Protective teaching mechanisms in case of mild perinatal adversity

Merkelbach, I.

Author: Merkelbach, I.
Title: Protective teaching mechanisms in case of mild perinatal adversities
Issue Date: 2018-11-07

Protective Teaching Mechanisms in Case of Mild Perinatal Adversity

Inge Merkelbach

ISBN: 978-94-92679-59-8
Lay-out by: Proefschriftenprinten.nl – The Netherlands
Publisher: Print Service Ede: Ede, The Netherlands

Protective Teaching Mechanisms in Case of Mild Perinatal Adversity

Proefschrift

ter verkrijging van
de graad van Doctor aan de Universiteit Leiden,
op gezag van Rector Magnificus prof.mr. C.J.J.M. Stolker,
volgens besluit van het College voor Promoties
te verdedigen op woensdag 7 november 2018
klokke 11.15 uur

door

Inge Merkelbach

geboren te Rotterdam
in 1991

Promotor Prof. dr. C. A. Espin
Co-promotor Dr. R. C. A. Rippe

Reading committee Prof. dr. E. J. P. G. Denessen
Prof. dr. D. Dickinson
Prof. dr. L. T. W. Verhoeven
Dr. T. M. Sikkema-De Jong

Table of Content

Chapter 1

General introduction

Jenny and Maria are both in kindergarten. Both were born *late preterm*, that is, both were born about five weeks too early. Both children seemed to have developed reasonably well in their first five years; however, Jenny is beginning to struggle academically, and is falling somewhat behind her peers. She cannot yet write her own name or count to twenty, skills that most children her age have mastered. Maria, on the other hand, is outperforming many of her peers. She can write several words and can even solve simple math problems. Providing Jenny with additional instruction may serve to reduce or to close the academic performance gap between her and her peers; however, what type of interventions are likely to be most effective, and under what circumstances are they likely to work best?

Answering such questions is of importance because Jenny and Maria are not alone – only 80% of children are born on time, that is, between the 38th and 40th week of pregnancy (Martin, et al., 2003). One in five children is born either 'too early' or 'too late.' Research has tended to focus on the outcomes of children who are born too early, that is, children born 'preterm' as these children are at risk for a range of adversities later in life, including health problems (Vohr, 2013) as well as behavioral and cognitive problems (Bhutta, Cleves, & Casey, 2012), which can result in academic delays. Not only children born very preterm (born before the 30th week of pregnancy) (Aarnoudse-Moens, Weisglas-Kuperus, Van Goudoever, & Oosterlaan, 2009), but also children born late preterm (born between the 34th and the 38th week of pregnancy) are at increased risk (Morse, Zheng, Tang, & Roth, 2009).

In this dissertation we examine the effects of interventions on the academic performance of children like Jenny, that is, children who are born late preterm and who struggle academically. However, we also are interested in children like Maria, that is, children who are born late preterm but do *not* struggle academically, and, in fact, seem to outperform their peers.

The difference in academic performance between children like Jenny and Maria might be explained by their *early programming* (Matthews, 2002). *Early programming* refers to the idea that under certain circumstances (i.e. high levels of maternal stress, a risk factor for late preterm birth (Dole, Savitz, Hertz-Picciotto, Siega-Riz, McMahon, & Bueskess,2002)) children are prenatally programmed to be more sensitive to the quality of their environment (Matthews, 2002). This sensitivity can result in either risk-augmenting or risk-protective effects, depending on the specific context (Boyce & Ellis, 2005). For children born late preterm, thus, the environment can exert a formidable influence on their learning and development. In suboptimal environments, the children will fall behind peers, but in optimal environments, the children will perform equal to or better than peers (Pluess & Belsky, 2010)The idea of responsivity to the environment – for better and for worse – is referred to as *differential susceptibility* (Belsky & Pluess, 2009).

A better understanding of the early programming of children born late preterm may help us to more clearly identify what type of academic interventions are likely to be effective for these children, and under what circumstances. Before discussing interventions for children born preterm, we discuss the causes of preterm birth, and the effects they have on the development of the fetus and later phenotype of the child.

Causes of preterm birth and early programming

Of all preterm births, spontaneous preterm births are the most common (Goldenberg, Culhane, Iams, & Romero, 2008). Spontaneous preterm birth refers to natural birth earlier than expected date. Despite the fact that such births are common, little is known about why and when they occur exactly. We only know that the causes of preterm birth are manifold and diverse (Goldenberg et al., 2008).The majority of research on spontaneous preterm births has been done in animals, but due to marked differences in hormone regulation between species, generalizing findings on birth inducing mechanisms to humans is difficult (Muglia & Katz, 2010). What *is* known about spontaneous preterm birth in humans is that intrauterine changes that are typically associated with stressful circumstances (e.g., an increase in stress hormone levels or a decrease in the production of nutrients for the fetus) increase the chances of preterm birth. Such changes may serve as a signaling function to the fetus, indicating that it will be born into a stressful or unpredictable environment (Pike, 2004), and may result in changes in the phenotype of the fetus in preparation for life outside of the womb. This process, which is referred to as early programming (Matthews, 2002), thus results in *adaptive phenotypic plasticity* (Agrawal, 2001), that is, the characteristics of the fetus are adapted to fit the expected postpartum environmental unpredictability and harshness (Belsky, Steinberg, & Draper, 1991).

In anticipation of this unpredictable and harsh environment fetuses can develop increased sensitivity of the hypothalamus-pituitary-adrenal axis (*HPA-axis*) (Matthews, 2002). The HPA-axis is the bodily mechanism responsible for the secretion of the stress hormone cortisol (Kolb & Whishaw, 2009). This increased sensitivity leads to higher levels of stress reactivity (Boyce & Ellis, 2005). A higher level of stress reactivity is a risk factor for a range of negative outcomes in children, including academic delays (Bhutta et al., 2002). From this point of view, increased stress reactivity is considered to be a vulnerability factor. However, increased stress reactivity also can be considered to be adaptive in uncertain environments because it may allow individuals to react rapidly to changes in the environment (Sapolsky, 2015).

However, mostly due to human intervening, the modern world often changes faster than evolution can keep up with. As a result, prenatal signaling cues do not always lead to accurate 'predictions' about the actual living environment: we often see a mismatch between early programming cues and the postnatal environment in humans, culminating into negative outcomes (Nederhof & Schmidt, 2012). This mismatch between early programming cues and the environment might explain part of the (academic) problems children born late preterm often experience. In normal (non-threatening) learning environments, such as the typical kindergarten classroom, increased levels of arousal are bound to be ineffectual and can stand in the way of effective learning.

