/978-3-319-17341-2_2

Wigboldus, D. H. J., Semin, G. R., & Spears, R. (2000). How do we communicate stereotypes? Linguistic bases and inferential consequences. *Journal of Personality and Social Psychology*, *78*(1), 5–18. https://doi.org/10.1037/0022-3514.78.1.5

Wigboldus, D. H. J., Spears, R., & Semin, G. R. (2005). When do we communicate stereotypes? Influence of the social context on the linguistic expectancy bias. *Group Processes & Intergroup Relations*, *8*(3), 215–230. https://doi.org/10.1177/1368430205053939

Wilkins, A. C. (2012). Stigma and status: Interracial intimacy and intersectional identities among black college men. *Gender & Society*, *26*(2), 165–189. https://doi.org/10.1177/0891243211434613

Wilkinson, L. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, *54*(8), 594–604. https://doi.org/10.1037/0003-066X.54.8.594

Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., … Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, *3*(1), Article 1. https://doi.org/10.1038/sdata.2016.18

Wilson, B. J., Martins, N., & Marske, A. L. (2005). Children's and parents' fright reactions to kidnapping stories in the news. *Communication Monographs*, *72*(1), 46–70. https://doi.org/10.1080/0363775052000342526

Wilson, E. K., & Koo, H. P. (2010). Mothers, fathers, sons, and daughters: Gender differences in factors associated with parent-child communication about

sexual topics. *Reproductive Health*, *7*(1). https://doi.org/10.1186/1742-4755-7-31

Windzio, M., & Kleimann, M. (2009). Criminal society and criminal immigrants: A social construction of reality by mass media? In M. E. Oswald, S. Bieneck, & J. Hupfeld-Heinemann (Eds.), *Social Psychology of Punishment of Crime* (pp. 93–112). Wiley-Blackwell.

Winer, B. J. (1962). *Statistical principles in experimental design*. McGraw-Hill. https://doi.org/10.1037/11774-000

Wonneberger, A., Schoenbach, K., & Van Meurs, L. (2012). Staying Tuned: TV News Audiences in the Netherlands 1988-2010. *Journal of Broadcasting and Electronic Media*, *56*(1), 55–74. https://doi.org/10.1080/08838151.2011.648684

World Economic Forum. (2021). *Global gender gap report* [Insight report]. World Economic Forum. https://www.weforum.org/reports/global-gender-gap-report-2021/

World Health Organization. (2002). *Defining sexual health: Report of a technical consultation on sexual health*. World Health Organization. https://www.cesas.lu/perch/resources/whodefiningsexualhealth.pdf

Wouters, C. (2013). Dating in de VS - verloving en verkering in Nederland [Dating in the US – engagement and *verkering* in the Netherlands]. *Sociologie*, *9*(3/4), 378–405. https://doi.org/10.5117/SOC2013.3.WOUT

Wu, H., Luo, S., Espinosa-Hernández, G., Klettner, A., White, T. D., & Li, H. (2021). Relating gender to sex: Gendered attitudes, sexual double standard, sexual intentions and behaviors in two chinese adolescent samples. *The Journal of Sex Research*, *58*(1), 29–40. https://doi.org/10.1080/00224499.2019.1695098

Y

Yigit, S., & Mendes, M. (2018). Which effect size measure is appropriate for one-way and two-way ANOVA-models? A Monte Carlo simulation study. *RevStat Statistical Journal*, *16*(3), 295–313.

Young, M., Cardenas, S., Donnelly, J., & J. Kittleson, M. (2016). Perceptions of Peer Sexual Behavior: Do Adolescents Believe in a Sexual Double Standard? *Journal of School Health*, *86*(12), 855–863. https://doi.org/10.1111/josh.12455

Z

Zaikman, Y., & Marks, M. J. (2017). Promoting theory-based perspectives in sexual double standard research. *Sex Roles*, *76*, 407–420. https://doi.org/10.1007/s11199-016-0677-z

Zambarano, R. (1992). Testing the additivity assumption in repeated measures ANOVA. *South-Central SAS Users Group*, 375–385.

Zhou, Y., & Skidmore, S. T. (2017). A reassessment of ANOVA reporting practices: A review of three APA journals. *Journal of Methods and Measurement in the Social Sciences*, *8*(1), 3–19. https://doi.org/10.2458/v8i1.22019

Appendices

Appendix A – Chapter 2. Disambiguation Table

Table A

Comparison of the Terms Used for ANOVA Components Across Research Literature

Concept	Current paper	D&S	K&W	M,D&K	O&A
Number of levels of factor A, B, C	a, b, c	J, K, L	a, b, c	a, b, c	J, K, L
One-way between-subjects design (A)					
Factor A (variance due to treatment A)	effect	between	A	between, B	effect
Error (variance within groups)	error	within	S/A	within, W	error
Multi-factor between-subjects design (A×B)					
Factor A or B (variance due to treatment A or B)	effect	effect, A, B	A, B	effect	effect
Factor A×B (variance due to interaction A×B)	effect	Effect, AB	A×B	effect	effect
Error (variance within groups)	error	error	S/AB	W	error
One-way within-subjects design ($\overline{A}$×S)					
Factor $\overline{A}$ (variance due to treatment $\overline{A}$)	effect	effect	A	A	effect
Factor $\overline{A}$×S (variance due to random error and interaction $\overline{A}$×S)	effect × subject	A×S	A×S	A×S	error
Error (variance due to subjects)	subject	S	S	S	subjects
Multi-factor within-subjects design ($\overline{A}$×$\overline{B}$×S)					
Factor $\overline{A}$ or $\overline{B}$ (variance due to treatment $\overline{A}$ or $\overline{B}$)	effect	effect	A, B	effect	-
Factor $\overline{A}$×$\overline{B}$ (variance due to interaction $\overline{A}$×$\overline{B}$)	effect	Effect, AB	A×B	effect	-
Error due to $\overline{A}$×S (variance due to random error and interaction $\overline{A}$×S)	effect × subject	A×S	A×S	effect × S	-
Error (variance due to subjects)	subject	S	S	S	-

Table A (continued)

Comparison of the Terms Used for ANOVA Components Across Research Literature

Concept	Current paper	D&S	K&W	M,D&K	O&A
Split-plot design (A×$\overline{B}$×S)					
Factor A or $\overline{B}$ (variance due to treatment A or $\bar{B}$)	effect	effect	A, B	effect	A, B
Factor $\overline{A}$×$\overline{B}$ (variance due to interaction $\overline{A}$×$\overline{B}$)	effect	Effect, AB	A×B	effect	AB
Error due to A (variance due to subjects within groups of A)	subject/A	S	S/A	S/A	S/A
Error due to $\bar{B}$×S (variance due to random error and interaction $\overline{B}$×S)	effect ×subject/A	B×S	B×S/A	B×S/A	BS/A

Note. Components that are used in formulas for omega squared and/or in ANOVA tables. D&S = Dodd & Schultz, 1973; K&W = Keppel & Wickens, 2004; M,D&K = Maxwell, Delaney & Kelley, 2018; O&A = Olejnik & Algina, 2000.