Academic performance

Increased levels of stress reactivity might thus (partly) explain, why late preterm children, who comprise the vast majority of all children born preterm (> 75%; Goldenberg et al., 2008), experience more academic adversities than children born full term. In this dissertation, only children born late preterm are considered because the majority of preterm children are born late preterm. We did also not include children born before the 34th week of pregnancy because in this group comorbidity of both physical and cognitive problems is very high (e.g. Wood, Marlow, Costeloe, Gibson, & Wilkinson, 2000)

Although most late preterm children do not experience severe school related problems (70% to 80% of these children show normative or above-average academic results (Talge, et al., 2010)), they do still form a substantial group of possibly vulnerable children who, up until now, has received little scientific attention. Academic problems associated with late preterm birth cover a broad range of skills; delays are found in overall achievement (Quigley, et al., 2012; Chan & Quigley, 2014), creative development (Quigley et al., 2012), literacy and language skills (e.g. Nepomnyaschy, Hegyi, Ostfeld, & Reichman, 2012), and numeracy and visuospatial skills (e.g. Lipkind, Slopen, Pfeiffer, & McVeigh, 2012; Baron, Erickson, Ahronovich, Coulehan, Baker, & Litman, 2009).

Interventions: adjusting the environment to accommodate learning in children born late preterm

Reasoning from the differential susceptibility model, the environment is extremely important in the development of sensitive children who show increased levels of stress reactivity, like presumably those born late preterm. To support learning in these children, the learning environment should be shaped to their specific needs. Such learning environments should for example be designed to offer a safe and supportive setting in order to optimally facilitate learning. High levels of verbal responsiveness, that is offering responses which are "*prompt*" (temporally contiguous), "*contingent*" (dependent on children's behavior), and "*appropriate*" (positively connected to children's behavior) (Tamis-LeMonda, Cristofaro, Rodriguez, & Bornstein, 2006) to all actions of the child, can provide children with safety, support, and structure while learning. This approach has already proven to facilitate learning in children with an autism spectrum disorder who

are also, although for other reasons, particularly in need of high levels of safety, support, and structure in the environment (Venker, McDuffie, Weismer, & Abbeduto, 2011). By providing stress reactive children with a safe and non-threatening learning environment that offers high levels of verbal responsiveness, risk protective effects might be provoked, resulting in better academic achievements. For children born full term on the other hand, in general such environments might not hold as much special value because, in most cases, the normal school environment offers enough support for this group. Because the group of children born late preterm might thus benefit most from such additional support, differential effects of responsive and supportive environments are to be expected.

Opportunities of digital learning interventions

Digital educational learning materials could play a key role in supporting stress reactive children. Such materials offer unique opportunities to provide children with learning environments that respond directly and appropriately to all the child's learning related actions. These digital environments could offer well-needed guidance, structure, and safety to stress reactive children. Digital learning environments can also provide continuous support and feedback, while this continuity is not achievable by classroom teachers or caretakers. Such support, or digital scaffolding (Mckenna, Reinking, Labbo, & Kieffer, 1999) can enhance learning by helping children to perform on a level they would not yet reach on their own.

Although little research has been done on what facilitates learning in children born late preterm, results from a small-scale experiment are promising (Van der Kooy-Hofland, Van der Kooy, Bus, Van IJzendoorn, & Bonsel, 2012). In this study children in the second year of kindergarten were assigned either to *Living Letters*, a digital, verbally responsive early literacy program, or to a control condition. Results showed that children who had experienced mild perinatal adversities (i.e. late preterm birth and/or being small for gestational age at birth) benefitted strongly from *Living Letters* (*Cohen's d* = 1.24): They outperformed their peers on alphabetic knowledge and phonemic awareness. However, these children fell behind when assigned to the non-responsive control condition. For children without mild perinatal adversities no effect of condition was found (*Cohen's d* = .34) (Van der Kooy-Hofland et al., 2012). From these findings we might thus conclude that supportive educational programs that offer structure as well as a sense of safety might facilitate learning in certain subgroups of children prenatally programmed to experience increased levels of stress reactivity, such as we expect children born late preterm to be. However, such programs will presumably not hold these learning facilitating affordances for all children.

Methodological challenges

In order to test if and why children born late preterm are susceptible to the effects of digital learning interventions the most powerful design would be a large scale randomized controlled trial in which two groups of children (either born late preterm or born full term) are randomly assigned to either a supportive learning environment or a non-supportive control condition. However, large scale RCT's often entail a range of methodological challenges that can undermine the reliability and thereby generalizability (Stuart, Cole, Bradshaw, & Leaf, 2010) of results, among which are missing data, the trade-off between the reliability and validity and the practical feasibility of collecting outcome measures in large groups. We should thus focus on overcoming these methodological challenges. Ways to improve external validity are the implementation of advanced statistical methods to account for missing data, for example by using multiple imputation (Royston, 2005) or maximum likelihood estimations (Little & Rubin, 2002), and to account for limited data quality via a planned missing data design (Little, Jorgensen, Lang, & Moore, 2013).

The current dissertation aims at providing new insights, as well as focusses on replicating results of a previous small scale study (Van der Kooy-Hofland et al., 2012) using a planned missing data design (Little et al., 2013). Currently, there is a replication crisis in the behavioral sciences (Lilienfield, 2017): Many well-established, highly valued phenomena, have not been replicated straightforwardly in new studies. This lack of replication indicates a high level of false positives in published work. Some go so far as to argue that the majority of published findings are likely to be false (Ioannidis, 2005). Whether or not this claim is accurate, the current crisis underlines the importance of not only carrying out studies in which new phenomena and theories are considered, but also devoting time and effort to replicating previous findings in order to contribute to a robust body of proven findings.

Aims of this dissertation

In this dissertation the effects of verbally responsive computer programs on the academic skills of kindergarten pupils born late preterm are studied. In addition, the neurobiological mechanisms possibly underlying effects are examined. Deploying an experimental approach, the studies address the following aims:

- Replicating the results of a small scale experimental study that examined the short- and long-term effects of digital literacy interventions on early literacy for kindergarten children with mild perinatal adversities (chapter 2)
- Exploring the potential of various methods to account for methodological challenges (i.e. missing data and reliability- and validity issues) inherently associated with large scales RCT's (chapter 3)

- Exploring whether teacher opinions about digital material are likely to contribute to the effectivity of digital learning interventions (chapter 4)
- Establishing whether digital interventions targeting academic areas other than early literacy (i.e. early numeracy) show similar potential for exerting differential effects in children with mild perinatal adversities (chapter 5)
- Examining differential effects of digital interventions for children born late preterm and children small for gestational age at birth (chapter 5 and chapter 6)
- Identifying the (neuro)biological mechanism underlying the differential susceptibility to supportive digital learning environments for children born late preterm (chapter 6)

The main focus of chapter 2 is on the short- and long-term effects of a verbally responsive early literacy program (*Living Letters*) on the early literacy skills of children who have experienced mild perinatal adversities (i.e. born late preterm or small for gestational age). To account for missing data, models are repeated using multiple imputation and maximum likelihood estimation.

The main focus of chapter 3 is on the effectiveness of a numeracy-focused verbally responsive digital program (*Clever Together*) in promoting the development of early numeracy skills in children who have experienced mild perinatal adversities (i.e. born late preterm or small for gestational age). Additionally, this chapter focusses on identifying differences in stress reactivity (i.e. differences in neurobiological reactions to a stressful situation) between children born late preterm and their full term peers. To account for missing data for both research questions, models are repeated using multiple imputation and maximum likelihood estimation.