Appendix B – Chapter 2. Quick Formula Reference for $\widehat{\omega}^2$

Table B

Quick Reference for Omega Squared Formulas

Design	Factor	Standard Omega Squared $\widehat{\omega}^2$	Partial Omega Squared $\widehat{\omega}_p^2$
1-way Between-subjects	A	$\frac{SS_A - (df_A \times MS_{error})}{SS_{total} + MS_{error}}$	
1-way Within-subjects	$\overline{A}$	$\frac{SS_A - (df_A \times MS_{A\times subject})}{SS_{total} + MS_{subject}}$	
2-way Between-subjects	A	$\frac{SS_A - (df_A \times MS_{error})}{SS_{total} + MS_{error}}$	$\frac{SS_A - (df_A \times MS_{error})}{SS_A + (N - df_A)MS_{error}}$
	B	$\frac{SS_B - (df_B \times MS_{error})}{SS_{total} + MS_{error}}$	$\frac{SS_B - (df_B \times MS_{error})}{SS_B + (N - df_B)MS_{error}}$
	AB	$\frac{SS_{AB} - (df_{AB} \times MS_{error})}{SS_{total} + MS_{error}}$	$\frac{SS_{AB} - (df_{AB} \times MS_{error})}{SS_{AB} + (N - df_{AB})MS_{error}}$

Table B (continued)

Quick Reference for Omega Squared Formulas

Design	Factor	Standard Omega Squared $\hat{\omega}^2$	Partial Omega Squared $\hat{\omega}_p^2$
2-way Within-subjects	$\overline{A}$	$\frac{SS_A - (df_A \times MS_{A\times subject})}{SS_{total} + MS_{subject}}$	$\frac{SS_A - (df_A \times MS_{A\times subject})}{SS_A + SS_{A\times subject} + SS_{subject} + MS_{subject}}$
	$\overline{B}$	$\frac{SS_B - (df_B \times MS_{B\times subject})}{SS_{total} + MS_{subject}}$	$\frac{SS_B - (df_B \times MS_{B\times subject})}{SS_B + SS_{B\times subject} + SS_{subject} + MS_{subject}}$
	$\overline{AB}$	$\frac{SS_{AB} - (df_{AB} \times MS_{AB\times subject})}{SS_{total} + MS_{subject}}$	$\frac{SS_{AB} - (df_{AB} \times MS_{AB\times subject})}{SS_{AB} + SS_{AB\times subject} + SS_{subject} + MS_{subject}}$
2-way Split-plot	A	$\frac{SS_A - (df_A \times MS_{subject/A})}{SS_{total} + MS_{subject/A}}$	$\frac{SS_A - (df_A \times MS_{subject/A})}{SS_A + SS_{subject/A} + MS_{subject/A}}$
	$\overline{B}$	$\frac{SS_B - (df_B \times MS_{B\times subject/A})}{SS_{total} + MS_{subject/A}}$	$\frac{SS_B - (df_B \times MS_{B\times subject/A})}{SS_B + SS_{B\times subject/A} + SS_{subject/A} + MS_{subject/A}}$
	$A\overline{B}$	$\frac{SS_{AB} - (df_{AB} \times MS_{B\times subject/A})}{SS_{total} + MS_{subject/A}}$	$\frac{SS_{AB} - (df_{AB} \times MS_{B\times subject/A})}{SS_{AB} + SS_{AB\times subject/A} + SS_{subject/A} + MS_{subject/A}}$

Table B (continued)

Quick Reference for Omega Squared Formulas

Design	Factor	Standard Omega Squared $\hat{\omega}^2$	Partial Omega Squared $\hat{\omega}_p^2$
3-way Between-subjects	A	$\frac{SS_A - (df_A \times MS_{error})}{SS_{total} + MS_{error}}$	$\frac{SS_A - (df_A \times MS_{error})}{SS_A + (N - df_A)MS_{error}}$
	B	$\frac{SS_B - (df_B \times MS_{error})}{SS_{total} + MS_{error}}$	$\frac{SS_B - (df_B \times MS_{error})}{SS_B + (N - df_B)MS_{error}}$
	C	$\frac{SS_C - (df_C \times MS_{error})}{SS_{total} + MS_{error}}$	$\frac{SS_C - (df_C \times MS_{error})}{SS_C + (N - df_C)MS_{error}}$
	AB	$\frac{SS_{AB} - (df_{AB} \times MS_{error})}{SS_{total} + MS_{error}}$	$\frac{SS_{AB} - (df_{AB} \times MS_{error})}{SS_{AB} + (N - df_{AB})MS_{error}}$
	AC	$\frac{SS_{AC} - (df_{AC} \times MS_{error})}{SS_{total} + MS_{error}}$	$\frac{SS_{AC} - (df_{AC} \times MS_{error})}{SS_{AC} + (N - df_{AC})MS_{error}}$
	BC	$\frac{SS_{BC} - (df_{BC} \times MS_{error})}{SS_{total} + MS_{error}}$	$\frac{SS_{BC} - (df_{BC} \times MS_{error})}{SS_{BC} + (N - df_{BC})MS_{error}}$
	ABC	$\frac{SS_{ABC} - (df_{ABC} \times MS_{error})}{SS_{total} + MS_{error}}$	$\frac{SS_{ABC} - (df_{ABC} \times MS_{error})}{SS_{ABC} + (N - df_{ABC})MS_{error}}$

Table B (continued)