The main focus of chapter 4 is replicating results of a previous study (i.e. Van der Kooy-Hofland et al., 2012) and of the study reported in chapter 2 of the dissertation using a planned missing data approach: An advanced statistical method used to account for reduced data quality, that is, reduced validity and reliability of the outcome measure (Little & Rhemtulla, 2013).

The main focus of chapter 5 is exploring if differential effects for late preterm children would also be present when these children would interact with a digital intervention program, targeting a different academic area than early literacy skills, in this case early numeracy.

The main focus of chapter 6 is to identify potential differences in stress reactivity in an educational setting between late preterm children and their full term peers, because this mechanism might explain increased susceptibility in the late preterm group.

In chapter 7 the results of this thesis are critically reflected on and recommendations for future research are discussed.

References

Aarnoudse-Moens, C., Weisglas-Kuperus, N., Van Goudoever, J., & Oosterlaan, J. (2009). Meta-analysis of neurobehavioral outcomes in very preterm and/or low birth weight children. *Pediatrics*, 717-728.

Agrawal, A. (2001). Phenotypic plasticity in the interactions and evolution of species. *Science*, 321-326.

Anders, Y., Rossbach, H.-G., Weinert, S., Ebert, S., Kuger, S., Lehrl, S., & Von Maurice, J. (2012). Home and preschool environments and their relations to the development of early numeracy skills. *Early Childhood Research Quarterly*, 231-244.

Baron, I., Erickson, K., Ahronovich, M., Coulehan, K., Baker, R., & Litman, F. (2009). Visuospatial and verbal fluency relative deficits in 'complicated' late-preterm preschool children. *Early Human Development*, 751-754.

Belsky, J., & Pluess, M. (2009). Beyond diathesis stress: Differential susceptibility to environmental influences. *Psychological Bulletin*, 885-908.

Belsky, J., Steinberg, L., & Draper, P. (1991). Childhood experience, interpersonal development, and reproductive strategy: An evolutionary theory of socialization. *Child Development*, 647-670.

Bhutta, A., Cleves, M., & Casey, P. (2012). Cognitive and behavioral outcomes of school-aged children who were born preterm: a meta-analysis. *Journal of the American Medical Association*, 728-737.

Bickel, R. (2007). *Multilevel Analysis for Applied Research It's just regression!* New York: The Guilford Press.

Blair, C., Granger, D., & Peters Razza, R. (2005). Cortisol reactivity is positively related to executive function in preschool children attending Head Start. *Child Development*, 554-567.

Boyce, W., & Ellis, B. (2005). Biological sensitivity to context: An evolutionairy-developmental theory of the origins and functions of stress reactivity. *Development and Psychopathology*, 271-301.

Centraal Bureau voor de Statistiek. (2013). *Documentatierapport Perinatale Registratie (PRN).*

Chan, E., & Quigley, M. (2014). School performance at age 7 years in late preterm and early term birth: a cohort study. *Archives of Disease in Childhood Fetal and Neonatal Edition*, 451-457.

Desoete, A., Ceulemans, A., De Weerdt, F., & Pieters, S. (2010). Can we predict mathematical learning disabilities from symbolic and non-symbolic comparison tasks in kindergarten? Findings from a longitudinal study. *British Journal of Educational Psychology*, 64-81.

Dole, N., Savitz, D., Hertz-Picciotto, I., Siega-Riz, A., McMahon, M., & Bueskess, P. (2002). Maternal stress and preterm birth. *American Journal of Epidemiology*, 14-24.

Goldenberg, R., Culhane, J., Iams, J., & Romero, R. (2008). Epidemiology and causes of preterm birth. *The Lancet*, 75-84.

Ioannidis, J. (2005). Why most published research findings are false. *PLoS Medicine*, 696-701.

Kegel, C. A., Bus, A. G., & Van IJzendoorn, M. H. (2011). Differential susceptibility in early literacy instruction through computer games: The role of the dopamine D4 receptor gene (DRD4). *Mind, Brain, and Education*, 71-78.

Kolb, B., & Whishaw, I. Q. (2009). *Fundamentals of Human Neuropsychology.* New York : Worth Publishers.

Lilienfield, S. (2017). Psychology's replication crisis and the grant culture: Righting the ship. *Perspectives on Psychological Science*, 660-664.

Lipkind, H., Slopen, M., Pfeiffer, M., & McVeigh, K. (2012). School-age outcomes of late preterm infants in New York City. *American Journal of Obstetrics and Gynecology*, 222.e1 - 222.e6.

Little, R. (1986). A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association* , 1198-1202.

Little, R., & Rubin, D. (2002). *Statistical Analysis with Missing Data*. Hoboken : John Wiley & Sons.

Little, T. D., & Rhemtulla, M. (2013). Planned missing data designs for developmental researchers. *Child Development Perspectives*, 199-204.

Little, T., Jorgensen, T., Lang, K., & Moore, W. (2013). On the joys of missing data. *Journal of Pediatric Psychology*, 151-162.

Martin, J., Hamilton, B., Sutton, P., Ventura, S., Menacker, F., & Munson, M. (2003). Births: Final data 2002. *National Vital Statistics Report*.

Matthews, S. (2002). Early programming of the hypothalamo-pituitary-adrenal axis . *Trends in Endocrinology & Metabolism*, 373-380.

Mckenna, M., Reinking, D., Labbo, L., & Kieffer, R. (1999). The electronic transformation of literacy and its implications for the struggling reader. *Reading & Writing Quarterly*, 111-126.

Morse, S. B., Zheng, H., Tang, Y., & Roth, J. (2009). Early school-age outcomes of late preterm infants. *Pediatrics*, 622-629.

Muglia, L., & Katz, M. (2010). The enigma of spontaneous preterm birth. *The New England Journal of Medicine*, 529-535.

Nederhof, E., & Schmidt, M. (2012). Mismatch or cumulative stress: towards an integrated hypothesis of programming effects. *Physiology & Behavior*, 691-700.

Nepomnyaschy, L., Hegyi, T., Ostfeld, B., & Reichman, N. (2012). Developmental outcomes of late-preterm infants at 2 and 4 years. *Maternal and Child Health Journal*, 1612-1624.

Pike, I. (2004). Maternal stress and fetal responses: Evolutionary perspectives on preterm delivery. *American Journal of Human Biology*, 55-65.

Pluess, M., & Belsky, J. (2010). Differential susceptibility to parenting and quality child care. *Developmental Psychology*, 379-390.

Poulsen, G., Wolke, D., Kurinczuk, J. J., Boyle, E. M., Field, D., Alfirevic, Z., & Quigley, M. A. (2013). Gestational age and cognitive ability in early childhood: A population based cohort study. *Paediatric and Perinatal Epidemiology*, 371-379.

Quigley, M., Poulsen, G., Boyle, E., Wolke, D., Field, D., Alfirevic, Z., & Kurinczuk, J. (2012). Early term and late preterm birth are associated with poorer school performance at age 5 years: a cohort study. *Archives of Disease in Childhood Fetal and Neonatal Edition*, 167-173.

Royston, P. (2005). Multiple imputation of missing values: Update of ice. *The Stata Journal*, 527-536.

Sapolsky, R. (2015). Stress and the brain: Individual variability in the inverted U. *Nature Neuroscience*, 1344-1346.

Stichting Perinatale Registratie Nederland. (2011). *Grote lijnen 10 jaar perinatale registratie Nederland*. Utrecht: Stichting Perinatale Registratie Nederland.

Stuart, E., Cole, S., Bradshaw, C., & Leaf, P. (2010). The use of propensity scores to assess the generalizability of results from randomized trials. *Statistics in Society*, 369-386.

Talge, N., Holzman, C., Wang, J., Lucia, V., Gardiner, J., & Breslau, N. (2010). Late preterm birth and its association with cognitive and socioemotional outcomes at 6 years of age. *Pediatrics*, 1124-1131.

Tamis-LeMonda, C., Cristofaro, T., Rodriguez, E., & Bornstein, M. (2006). Early language development: Social influences in the first years of life. In L. Balter, & C. Tamis-LeMonda, *Child psychology: A handbook of contemporary issues* (pp. 79-108). New York: Psychology Press.

Van der Kooy-Hofland, V. A., van der Kooy, J., Bus, A. G., van IJzendoorn, M. H., & Bonsel, G. J. (2012). Differential susceptibility to early literacy intervention in children with mild perinatal adversities: Short- and long-term effects of a randomized controlled trial. *Journal of Educational Psychology*, 337-349.

Venker, C., McDuffie, A., Weismer, S., & Abbeduto, L. (2011). Increasing verbal responsiveness in parents of children with autism: a pilot study. *Autism*, 568-585.

Vohr, B. (2013). Long-term outcomes of moderately preterm, late preterm, and early term infants. *Clinics in Perinatology*, 739-751.

Wood, N., Marlow, N., Costeloe, K., Gibson, A., & Wilkinson, A. (2000). Neurologic and developmental disability after extremely preterm birth. EPICure study group. *The New England Journal of Medicine*, 378-384.

Chapter 2

Reproducibility of Young Learners' Susceptibility to the Learning Context

Introduction: The current study tests if mild perinatal adversities imply increased susceptibility to quality of instruction in early literacy skills. *Method:* In a large-scale experiment (*N* = 981) preschool children were randomly assigned to a digital intervention condition offering guidance and continuous feedback (*Living Letters*) or to a digital control condition that did not contain these features. Effects of the program on short- and long–term literacy outcomes were assessed; for the group as a whole and for children with and without differential susceptibility markers. *Results:* No main effects of the intervention program were found for the group as a whole. Previous findings of susceptibility of children with mild perinatal adversities to *Living Letters* were not replicated. Further exploration of the data revealed, however, increased susceptibility in children born late preterm. Both directly after the intervention and a year later, children born late preterm outperformed their full term born peers if they had received *Living Letters* in kindergarten, but fell behind if they had received the control program. *Conclusion:* An extra program that typically provides continuous guidance and feedback can benefit children born late preterm, but does not benefit children born full term. An increased level of stress reactivity is proposed to be the mechanism underlying the susceptibility to the program found in children born late preterm.

Mild perinatal adversity is generally conceived as a vulnerability factor because of the well-established association between mild perinatal adversities and higher risk of learning problems (Van Baar, Vermaas, Knots, De Kleine, & Soons, 2009). As implied by the emerging notion of differential susceptibility, however, a so-called vulnerability factor may actually be a plasticity factor. Vulnerable individuals, such as children with mild perinatal adversities, may be more susceptible to qualities of instructional programs, for better and for worse. In a prior study, it was shown that children with mild perinatal adversities were at risk for early reading problems, but when their emerging alphabetic skills were stimulated by a computer program targeting these skills, these children reached a higher level of early reading skills compared to their non-risk peers, an advantage that remained a year later (Van der Kooy-Hofland, Van der Kooy, Bus, Van IJzendoorn, & Bonsel, 2012).

In the current study, we test the reproducibility of Van der Kooy-Hofland et al.'s (2012) results and conclusions. In the Van der Kooy-Hofland et al. (2012) sample, there was only a small number of children with perinatal adversities (N = 21). It is important to examine the inferential reproducibility (Goodman, Fanelli, & Ioannidis, 2016) in other, preferably larger samples. The current study was part of an ongoing large-scale extensive experiment that took place in 172 Dutch schools for primary education. The primary aim of the large-scale study was to test a gene x environment interaction targeting genes related to the dopamine-system. With rather modest additional costs and efforts this experiment allowed for testing the reproducibility of the hypothesis that children with perinatal adversities were more susceptible to a program that offers guided practice to learn alphabetic skills, that is, to the *Living Letters* program, a computer-based remedial intervention with an adaptive feedback regime. The current study was similar to the study carried out by Van der Kooy-Hofland et al. (2012) except for small details of experimentation. The large sample guaranteed that sufficiently large numbers of pupils with low base rate perinatal adversities could be sampled and included in the experiment. It also allowed for examination of the effects of the Living Letters program on subsamples of children with perinatal adversity, specifically children born late preterm and children small for gestational age.

The current line of research was inspired by a study by Boyce et al. (1995), who found that biological reactivity makes children more sensitive to the context, both for better and for worse. That is, highly biologically reactive children who were in high-adversity childcare settings or home environments had substantially higher illness rates than other groups of children, however biologically reactive children who were in more supportive childcare or family settings had the lowest illness rates. It may be that mild perinatal adversities lead to higher cardiovascular and HPA-axis reactivity to context, which, according to the pioneering study of Boyce et al. (1995), would make children more sensitive to context, for better and for worse. Due to heightened stress reactivity, children with perinatal adversities may easily shut themselves off from learning experiences, especially when those experiences are unstructured. The concept of biological reactivity, for better and for worse, can be applied to an educational context as well. For example during the preschool years, children learn alphabetic skills, but the learning is often unstructured. That is, rather than receiving systematic instruction, children learn through accidental events such as attempts to write their name or 'mama', a parent informally instructing letters or phonemic awareness saying "See that is the letter P from Peter", and so forth. However, it may be that biological reactive children would profit from systematic instruction in alphabetic skills. The target program, *Living Letters* provides such systematic instruction. *Living Letters* makes use of guided practice and provides continuous feedback, features that may be particularly helpful for children suffering from an increased biological reactivity to stress.

Perinatal adversity and academic performance

Both low birth weight and preterm birth have been associated with negative cognitive and academic outcomes later in life. Children who are small for gestational age at birth are found to have lower IQ-scores (Hutton, Pharoah, Cooke, & Stevenson, 1997; Sommerfelt, et al., 2000) and poorer cognitive performance (McCarton, Wallace, Divon, & Vaughan, 1996), and are at risk for developmental delays and language problems (Gutbrod, Wolke, Soehne, Ohrt, & Riegel, 2000). Compared to full term children, children born late preterm have twice the risk for enrollment in special education at all grade levels (Van Baar et al., 2009), are at increased risk for developmental delays and school-related problems (Morse, Zheng, Tang, & Roth, 2009; Quigley, et al., 2012), and are at increased risk for literacy problems or disabilities (e.g. Guarini, Sansavini, Fabbri, & Savini, 2010; Kirkegaard, Obel, Hedegaard, & Henriksen, 2006).