Quick Reference for Omega Squared Formulas

Design	Factor	Standard Omega Squared $\hat{\omega}^2$	Partial Omega Squared $\hat{\omega}_p^2$
3-way Within-subjects	$\overline{A}$	$\frac{SS_A - (df_A \times MS_{A\times subject})}{SS_{total} + MS_{subject}}$	$\frac{SS_A - (df_A \times MS_{A\times subject})}{SS_A + SS_{A\times subject} + SS_{subject} + MS_{subject}}$
	$\overline{B}$	$\frac{SS_B - (df_B \times MS_{B\times subject})}{SS_{total} + MS_{subject}}$	$\frac{SS_B - (df_B \times MS_{B\times subject})}{SS_B + SS_{B\times subject} + SS_{subject} + MS_{subject}}$
	$\overline{C}$	$\frac{SS_C - (df_C \times MS_{C\times subject})}{SS_{total} + MS_{subject}}$	$\frac{SS_C - (df_C \times MS_{C\times subject})}{SS_C + SS_{C\times subject} + SS_{subject} + MS_{subject}}$
	$\overline{AB}$	$\frac{SS_{AB} - (df_{AB} \times MS_{AB\times subject})}{SS_{total} + MS_{subject}}$	$\frac{SS_{AB} - (df_{AB} \times MS_{AB\times subject})}{SS_{AB} + SS_{AB\times subject} + SS_{subject} + MS_{subject}}$
	$\overline{AC}$	$\frac{SS_{AC} - (df_{AC} \times MS_{AC\times subject})}{SS_{total} + MS_{subject}}$	$\frac{SS_{AC} - (df_{AC} \times MS_{AC\times subject})}{SS_{AC} + SS_{AC\times subject} + SS_{subject} + MS_{subject}}$
	$\overline{BC}$	$\frac{SS_{BC} - (df_{BC} \times MS_{BC\times subject})}{SS_{total} + MS_{subject}}$	$\frac{SS_{BC} - (df_{BC} \times MS_{BC\times subject})}{SS_{BC} + SS_{BC\times subject} + SS_{subject} + MS_{subject}}$
	$\overline{ABC}$	$\frac{SS_{ABC} - (df_{ABC} \times MS_{ABC\times subject})}{SS_{total} + MS_{subject}}$	$\frac{SS_{ABC} - (df_{ABC} \times MS_{ABC\times subject})}{SS_{ABC} + SS_{ABC\times subject} + SS_{subject} + MS_{subject}}$

Table B (continued)

Quick Reference for Omega Squared Formulas

Design	Factor	Standard Omega Squared $\widehat{\omega}^2$	Partial Omega Squared $\widehat{\omega}_p^2$
3-way Split-plot (2 Between-subjects Factors)	A	$\frac{SS_A - (df_A \times MS_{subject/AB})}{SS_{total} + MS_{subject/AB}}$	$\frac{SS_A - (df_A \times MS_{subject/AB})}{SS_A + SS_{subject/AB} + MS_{subject/AB}}$
	B	$\frac{SS_B - (df_B \times MS_{subject/AB})}{SS_{total} + MS_{subject/AB}}$	$\frac{SS_B - (df_B \times MS_{subject/AB})}{SS_B + SS_{subject/AB} + MS_{subject/AB}}$
	$\bar{C}$	$\frac{SS_C - (df_C \times MS_{C\times subject/AB})}{SS_{total} + MS_{subject/AB}}$	$\frac{SS_C - (df_C \times MS_{C\times subject/AB})}{SS_C + SS_{C\times subject/AB} + SS_{subject/AB} + MS_{subject/AB}}$
	AB	$\frac{SS_{AB} - (df_{AB} \times MS_{subject/AB})}{SS_{total} + MS_{subject/AB}}$	$\frac{SS_{AB} - (df_{AB} \times MS_{subject/AB})}{SS_{AB} + SS_{subject/AB} + MS_{subject/AB}}$
	A$\bar{C}$	$\frac{SS_{AC} - (df_{AC} \times MS_{C\times subject/AB})}{SS_{total} + MS_{subject/AB}}$	$\frac{SS_{AC} - (df_{AC} \times MS_{C\times subject/AB})}{SS_{AC} + SS_{C\times subject/AB} + SS_{subject/AB} + MS_{subject/AB}}$
	B$\bar{C}$	$\frac{SS_{BC} - (df_{BC} \times MS_{C\times subject/AB})}{SS_{total} + MS_{subject/AB}}$	$\frac{SS_{BC} - (df_{BC} \times MS_{C\times subject/AB})}{SS_{BC} + SS_{C\times subject/AB} + SS_{subject/AB} + MS_{subject/AB}}$
	AB$\bar{C}$	$\frac{SS_{ABC} - (df_{ABC} \times MS_{subject/AB})}{SS_{total} + MS_{subject/AB}}$	$\frac{SS_{ABC} - (df_{ABC} \times MS_{C\times subject/AB})}{SS_{ABC} + SS_{C\times subject/AB} + SS_{subject/AB} + MS_{subject/AB}}$

Table B (continued)

Quick Reference for Omega Squared Formulas

Design	Factor	Standard Omega Squared $\widehat{\omega}^2$	Partial Omega Squared $\widehat{\omega}_p^2$
3-way Split-plot (1 Between-subjects Factor)	A	$\frac{SS_A - (df_A \times MS_{subject/A})}{SS_{total} + MS_{subject/A}}$	$\frac{SS_A - (df_A \times MS_{subject/A})}{SS_A + SS_{subject/A} + MS_{subject/A}}$
	$\overline{B}$	$\frac{SS_B - (df_B \times MS_{B\times subject/A})}{SS_{total} + MS_{subject/A}}$	$\frac{SS_B - (df_B \times MS_{B\times subject/A})}{SS_B + SS_{B\times subject/A} + SS_{subject/A} + MS_{subject/A}}$
	$\overline{C}$	$\frac{SS_C - (df_C \times MS_{C\times subject/A})}{SS_{total} + MS_{subject/A}}$	$\frac{SS_C - (df_C \times MS_{C\times subject/A})}{SS_C + SS_{C\times subject/A} + SS_{subject/A} + MS_{subject/A}}$
	$A\overline{B}$	$\frac{SS_{AB} - (df_{AB} \times MS_{B\times subject/A})}{SS_{total} + MS_{subject/A}}$	$\frac{SS_{AB} - (df_{AB} \times MS_{B\times subject/A})}{SS_{AB} + SS_{B\times subject/A} + SS_{subject/A} + MS_{subject/A}}$
	$A\overline{C}$	$\frac{SS_{AC} - (df_{AC} \times MS_{C\times subject/A})}{SS_{total} + MS_{subject/A}}$	$\frac{SS_{AC} - (df_{AC} \times MS_{C\times subject/A})}{SS_{AC} + SS_{C\times subject/A} + SS_{subject/A} + MS_{subject/A}}$
	$\overline{B}\overline{C}$	$\frac{SS_{BC} - (df_{BC} \times MS_{BC\times subject/A})}{SS_{total} + MS_{subject/A}}$	$\frac{SS_{BC} - (df_{BC} \times MS_{BC\times subject/A})}{SS_{BC} + SS_{BC\times subject/A} + SS_{subject/A} + MS_{subject/A}}$
	$A\overline{B}\overline{C}$	$\frac{SS_{ABC} - (df_{ABC} \times MS_{BC\times subject/A})}{SS_{total} + MS_{subject/A}}$	$\frac{SS_{ABC} - (df_{ABC} \times MS_{BC\times subject/A})}{SS_{ABC} + SS_{BC\times subject/A} + SS_{subject/A} + MS_{subject/A}}$

Appendix C – Chapter 3. Categories and codes

In the following table all categories and codes are listed, accompanied by short descriptions. In the table, the five themes are represented by numbers; 1) normative sexual development, 2) promiscuity and impulsivity, 3) dependent and romantic creatures, 4) social penalties and praises, 5) ambiguous and unclear answers. In the full codebook, descriptions are more elaborate and each code is followed by examples of actual parental responses. The full codebook will be made available through the supplemental materials upon publication. Each response (i.e., unit) could receive multiple codes. Parents responded to the following vignettes:

1. Thijs is in a relationship with his sixth girlfriend. What type of person do you think Thijs is?
2. Lieke kissed someone she doesn't know while going out. What type of person do you think Lieke is?

Table C

Themes and Descriptions of Categories and Codes

Theme	Category	Description of category	Code	Description of code
1	Normalization	The behavior is described as normal, age-appropriate, fitting with the developmental stage, acceptable.	*Gewoon*	*Gewoon* (Dutch word meaning normal, standard, just, acceptable), typical.
			Can happen	That can happen, happens to the best of us, common.
			Fine	Fine, okay.
			A boy	Thijs is described as (just) a boy.
	Age-related	The situation is linked to the age or developmental stage the adolescent is in, either explicitly or implicitly.	*Puber*	*Puber* (Dutch for a person who is going through puberty).
			Young	Adolescent is described as young, or as a typical 15-year old.
	Exploration	The parent's answer contains elements that allude to exploration and discovery, including investigating boundaries.	Discover	Explore, experiment, discover.
			Adventure	Adventurous, enjoys/pursues adventures.
			Curious	Likes new things, curious.
			Enterprising	Adolescent described as enterprising.
			Excitement	Looks for exciting things, excitement, explores boundaries/risks.

Table C (continued)

Themes and Descriptions of Categories and Codes

Theme	Category	Description of category	Code	Description of code
1	Exploration (continued)		Free	Free, freed, free-spirited.
	In search of	Parent describes searching, discovery of the self, not being sure what one wants in life.	Doubter	Has trouble choosing, can't choose, indecisive.
			Searching	Is looking for something, is still finding themselves, searching.
			Doesn't know what they want	Doesn't know what they want/what they are looking for.
	Pleasure	The answer includes words that express enjoyment and pleasure.	Enjoy	Descriptors related to enjoyment and pleasure. Party-goer, hedonist, likes fun/parties.
			Likes	Adolescent is described as liking girls/boys/women/men/flirting. The other is described generally or passively (e.g. "Lieke likes boys.")
			Cute person	Adolescent is described as having met a cute/nice/prettily person.
2	Promiscuity	Descriptions of promiscuity, capriciousness, and lack of attachment.	Reveler	Loose, promiscuous, unserious, fleeting, libertine.
			Capricious	Inconstant, fickle, volatile, bored quickly, unreliable, on-and-off.

Table C (continued)

Themes and Descriptions of Categories and Codes

Theme	Category	Description of category	Code	Description of code
2	Promiscuity (continued)		Attachment (lack of)	Has trouble attaching, doesn't give themselves fully.
			Not critical	Not picky, likes all and sundry.
	Impulsivity	Descriptors of thoughtlessness, acting without thinking.	Spontaneous	Spontaneous (only literal instances and derivations of the word spontaneous).
			Impulsive	Impulsive (only literal instances and derivations of the word impulsive).
			Follow gut	Follow feelings, gut.
			Unthinking	Unthinking, acting without thinking, not well-considered.
3	Dependent	Descriptors that imply dependence, insecurity or neediness are used. Includes submissiveness and passivity.	Attention	Seeks attention, confirmation, is needy.
			Insecure	Descriptors related to insecurity.
			Not alone	Has trouble being alone.
			Passive	Adolescent is described as kissing/dating because someone else wanted them to.
			Dependent (other)	Submissive descriptors that do not fit other codes in this category, like lonely or unhappy.

Table C (continued)

Themes and Descriptions of Categories and Codes

Theme	Category	Description of category	Code	Description of code
3	Love	Parent connects the behavior with being in/looking for love.	Looking for love	Is looking for love, the (right) one, a partner. Or hasn't found love, the (right) one, partner yet.
			Infatuated quickly	Touched/infatuated (too) quickly, fiery, lost their head over someone.
			Infatuated	Adolescent described as infatuated, having butterflies in their stomach.
4	Slut	Explicit use of derogatory terms that allude to sluttiness.	Slut	Skank, whore, slut, etc.
			Easy	Easy (negative connotation).
	Player	Explicit use of (synonyms of) player.	Player	Player, *mooiboy* (pretty boy) lady killer, Don Juan, womanizer, skirt-chaser.
	Positive	Parent uses descriptors that are explicitly positive (and do not fit in another category).	Popular/ Attractive	Pretty, handsome, attractive, popular/attractive (with boys/girls).
			Nice	Nice/cute (*leuk*).
			Social	Social, outgoing, likes to be with people, is good in making contact, open, easy to make contact with.
			Gezellig	*Gezellig* - sociable, enjoyable.
			Positive (other)	Other positive descriptors, e.g., cheerful, creative, a good friend, sweet.

Table C (continued)

Themes and Descriptions of Categories and Codes

Theme	Category	Description of category	Code	Description of code
4	Negative	Parent uses descriptors that are explicitly negative (and do not fit in another category).	Doesn't care about others	Isn't interested in others, doesn't care what others think/feel.
			Foolish	Foolish, dumb, not clever.
			Wants to be tough	Wants/tries to be tough, shows off, macho.
			Alcohol	Had too much to drink, was drunk, drinking problem
			Negative (other)	Other negative descriptors, e.g., annoying, loser.
5	Ambiguous and unclear answers	Parent gives an ambiguous answer or doesn't give a clear answer, by saying they don't know/are unsure, or by giving multiple options.	I don't know	Parent says they don't know a/the answer.
			No comment	Parent says they have no comment/opinion/judgment.
			Depends	Parent indicates that 'it depends'. Parent gives multiple options, which sometimes contradict each other.