Perinatal adversity and stress

Being born (late) preterm is associated with dysfunctioning of the hypothalamic-pituitary-adrenal axis (HPA-axis) (e.g. Buske-Kirschbaum, et al., 2007; Bolt, Van Weissenbruch, Lafeber, & Delemarre-Van de Waal, 2001). The HPA-axis controls the secretion of the stress-hormone cortisol (Kolb & Whishaw, 2009)) and may therefore be essential for coping with stress (Aisa, Tordera, Lasheras, Del Río, & Ramírez, 2007). The preterm group may easily feel stressed, and the stress may interfere with their ability to attend to information (Gotlib, Joormann, Minor, & Hallmayer, 2008). Hence, they may need external support to control extreme stress reactivity to the environment in order to benefit from a program such as *Living Letters* that provides guided practice and continuous feedback.

Being born small for gestational age has also been shown to be related to the functioning of the HPA-axis (e.g. Bolt et al., 2001). For instance, low-birth-weight babies showed increased cortisol concentrations in umbilical cord blood, and raised urinary cortisol excretion in childhood (Economides, Nicolaides, Linton, Perry, & Chard, 1988). In adult life, they have higher pulse rates, an index of sympathetic activity, and increased fasting cortisol concentrations (Phillips, et al., 1998; Reynolds, et al., 2001). Studies have shown an enhanced plasma cortisol response to synthetic adrenocorticotropic hormone (Levitt, et al., 2000). Further, an increased stress response has been observed in low-birth-weight children (Phillips & Jones, 2006). Thus, *Living Letters*, may fit the needs of this subsample as well, because the program may help to control extreme stress reactivity to the environment.

Aims of current research

The main aim of the current study was to replicate and extend a previous small-scale prior experiment that demonstrated an increased susceptibility to a computer program, *Living Letters*, compared to a control program (*Living Books*) for a group of children with mild perinatal adversities (Van der Kooy-Hofland et al., 2012). In the previous study, a large effect size was found for the susceptible group (d = 1.5, 84% CI = .74, 2.15), and a small effect size for the non-susceptible group (d = .00, 84% CI = -.33, .33). We also examined the long-term effects of *Living Letters* using standardized tests assessing word recognition about one year later (i.e. rapid word reading). In the previous study, the effect size was large for the susceptible group (d = 1.17, 84% CI = .44, 1.8) but small for the non-susceptible group (d = -.04, 84% CI = -.40, .31). Lastly, we extended the previous research by examining effects separately for children who were small for gestational age and children who were born late preterm.

Methods

Design

The purpose of this study to replicate the small-scale study carried out by Van der Kooy-Hofland et al. (2012), but with a larger sample size. We thus designed the study to similar to the previous study, with some small changes had due to the larger sample size. The current study used data collected in two successive research waves (2013-2014 and 2014-2015) in which in total 147 different Dutch schools participated. In 2013, the experiment was carried out at 57 schools. Teachers selected, with the help of a commonly applied standardized test (*Cito Kindergarten Test*), children who were delayed in basic knowledge skills essential for learning to read. Since teachers were in control of selecting children, some bias could have been introduced (Ready & Chu, 2015). Children were randomly assigned to different treatment conditions, as proposed in (Parker, 1990). A similar rigorous procedure was followed a year later with 118 schools, resulting in a total of 981 participants across both research waves. There was only a small overlap of schools between the two waves (*k* = 28 schools). The short-term post-test was a digital literacy test designed by the researchers. The test included three subtests, and was administered individually and computer-assisted by the teacher. The long term post-tests were standardized literacy tests that are commonly administered to first graders in the eighth month of school. The tests target beginning word reading (accuracy and rate).

Participants

Based on the 20% perinatal adversities in the prior experiment (Van der Kooy-Hofland et al., 2012), we estimated that a sample of 450 children might include approximately 90 children with perinatal adversities. A sample this large would allow for examination of low birth weight children and preterm children separately. The initial sample for the current study consisted of 981 five-year-old children. Participants were excluded from analysis due to missing pretest or posttest information or incomplete perinatal information (see flow diagram in Figure 1). Two children born (very) preterm (before 34 weeks of

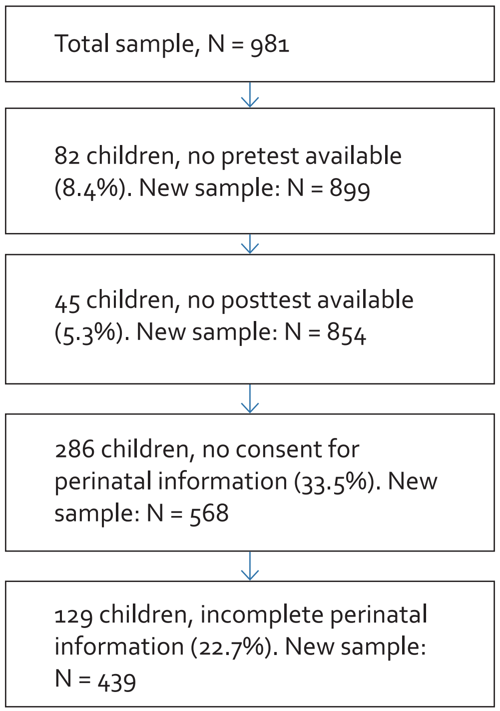


Figure 1. Participant selection scheme

pregnancy) were also excluded from analyses. The final sample consisted of 439 children from 147 different schools. Of these children, 55 children were born late preterm and 102 were small for gestational age at birth. The 55 children born late preterm were from 44 different schools. None of the participating schools provided more than three children born late preterm to the final sample. The 102 children who were small for gestational age at birth were from 78 different schools. Most schools provided only one to two pupils small for gestational age at birth. Of all participants, 49.5% scored below average on a standardized literacy test (*Cito Kindergarten test*) while the rest scored mid-range.

Procedure

The study protocol was approved by the Ethics Committee of the department of Child and Educational Studies of Leiden University, and was carried out in accordance with its codes of conduct.

Data collection took place over two consecutive school years (2012/2013 and 2013/2014). From August to February, schools were recruited by sending out flyers and letters containing information about the content and purpose of the study through both email and mail. We offered participating schools three months of free access to all intervention programs. These programs normally require a paid subscription (http://www.bereslim.nl). If teachers agreed to participate, they were asked to select pupils from their classroom who were achieving poorly in language/literacy. This process was the same as the one used in the Van der Kooy-Hofland et al. (2012) study. Initial eligibility for pupils was determined by their ability to write their proper name, to rhyme, to name a few letters, and to identify sounds in words. In addition these children had to score in the lowest ranges -between 0 and 59- on the standardized language/literacy test CLT administered in January (Lansink & Hemker, 2010). However, if there were not enough children scoring below the 40th percentile, teachers were asked to include other children who they believed were in need for additional guidance in the field of early literacy. For the first wave of data collection, parents were asked 'after the fact' – that is, at the end of the study – for consent for retrieval of perinatal information. The response rate for the first wave was fairly low (43% consent). For the second wave of data collection, consent for retrieving of perinatal information was asked for prior to the beginning of the study. The response rate for the second wave was much (94%) higher.