Appendix D – Supplemental Materials and Data Statement

For advancements in research (and the policies and practices that rely upon research), it is important that research is accessible, replicable, and cumulative. To this end, we aim for open access publication or our studies, and follow the FAIR Guiding Principles for scientific data management (M. D. Wilkinson et al., 2016). FAIR is an acronym that stands for Findability, Accessibility, Interoperability, and Reuse of digital assets.

For Chapter 2, the data and materials have already been made publicly available via Open Science Framework. The data and materials can be found through the following stable link: https://osf.io/a2vqe/.

At this moment Chapters 3 and 4 have not yet been published. Upon publication, the data and other supplemental materials that are part of these studies will be made available in publication packages. The publication packages will be openly available at the DataverseNL repository, accompanied by a DOI. For Chapter 5, we did not share a publication package, as this chapter involves data that was collected during a previous project, when the FAIR Guiding Principles were not yet established, and for which procedures and communication towards participants about the handling of the data were different.

Addenda

Nederlandstalige Samenvatting (Dutch Summary)

Bias (vooringenomenheid), vooroordelen en discriminerende ongelijke behandeling kunnen leiden tot structurele ongelijkheid in de samenleving. Dit zien we bijvoorbeeld in de benadeling van vrouwen ten opzichte van mannen wanneer het aankomt op salaris, seksueel- en huiselijk geweld (Dutch Central Bureau of Statistics, 2022a, 2022d; Laan et al., 2021; Leaper & Brown, 2018). Wanneer leerlingen op school benadeeld worden door bias, heeft dit een negatief effect op hun prestaties, gedrag, zelfvertrouwen en motivatie (Wang et al., 2018). Er heerst tegelijkertijd onder veel mensen het idee dat onze samenleving al voldoende gelijkwaardig is en dat aandacht voor ongelijkheid en emancipatie niet (meer) belangrijk is (Kanne & van der Schelde, 2022). Het is daarom belangrijk om te onderzoeken in welke mate bias bestaat in verschillende contexten en hoe bias gecommuniceerd wordt. Bias verandert met de tijd en is vaak subtiel en impliciet van vorm (Leaper & Brown, 2018). In dit proefschrift onderzoek ik dit aan de hand van drie belangrijke *social agents* (sociale tussenpersonen): ouders, leerkrachten en de media. *Social agents* zijn mensen en entiteiten die door middel van verschillende sociale mechanismen aan ontwikkelende individuen overbrengen welke kennis, attituden en gedragingen gepast zijn en welke niet (Bussey & Bandura, 1999; C. Connell & Elliott, 2009; Endendijk et al., 2022; Zaikman & Marks, 2017). Met gebruik van vernieuwende methodologieën heb ik onderzoek gedaan naar de inhoud en mechanieken van bias. Ik heb op deze manier getracht subtiele vormen van bias te documenteren. De focus lag hierbij voornamelijk op genderbias en in mindere mate op etnische bias.

Praktische significantie

In Hoofdstuk 4 en Hoofdstuk 5 maak ik gebruik van variantieanalyses, een statistische methode waarbij gemiddelden van verschillende groepen worden vergeleken. Bij het uitvoeren van kwantitatief statistisch onderzoek is het belangrijk om een maat voor de grootte van een effect te rapporteren. Deze maat geeft een indicatie van hoe sterk een effect is en

welke praktische impact het effect heeft (Ellis, 2010; Fritz et al., 2012; Kirk, 1996). Bij variantieanalyses is de meest gebruikte maat hiervoor èta kwadraat (Peng et al., 2013; Zhou & Skidmore, 2017). Uit onderzoek blijkt echter dat deze maat effecten overschat, zeker wanneer er gebruik wordt gemaakt van een kleine steekproef (Albers & Lakens, 2018; Keppel & Wickens, 2004; Lakens, 2015; Levine & Hullett, 2002). Dit zorgt voor een vertekend beeld van hoe groot een effect is. Er is een alternatief dat effecten niet overschat: omega kwadraat, maar dit alternatief wordt nog maar weinig gebruikt. In Hoofdstuk 2 presenteer ik daarom een uitgebreide uitleg over wat effectgroottes zijn, wat er mis is met èta kwadraat en hoe omega kwadraat berekend kan worden.

Ouders en de seksuele dubbele standaard

Ouders zijn belangrijke sociale agenten, die in sociale interacties met hun kinderen bias kunnen doorgeven (Bandura & Walters, 1963; Bussey & Bandura, 1999; C. Connell & Elliott, 2009; Kreager et al., 2016; Zaikman & Marks, 2017). In Hoofdstuk 3 ligt de focus op ouders als sociale agenten en op een specifieke vorm van bias: de seksuele dubbele standaard (SDS). De SDS houdt in dat er in relatie tot seksualiteit andere verwachtingen zijn voor vrouwen dan voor mannen op het gebied van sociale rollen, gedragingen en karaktereigenschappen (Axinn et al., 2011; Zaikman & Marks, 2017). Recent hebben meerdere onderzoekers zich afgevraagd of deze dubbele standaard nog wel bestaat, uit onderzoek blijkt dat de SDS niet consequent gevonden wordt, mogelijk door groeiende gendergelijkheid en veranderende seksuele moralen (Bordini & Sperb, 2013; Endendijk et al., 2020; Zaikman & Marks, 2017). Deze onzekerheid maakt het onderzoeken van de SDS interessant.

In het huidige onderzoek moesten ouders reageren op hun adolescente kind die twee korte situatiebeschrijvingen voorlas. De ene situatie ging over een jongere die voor de vijfde keer verkering heeft en de andere situatie over een jongere die tijdens het uitgaan zoent met een vreemde. Het gender van de jongeren in de situaties werd afgewisseld, zodat ik kon

onderzoeken of de reacties van ouders genderverschillen vertonen. In eerder onderzoek naar de SDS is vaak gebruik gemaakt van onderzoeksmethoden die sociaal wenselijke antwoorden kunnen uitlokken, zoals vragenlijsten (Bordini & Sperb, 2013; Endendijk et al., 2020). Dit heeft mogelijk geleid tot onduidelijkheid over het voortbestaan van de SDS. Ik koos daarom voor een onderzoeksmethode die een relevante dynamische sociale setting nabootst: ouders die praten met hun adolescente kind over andere jongeren.