Similar to the Van der Kooy-Hofland et al. (2012) study we contrasted two interventions: *Living Letters* and *Living Books* (other conditions included in the larger study are not discussed in this manuscript). On average, one to two children per classroom participated in the study (*Mean* = 1.66 children per classroom, *SD* = .89). As in the previous study, children were randomly assigned to a condition by one of the researchers. The sessions took place once per week, and were spread out over a period of approximately two to three months. Except for logging in, children worked on their own without adult assistance during the sessions. Children wore headphones in order to prevent being disturbed by other children. Children worked with the mouse and did not have to make use of the keyboard. This procedure was similar to the procedure followed in Van der Kooy-Hofland et al. (2012). One key difference in the two studies was that in the current study teachers, rather than researchers, implemented the intervention.

Target programs

Living Letters is designed to promote understanding of the alphabetic principle and to improve phonemic awareness of young children. In the program two main characters, a boy and a girl, explain the assignments. An online tutor (the boy's teddy bear) provides the children with adaptive feedback. Lessons are provided in a game format. In the first 22 lessons of *Living Letters*, children practice recognizing their own written names (or the word 'mamma') between other symbol strings or scribbles. The following six lessons focus on the sound of the first letter of the child's name. In the last twelve lessons, children select pictures of words that start or end with the first letter of their own name. The tutor (the teddy bear) provides the children with the following feedback. For correct answers, the teddy bear confirms that the answer is correct and explains why it is correct. For incorrect answers, the teddy bear provides three levels of feedback: (1) repeating instructions; (2) providing cues to the answer if children answer a second time incorrectly; and (3) verbalizing how the correct solution can be found if children didn't find the correct solution themselves or when the online tutor had modeled the answer. The program thus provides not only feedback as to the accuracy of answers, but it also offers hints and explanations. The program is adaptive to children's needs. If children fail during their first attempt to complete the game, the assignments are repeated in the two subsequent sessions.

Control children received *Living Books* during the same period of time. This program includes eight digital, animated, age-appropriate stories based on high-quality children's books. Each story is 'read' twice. A picture storybook is read to the children by a computerized voice while children watch animations and listen to background sounds and music that support comprehension of the story content. Text is not presented as print on screen but only orally. Each reading session is interrupted four times so that children can answer two questions about the story events and two questions about difficult words in the text. After answering the questions, children receive immediate feedback, as well as positive reinforcement in the form of compliments.

Both the *Living Letters* and *Living Books* computer programs stored the number and duration of log-ins. Data revealed that children completed on average 33.62 out of 34 *Living Letters* games (*SD* = 2.50) and they "read" on average 14.80 out of 16 *Living Books* (*SD* = 1.80). Children worked on average 144.07 minutes (*SD* = 95.24) with *Living Letters* and 163.32 minutes (*SD* = 110.37) with *Living Books.* Time spent per individual child depended on both time required to come up answers and on how many retries and feedback rounds were needed.

Measures

Pretest

As pretest the Cito Literacy Test for Kindergarten Pupils (CLT, Lansink & Hemker, 2012) was used. The CLT is a group-administered test given by teachers in January/February. The test consists of 60 paper-pencil questions measuring a range of language and literacy skills: vocabulary, critical listening, rhyming, hearing the first or last word in a sentence, sound blending, writing conventions, and predicting book content based on book cover. Children's pretest scores were categorized as 'at risk' scores within the lowest 25% (score of 59 or below) or 'not at risk' (score of 59 and beyond).

Posttests, short term (directly after intervention)

Phonemic awareness. The Phonemic Awareness Task included five items. Children identified the first sound of five words (e.g. muis [mouse]) while pictures of the words were shown on the computer screen. *Cronbach's α* was .758 for the phonemic awareness test.

Letter knowledge. Children identified ten letters presented on screen by pointing to them (i.e. *s*, *k*, *a*, *p*, *r*, *o*, *v*, *m*, *t*, & *n*). *Cronbach's α* was .827 for the letter knowledge test.

Word Picture Task. Children matched a printed word with picture. For each of six words (e.g. dak [roof]) there were four options from which children could choose: correct (dak), first letter correct (dor), first and last letter correct (dek), and entire word incorrect (lom). *Cronbach's α* was .827 for the word picture task.

Aggregate measure. Principal component analysis (PCA) applied to the three tests resulted in one component explaining 67.59% of the variance. Component loadings ranged from .74 to .86. Scores were combined by calculating the average standardized score, with a higher score indicating better alphabetic skills.

Posttest, long term (eight months into first grade)

Three Minute Test (TMT). We selected a commonly applied standardized test to assess literacy development in first grade: The Three Minute Test (TMT) test. The TMT is designed by the Dutch educational institution Cito and assesses accuracy and speed in word reading. Children read aloud as many words as they can in three minutes from a set of reading cards, each containing 150 words. Teachers scored the number of correct words. Easy and difficult words were equivalently balanced per card.

Perinatal data

The Netherlands Perinatal Registry (Stichting Perinatale Registratie Nederland, 2011) contains comprehensive data on pregnancy, pregnancy care (interventions, referrals), and pregnancy outcomes. The variables are recorded by the health care provider during prenatal care, delivery and neonatal and lying-in period. The register covers approximately 96% of all deliveries in the Netherlands. The data from three registers (the National Obstetric Database by midwives, the National Obstetric Database by gynecologists, and the National Neonatal/Pediatric Database) are annually sent to the national registry office, where a number of range and consistency checks are conducted. The perinatal registry can be accessed by researchers, provided that they have the written permission of the mother. Missing values in our sample were largely due to non-consent for retrieving data (61%). A second reason was failure to connect data in the registry to the research database (39%). Criteria for assignment to the group with mild perinatal adversities were birth weight between the 2.5th and 10th percentile for the gestational age (small for gestational age group) or gestational age at birth between 34-37 weeks, 6 days (late preterm birth group). Thresholds for the small for gestational age group were those used by the Netherlands Perinatal Register, which are based on birth weight, duration of pregnancy, parity, and gender of the child. In the study by Van der Kooy-Hofland et al. (2012), the group of children with mild perinatal adversities was too small to test effects of Living Letters on subsamples.

Data analysis

Testing the differential susceptibility model

For effects on the short- and long term measurement, a multilevel approach using mixed models was applied in order to account for variation attributable to school-level characteristics (Luke, 2004). We employed a likelihood-ratio test for examining whether the model improved when intercepts or both intercepts and slopes were allowed to vary across schools. In all models the following variables were included: cohort (first or second tranche), pretest score, condition, perinatal adversity, and the two-way interaction, condition * perinatal adversity. If the interaction between the susceptibility marker and the intervention was significant, effect sizes (*Cohen's d*) and their 95% Confidence Intervals were calculated and compared for susceptible and non-susceptible groups. Estimates were based on mean outcome scores and standard errors ignoring covariates. Likewise, it was tested whether both criteria for perinatal adversities- being born preterm or being small for gestational age- were susceptibility markers.