De antwoorden van ouders werden thematisch onderzocht door middel van contentanalyse. Uit de analyse bleek dat de meeste voorkomende reacties het normaliseren van seksueel gedrag betroffen en geen SDS bevatten. Een op de vier reacties had wel elementen van de SDS, bijvoorbeeld door meisjes slet te noemen en jongens als player aan te duiden. Meisjes werden daarnaast regelmatig als onzeker en onafhankelijk omschreven, terwijl dat over jongens vrijwel nooit werd gezegd. Er waren ook verschillen tussen de antwoorden van vaders en moeders, en tussen antwoorden gegeven aan zoons en dochters. Zo waren vaders met dochters vaker negatief over jongens dan vader met zonen. Moeders gaven vaker dan vaders aan dat ze geen antwoord konden of wilden geven en gaven vaker tegenstrijdige antwoorden.

Leerkrachten en bias op basisschoolrapporten

Leerlingen ontvangen regelmatig en op allerlei wijzen evaluaties van hun leerkrachten. Deze evaluaties zijn een expressie van de verwachtingen die leerkrachten hebben van leerlingen. Deze verwachtingen kunnen de prestaties, het gedrag en psychische aspecten van leerlingen beïnvloeden (Wang et al., 2018). Wanneer de evaluaties van leerlingen verschillen op basis van persoonskenmerken, zoals gender en etniciteit, is er sprake van bias. Uit eerder onderzoek blijkt dat bevindingen op basis van gender en etniciteit inconsistent zijn (Geven et al., 2018; Wang et al., 2018), waardoor verder onderzoek hiernaar nodig is. In Hoofdstuk 4 onderzocht ik daarom de teksten die leerkrachten hebben geschreven op de schoolrapporten

van basisschoolleerlingen. Schoolrapporten zijn interessante authentieke artefacten voor onderzoek, onder andere omdat leerkrachten op het moment van schrijven niet op de hoogte waren van onderzoeksdoelen.

Ik keek bij dit onderzoek naar drie aspecten: 1) de valentie van de opmerkingen (positief, negatief of neutraal); 2) of de opmerkingen over schoolwerk of over iets anders gingen; 3) de abstractie van bepaalde woorden, aan de hand van het linguïstische categorieën model (Beukeboom, 2014; Maass et al., 1989; Wigboldus et al., 2000). Uit de variantieanalyses in Hoofdstuk 4 bleek dat leerkrachten gemiddeld positieve bias laten zien naar meisjes met een Nederlandse achtergrond dan naar andere leerlingen.

Grote nieuwsverhalen in de media en bias

Door de actualiteiten in de media te volgen kunnen de attituden van mensen richting specifieke groepen beïnvloed worden (Arendt & Northup, 2015; Mertens et al., 2019). Wanneer er tijdens een longitudinaal onderzoek een bepaald onderwerp veelvuldig in het nieuws is, kan dit de onderzoeksresultaten beïnvloeden. Het is daarom belangrijk om de mogelijke invloed van grote nieuwsverhalen te onderzoeken.

Nieuws over seksueel kindermisbruik krijgt disproportioneel veel aandacht in de media in vergelijking met ander nieuws (Grosholz & Kubrin, 2007; Hove et al., 2013), zeker wanneer extreme zaken aan het licht komen (L. R. Smith & Pegoraro, 2020; Weatherred, 2015, 2017). In dit soort nieuwsberichten worden stereotypen over daders vaak bestendigd (Cromer & Goldsmith, 2010b), zoals het idee dat mannen die met zeer jonge kinderen werken gezien worden als potentiële pedofielen (Eidevald et al., 2018; Rentzou, 2011; Sak et al., 2019). In ander onderzoek is gesuggereerd dat dit soort media-aandacht een negatieve invloed heeft op de attituden van mensen ten opzichte van mannen die met jonge kinderen werken (Eidevald et al., 2018; Fahmy & Ibrahim, 2022; Munk et al., 2013; Rentzou, 2011), maar dit is niet eerder onderzocht.

In Hoofdstuk 5 presenteer ik een methode om media-aandacht te onderzoeken en beschrijf ik een casestudie over hoe een grote zedenzaak de genderattituden van ouders in een longitudinaal onderzoeksproject beïnvloedde. Uit de bevindingen blijkt dat een specifieke zedenzaak inderdaad langdurig grote aandacht kreeg in de media. Ouders die aan het onderzoek meededen voordat deze zaak bekend werd waren positiever over mannelijke babysitters dan ouders die meededen nadat de zaak in het nieuws kwam. De negatieve invloed op de attituden van ouders bleek langdurig te zijn.

Conclusie

In de verschillende hoofdstukken in dit proefschrift heb ik bias gedocumenteerd in specifieke sociale contexten. Uit de bevindingen blijkt dat bias subtiel aanwezig kan zijn in de socialiserende boodschappen van verschillende sociale agenten. Deze bevindingen kunnen aanleiding zijn om in te zetten op werkzame interventies die in verschillende sociale contexten worden ingezet, om zo bias in de samenleving tegen te gaan. Omdat bias veranderlijk en subtiel is, is het voortdurend onderzoeken van bias belangrijk. Het gebruik van goede statistische technieken en het bestuderen van relevante en authentieke data in verschillende sociale contexten (zoals gesprekken, geschreven teksten en media-uitingen) gelden daarbij als belangrijke onderzoekscriteria en als aanbeveling voor toekomstig onderzoek.

Curriculum Vitae

Op 25 juni 1988 werd ik geboren in Alphen aan den Rijn en kreeg ik de naam Antoinette Dana Alice Kroes. In 2000 startte ik mijn middelbareschooltijd aan de evangelische school De Passie in Utrecht en ik behaalde mijn vwo-diploma uiteindelijk aan het Christelijk Lyceum Zeist in 2007. Dat jaar studeerde ik korte tijd met veel plezier maar weinig succes Religiewetenschappen aan de Vrije Universiteit in Amsterdam. Ter oriëntatie op het onderwijs liep ik daarna stage bij mijn oud-docent Beeldende Vorming op het Christelijk Lyceum in Zeist. In 2010 begon ik aan de universitaire lerarenopleiding aan de IPABO en de Vrije Universiteit, beide in Amsterdam. Vanaf 2014 stond ik voor de klas op verschillende basisscholen, terwijl ik dat jaar ook een minorprogramma rond gender en seksualiteit aan de Universiteit van Amsterdam doorliep. Daarnaast werd ik beheerder van The Feminist Club Amsterdam, waarvoor ik onder andere discussieavonden, filmscreenings, benefieten en demonstraties organiseerde. In 2015 behaalde ik mijn pabodiploma en mijn bachelorsdiploma in de pedagogische wetenschappen. Aansluitend doorliep ik de master onderwijspedagogiek. Ik was als stagiair en later als medewerker verbonden aan EduDivers, een expertisecentrum voor seksuele diversiteit en onderwijs. Ik gaf vanuit die positie trainingen aan onderwijsprofessionals, studenten en scholieren uit verschillende landen. Ook werkte ik aan een wetenschappelijke onderbouwing van een nieuw lespakket over seksuele diversiteit. In 2016 behaalde ik mijn master-diploma met het predicaat ‘met genoegen’.