Missing data

Based on Little's MCAR test (Little, 1988), we could reject the null hypothesis that data were not missing completely at random (χ^2 = 14.66, p = .066); therefore, as a first step, complete case analysis was applied, i.e., including only individuals with complete data. To further account for missing data, both models (short- and long-term), as fitted on complete data, were also estimated using a multiple imputation (MI) approach accounting for possible differences between the two cohorts. Using a MI-approach, missing values were imputed (m=100 datasets) via chained equations by using an imputation model which included all variables as well as all interactions (Graham, Olchowski, & Gilreath, 2007). Estimates of parameters and standard errors were pooled over all imputed datasets. This approach yields very precise parameter estimates, but has slightly increased standard errors to account for the estimation of missing information. In order to assess the robustness of the results, estimates and standard errors were compared between the applied approaches. Similarity of estimates would indicate robustness, while considerable differences would signal that results derived from complete case analysis might be strongly affected by bias due to missing data.

Results

After comparing sample characteristics for the *Living Letters* (experimental) and *Living Books* (control) groups, both short and long term effects of *Living Letters* will be considered. We first examined our results would replicate findings of Van der Kooy-Hofland et al., (2012) on the short term measures for children with mild perinatal adversities as one group. We then examined whether findings were different for children born late preterm and children who were small for gestational age at birth. The same procedure was then followed in examining the long term effects.

Comparison of sample characteristics for experimental and control groups

The experimental and control groups did not differ in age (t (432) =.22 p=.823), educational level of the father (t (421) =-.19, p=.848), and pretest score (t (432) = -.33, p=.743). Nor did the groups differ in gender (χ^2(1) = .47, p=.495), number of children with perinatal adversities (χ^2(1) = .07, p=.793), number of late preterm children (χ^2(1) = .01, p=.682), or number of children small for gestational age (χ^2(1) =.68, p = .944). Table 1 presents characteristics for the complete group and for subgroups broken down by condition (Rosenberg et al., 1992).

Table 1. *Sample characteristics for the complete group and broken down by condition*

	Complete group (*n* = 439)	Experimental *Living Letters* (*n* = 230)	Control *Living Books* (*n* = 209)	*p*
Male	55.4%	53.9%	56.9%	.524
Age (in months)	66.81 (*4.23*)	59.53 (*7.80*)	66.86 (*4.30*)	.793
Father's education (max = 6)	3.71 (*1.38*)	3.74 (*1.42*)	3.69 (*1.35*)	.721
Distribution of condition in first wave of data collection	23.9%	23.0%	24.9%	.652
Mild perinatal adversities	32.3%	33.0%	31.6%	.743
Late preterm	12.5%	12.6%	12.4%	.958
Small for gestational age	23.2%	22.6%	23.9%	.745
CLT* pretest (raw score)	59.85 (*8.06*)	59.53 (*7.80*)	60.22 (*8.35*)	.372
CLT pretest (percentage low)	49.7%	50.4%	48.8%	.733
Alphabetic knowledge posttest (z-score)	.00 (*1.00*)	-.04 (*1.00*)	.04 (*1.00*)	.389
CLT posttest word recognition (raw score)	31.24 (*20.30*)	29.81 (*22.59*)	32.72 (*17.57*)	.251

*CLT = Cito Literacy Test

Short-term effects of Living Letters, broken down by adversity groups

As an initial step in the analyses, we compared the short term effects of *Living Letters* for children with mild perinatal adversities vs. children without perinatal adversities. The fit of the null model significantly improved after adding a random intercept for school ($\chi^2(1) = 8.21$, $p < .01$). The fit of the model deteriorated significantly after adding a random slope for intervention, $\chi^2(2) = 6.46$, $p < .05$. Intra class correlation equaled 13%.

The CLT pretest was a significant predictor for the posttest score (t (430.17) = 6.96, $p<.001$). There was no main effect for perinatal adversities (t (424.42) = -1.60, p =.111). *Living Letters* (vs. *Living Books*) approached a main effect (t (379.42) = -1.83 p =.068), albeit in favor of the control condition. The interaction between condition and perinatal adversities approached but did not reach significance (t (418.05) = 1.84, p=.066), indicating that the non-susceptible group benefited more from the control condition whereas the susceptible group benefited more from the intervention condition. Table 2 describes the posttest scores per condition for different group definitions (general adversity, specific for (absence of) preterm birth, (absence of) being born small for gestational age and the total group). Repetition of analysis with imputed datasets yielded highly similar results (Supplementary Table 1): Estimates and standard errors strongly resembled those found in complete case analysis, including those for the interaction between mild perinatal adversities and intervention (Estimates for complete cases were: .35 (*.19*), for MI: .34 (*.15*)).

Table 2. *Means and Standard Deviations for post-test scores by Condition and Mild perinatal adversities, LP, and SGA*

	Alphabetic Knowledge & Phonemic Awareness			
	Living Letters	***n***	***Living Books***	***n***
No perinatal adversities	-.08 (*1.02*)	154	.08 (*.97*)	143
Mild perinatal adversities	.05 (*.95*)	76	-.04 (*1.07*)	66
Full term	-.07 (*1.01*)	201	.08 (*.98*)	183
Late Preterm	.16 (*.94*)	29	-.22 (*1.08*)	26
Not SGA*	-.04 (*1.01*)	178	.05 (*.99*)	159
SGA	-.03 (*.96*)	52	.02 (*1.04*)	50
Total	-.04 (*1.00*)	230	.04 (*1.00*)	209
	Word Recognition standardized			
	Living Letters	***n***	***Living Books***	***n***
No perinatal adversities	27.31 (*17.40*)	93	32.51 (*15.69*)	91
Mild perinatal adversities	35.95 (*30.99*)	39	33.29 (*21.97*)	35
Full term	28.19 (*17.97*)	118	33.07 (*17.31*)	109
Late Preterm	37.93 (*27.57*)	14	29.71 (*16.53*)	17
Not SGA	27.53 (*17.96*)	104	31.79 (*15.51*)	101
SGA	35.54 (*22.93*)	28	35.96 (*22.82*)	25
Total	29.23 (*19.31*)	132	32.62 (*17.18*)	126

*SGA = small for gestational age

In Table 3, main outcomes (*d*s, *n*s and 84% *CI*s) are summarized for susceptible and non-susceptible groups. The direction of the difference between the group with perinatal adversities and the control condition was in accordance with the differential susceptibility model: the adversity group benefited more from *Living Letters* when compared to the control condition (*Cohen's d* = .09) than did the group without perinatal adversities (*Cohen's d* = -.16), but not significantly so (p = .123).