Na een jaar voor de klas ging ik in 2017 aan de Universiteit Leiden aan de slag als promovendus op het onderzoeksproject *Girls in Science* van Judi Mesman, gefinancierd met de *Consolidator grant* van de European Research Council. Daarin werd onderzoek gedaan gericht op gendersocialisatie in de gezins- en schoolcontext. Een deel van de resultaten van dit project zijn beschreven in dit proefschrift. Naast de werkzaamheden binnen het project werkte ik samen met een taalkundige aan een onderzoek naar de

teksten op basisschoolrapporten, waarvoor ik deelnemers uit etnische minderheidsgroepen verwierf. Verder werkte ik samen met een Amerikaanse onderzoeker aan een artikel over effectgroottes, die belangrijk zijn bij het uitvoeren van statistische analyses.

Naast mijn werk als promovendus gaf ik in de afgelopen jaren verschillende workshops en lezingen aan studenten en onderwijsprofessionals over onderwerpen als inclusie, leerkracht-verwachtingen en vooroordelen. Vanaf het voorjaar van 2020 gaf ik als vrijwilliger ondersteuning aan een leerling die door COVID-19 niet naar school kon. Sinds 2022 maak ik deel uit van een klankbordgroep bij de ontwikkeling van een nieuwe lesmethode van het COC rond genderidentiteit voor de basisschool. Sinds het voorjaar van 2023 ben ik werkzaam als docent bij de masteropleiding *Educational Needs* aan het Onderwijscentrum Speciale Onderwijszorg, en als onderzoeker bij Kenniscentrum *Youth Education for Society*, beide van Hogeschool Fontys. In deze functies hoop ik mijn passie voor kansengelijkheid en onderwijs tot uiting te kunnen blijven brengen.

List of Publications

Publications

Van de Rozenberg, T. M., **Kroes, A. D. A.,** Van der Pol, L. D., Groeneveld, M. G., & Mesman, J. (2023). Same-sex kissing and having a gay or lesbian child: A bridge too far? Parent-child similarities in homophobic attitudes and observed parental discomfort. Journal of Homosexuality. Advance online publication. https://doi.org/10.1080/00918369.2023.2233658

Kroes, A. D. A., & Finley, J. R. (2023). Demystifying omega squared: Practical guidance for effect sizes in common ANOVA designs. Psychological Methods. Advance online publication. https://doi.org/10.1037/met0000581

Kroes, A. D. A., Jehle, A. M. C., Doornkamp, L., Groeneveld, M. G., Van der Pol, L. D., & Mesman, J. (2021). Dutch adaptation of the implicit association test: Gender arts science [Computer software]. https://app.gorilla.sc/openmaterials/497105

Kroes, A. D. A., Van der Pol, L. D., Groeneveld, M. G., & Mesman, J. (2021). Big news stories and longitudinal data collection: A prominent child sexual abuse case negatively affects parents' attitudes toward male caregivers: International Journal of Behavioral Development, 45(5), 1–8. https://doi.org/10.1177/01650254211042418

Work in progress

Kroes, A. D. A., Van de Rozenberg, T.M., Van der Pol, L. D., Van der Voort, A., Groeneveld, M. G., & Mesman, J. (*submitted for publication)*. Sexual double standards? How parents talk to their child about adolescents who date and kiss.

Van de Rozenberg, T.M., Van der Pol, L.D., Van Esveld, S., Groeneveld, M. G., **Kroes, A.D.A.,** Mesman, J. (*under review*). Nurse, teacher or babysitter: Not a man's job? The role of parents in predicting

gender-stereotypic attitudes towards HEED occupations and gender-stereotypic interest in these careers.

Kroes, A. D. A., Groeneveld, M. G., De Boer, M. M., Van der Pol, L. D., & Mesman, J. (*in preparation*). Gender and ethnicity: Bias in written evaluations of primary school pupils.

Van Esveld, S., Groeneveld, M. G., Van de Rozenberg, T. M., Van der Pol, L. D., **Kroes, A. D. A.,** & Mesman, J. (*in preparation*). A longitudinal study on parents' perceptions, gender stereotypes, and occupations in relation to adolescents' STEM self-efficacy and occupational interests.

Van Esveld, S., Van der Pol., L. D., Toshkov, D. D., **Kroes, A. D. A.,** Groeneveld, M. G., & Mesman J. (in preparation). A longitudinal study on parent's gender stereotypes and family-career involvement in relation to adolescents' gender stereotypical ideas about family and career.

Presentations

Kroes, A. D. A., Groeneveld, M. G., De Boer, M. M., Van der Pol, L. D., Mesman, J. (2022). Gender and ethnicity: Bias in written evaluations of primary school pupils. (Paper presentation). Onderwijs Research Dagen, Hasselt, Belgium.

Van Leest, A. M. C., Dijks, M., **Kroes, A. D. A.**, Korpershoek, H., Douma, N. H. (2022). Basisschooladviezen en hun voorspellende waarde. [Primary school track allocations and their predictive value] (Symposium - Discussant). Onderwijs Research Dagen, Hasselt, Belgium.

Kroes, A. D. A., Van der Pol, L. D., Groeneveld, M. G., & Mesman, J. (2020). Big news stories and longitudinal data collection: A prominent child sexual abuse case negatively affects parents' attitudes toward male caregivers. (Paper presentation). VNOP-CAS Research Days, Utrecht, The Netherlands.

Other publications

Written

Kroes, A. D. A. (2023). (Video)observatie. [(Video) Observation] *Samen Onderwijs Ontwikkelen: Werkwijzen voor participatief actieonderzoek.* Fontys Lectoraat Waarderen van Diversiteit. https://www.fontys.nl/Onderzoek/Waarderen-van-diversiteit/Magazine/Magazine-inhoud.htm

Kroes, A. D. A. (2023, July 3). Het allerbeste docentencollectief [The very best teacher collective]. Sargasso. https://sargasso.nl/het-allerbeste-docentencollectief/

Kroes, A. D. A. (2023, April 10). Kriebels, kleuters, clitoris [Spring fever, kindergarteners, clitoris]. Sargasso. https://sargasso.nl/kriebels-kleuters-clitoris/

Kroes, A. D. A. (2022, October 19). Noem de verwijderde prof gewoon bij naam [Just use the name of the fired professor]. Mare. https://www.mareonline.nl/opinie/noem-hem-gewoon-bij-naam/

Kroes, A. D. A. (2022). De inclusieve klas [The inclusive classroom]. *Praxis Bulletin.* 22-25. https://antoinette.kro.es/wp-content/uploads/2022/04/De-Inclusieve-Klas.pdf

Video

Kroes, A. D. A. (2021, December 8). Lesmethode seksuele vorming PO aflevering 3 - vragen aan een wetenschapper [Lesson program sexual education primary education episode 3 – questions for a scientist]. (Video). Youtube. https://youtu.be/mK27osYYz2Q

Kroes, A. D. A. (2021, December 8). Lesmethode seksuele vorming VO aflevering 3 - vragen
aan een wetenschapper [Lesson program sexual education secundary education episode 3 – questions for a scientist]. (Video). Youtube. https://youtu.be/W5bknOkaKfI

Dankwoord (Acknowledgements)

Het is onmogelijk om iedereen die heeft bijgedragen bij het schrijven van dit proefschrift bij naam te noemen, er zijn te veel mensen die mij hebben geïnspireerd, geholpen en gesteund. Ik begin dit dankwoord daarom in algemene termen: heel veel dank aan iedereen die betrokken is geweest.

Duizenden mensen hebben mij en mijn collega's hun vertrouwen geschonken om intieme details van hun leven te bestuderen. Ik wil alle ouders en kinderen bedanken voor hun deelname aan de onderzoeken in dit proefschrift.

Mijn (co)promotoren Judi, Marleen en Lotte, ik wil jullie bedanken voor de kans om te promoveren en voor jullie begeleiding bij het onderzoeksproject en het schrijven van mijn proefschrift. Ik ben speciaal dankbaar voor het vertrouwen en de ruimte die jullie mij gaven om steeds mijn eigen afwegingen te maken. Tessa, Stefanie, Laura en Astrid: ik heb genoten van jullie als mijn directe collega's, jullie waren onmisbaar. Yiran, Fadime, Daudi en Rosanneke: ik heb veel geleerd van onze uitwisselingen. Mathilde, hoewel je er veel korter bij was dan we hadden voorzien, wil ik je graag bedanken voor hoeveel je hebt gegeven aan dit project en aan mij. Ymke, zonder jou had ik het opgegeven. Dankjewel dat je dagelijks voor me klaarstond, ik ben blij dat ik tijdens dit proces zo'n goede vriendin gevonden heb.

Een groot aantal studenten en assistenten heeft bijgedragen aan het verzamelen, verwerken en coderen van de data. Ik wil jullie allemaal bedanken. Larissa, geen student was zo betrokken bij mijn werk als jij. Marina, Irisa, Dagmar en Annette, bedankt voor het coderen van de schoolrapporten. Dominique, wat kan jij ongelofelijk veel!

Anja en Meike: dank voor jullie inzichten en co-auteurschap, jullie hebben me geïnspireerd. Jason, you have become my far away friend. You were an essential driving force, not just in figuring out omega squared, but also by bringing positivity and light into my entire process. Having a neuro-divergent co-author with whom I could exchange experiences on being an ADHD academic was comforting and fun.

Ik heb me altijd gesteund geweten door mijn collega's van de facilitaire en administratieve diensten. Mijn collega's van de receptie ben ik in het bijzonder dankbaar. Jullie waren de eersten en de laatsten die ik elke dag zag. Met jullie interesse en behulpzaamheid wisten jullie me altijd op te vrolijken en hoop te geven. Ik hoop dat jullie weten dat jullie ontzettend belangrijk zijn en dat de universiteit niet kan draaien zonder jullie. Jullie vertrouwen in mij heeft mij heel diep geraakt. Daarom draag ik mijn proefschrift op aan jullie.

Voor het ontwikkelen van meetinstrumenten en (voor de werving voor) het onderzoek naar schoolrapporten ben ik geholpen door honderden mensen. Ik ben hen allen zeer dankbaar. Dank aan degenen zoals mijn buurman, de IMC Weekendscholen en jeugdwerkers die mij hielpen onderzoeksdeelnemers te vinden. Riiziane, bedankt voor jouw expertise. Dank aan de leerlingen en mijn oud-collega's van de Dr. Plesmanschool in Badhoevedorp. Grote dank gaat uit naar de online community waarmee ik via platforms als Twitter en Bluesky in verbinding sta. Dankzij jullie konden vragenlijsten, vignetten en stellingen in een oogwenk worden getest. Veel mensen die ik nooit heb ontmoet waren bereid mij daarmee en met allerlei inhoudelijke vraagstukken te helpen, ik voel mij daardoor een gezegend mens.

Voorafgaand aan en gedurende mijn promotietraject ben ik enorm geïnspireerd door een groot aantal mensen. Ik heb met veel academici van gedachten mogen wisselen, dat was erg bijzonder. Ik wil ook mijn oud-

docenten van de IPABO, VU en UVA bedanken. Claudia, bedankt dat je in me geloofde, zonder jou was ik nooit aan mijn promotietraject begonnen. Aan alle leden van de leesclub en Nina in het bijzonder: ik ben blij dat ik jullie heb gevonden.

Ik ben mijn nieuwe collega's bij Fontys dankbaar voor het warme bad waar ik in ben terechtgekomen en voor de ruimte die ik heb gekregen om de laatste hand aan mijn proefschrift te leggen.

Ik prijs mezelf gelukkig met de hulp en steun die ik in mijn persoonlijke omgeving heb mogen ontvangen. Ik wil mijn (schoon)familie en vrienden graag bedanken. Sophie, dank dat jij de vacature in 2017 naar mij doorstuurde. Mijn vrienden die ik via Studievereniging Icarus en de pabo heb leren kennen, mijn 'vrienden van de chat', en alle andere vrienden die ik gaandeweg heb gemaakt: jullie hebben me allemaal geraakt, gesteund, geholpen, aangescherpt, aan het lachen gemaakt, uitgedaagd en er doorheen gesleept. Het liefst had ik nog een paar pagina's genomen om jullie allemaal bij naam te noemen, maar het past niet binnen de woordlimiet.

Papa en mama: bedankt dat jullie mij de steun en mogelijkheden hebben gegeven om te worden wie ik ben en om door te leren. Ik ben me meer dan ooit bewust van mijn bevoorrechte positie. Gerald, je bent mijn lievelingsbroer en -pubquizpartner.

Rik, ik ben zo blij dat jij er bent.