Table 3. *Effect sizes and 84% confidence intervals in susceptible and non-susceptible groups*

	Susceptible				Non-susceptible					
	d	84%*CI*	n_e*	n_c*	*d*	84%*CI*	n_e	n_c	*z*	p_1
	Short term effect									
Perinatal adversity vs. no perinatal adv.	.09	-.15/ .33	76	66	-.16	-.32/.00	154	143	1.16	.123
Late preterm vs. full term	.38	-.01/.75	29	26	-.15	-.29/- .01	201	183	1.75	.040
	Long term effects									
Perinatal adv. no perinatal adv.	.10	-.23/.42	39	35	-.31	-.52/-.10	93	91	1.48	.068
Late preterm vs. full term	.37	-.15/.87	14	17	-.28	-.46/-.09	118	109	1.69	.045

$_1$ one-tailed * n_e = number of participants in experimental condition; * n_c = number of participants in control condition

Exploratory secondary analyses

The analyses were repeated with late preterm (LP) and small for gestational age (SGA) as markers for susceptibility (Table 4). For both adversities a dummy variable was created. LP and SGA were not mutually exclusive, children could be both LP and SGA as was the case for 15 children (3.4%). Thus, children could fall in both groups simultaneously. The regression was carried out with a random intercept for school because the fit of the null model significantly improved after adding a random intercept for school ($\chi^2(1)$ = 8.56, p< .01). A random slope (for condition) diminished model fit ($\chi^2(2)$ = 6.69, p <.050). CLT pretest (t (428.21) = 6.86, p <.001) was a significant predictor for the posttest score. There were no main effects for *Living Letters* (t (375.18) = 1.67, p=.095), SGA (t (423.58) = -.64, p=.524), or LP (t (424.53) = -1.48, p=.140), nor for the interaction between condition and SGA (t (420.68) =.51, p = .612). The interaction between condition and LP, however, reached significance (t (420.68) = 1.98, p=.048), indicating that late preterm children benefited most from the intervention. As can be concluded from inspection of the graph presented in Figure 2, children born late preterm outperformed their peers without mild perinatal adversities when assigned to *Living Letters*, and fell behind when assigned to the control condition.

Table 4. *Regressing the aggregate measure of alphabetic knowledge on CLT pretest, Living Letters, SGA age, and LP, controlling for age, sex, and father's education*

Measure	Estimate (*SE*)	95% CI	*t*	*p-value*	*df*
Fixed effects					
Intercept	-.26 (*.21*)	-.68 - .15	-1.26	.207	402.00
Main effects					
Cohort	.03 (*.11*)	-.18 - .25	.31	.756	393.87
CLT* pretest	.62 (*.09*)	.44 - .80	6.86	<.001	428.21
Living Letters (vs. *Living Books*)	-.18 (*.10*)	-.38 - .03	1.67	.095	375.18
Late preterm	-.29 (*.20*)	-.68 - .10	1.48	.140	424.53
Small for gestational age	-.10 (*.15*)	-.40 - .21	-.64	.524	423.58
Two-way interactions					
LP* X *Living Letters*	.54 (*.27*)	.00 - .11	1.98	.048	420.68
SGA* X *Living Letters*	.11 (*.21*)	-.31 - .53	.51	.612	420.68
Measure	**Estimate (*SE*)**	**Wald Z**	***p*-value**		
Random effects					
Level Child	.79 (*.06*)	12.24	<.001		
Level School	.12 (*.05*)	2.36	.018		

*CLT = Cito Literacy Test; SGA = small for gestational age

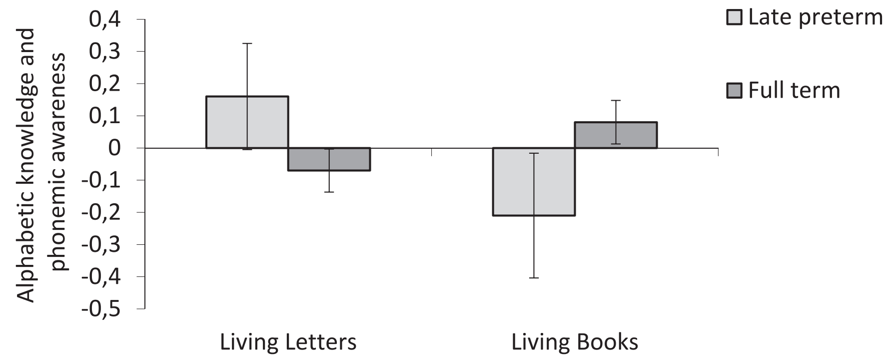


Figure 2. Interaction between late preterm and condition with alphabetic knowledge and phonemic awareness (short term) as outcome measure

Repetition of the analysis using MI yielded similar results and similar conclusions. Estimates and standard errors where highly comparable across all parameters (Supplementary Table 2), including the interaction between LP and condition. Estimates (standard errors) for complete cases were: .54 (*.27*); for MI: .44 (*.37*).

The effect sizes are in accordance with differential susceptibility; see Table 3. In the late preterm group, *Living Letters*, as compared to the control condition had a larger effect (d = .38) than in the full-term group (d = -.15). Because we expected deviations in one direction, we carried out a one-tailed test which was significant (p < .04).

Long term effects of Living Letters at eight months into first grade

Word recognition scores administered in May/June in first grade were available for 258 children (58.8% of total sample) of which 74 were children with perinatal adversities. A random intercept offered the best fit, as compared to a random slope ($\chi^2(2)$ = 1.50, p> .050) or an ordinary least squares (OLS) model ($\chi^2(1)$ = 2.55, p> .050). The intra class correlation equaled 8%. Scores of two children were winsorized at 3 SD's from the mean. A main effect was found for pretest (t (251.20) = 3.16, p = .002), and condition (t (238.35) = -2.35, p= .026): Children in the *Living Books* condition had higher mean scores (*Mean* = 32.62, *SD* = 17.18) than children in the *Living Letters* condition (*Mean* = 29.23, *SD* = 19.31). Perinatal adversities (t (249.73) = -.26, p = .797) did not result in a main effect, nor did the interaction between perinatal adversities and condition (t (248.05) = 1.52, p =.130) reach significance. Repetition of analysis in imputed datasets yielded similar results (Supplementary Table 1). When we included LP and SGA, instead of mild perinatal adversities, as markers for differential susceptibility the model with only school as random intercept again fitted best. In this analysis, the interaction between condition and LP reached significance (t

(247.46) = 2.16, p = .032). Inspection of the interaction depicted in Figure 3 reveals that children born late preterm benefited from *Living Letters* and outperformed their peers when assigned to this condition, however they did not fall behind when assigned to the control condition (*Living Books*). Children born full term, on the other hand, had higher scores when assigned to the control condition (*Living Books*) than when assigned to the target program. After working with *Living Letters* late preterm children showed an average score of 43.93 (SD = 44.30), which was between the 60th and 80th quartile (ranging from 39 to 50). Late preterm children thus performed above average. All other groups included in this analysis on average scored (just) within the average range, that is, between the 40th and 60th percentile, showing no effect of condition on performance.

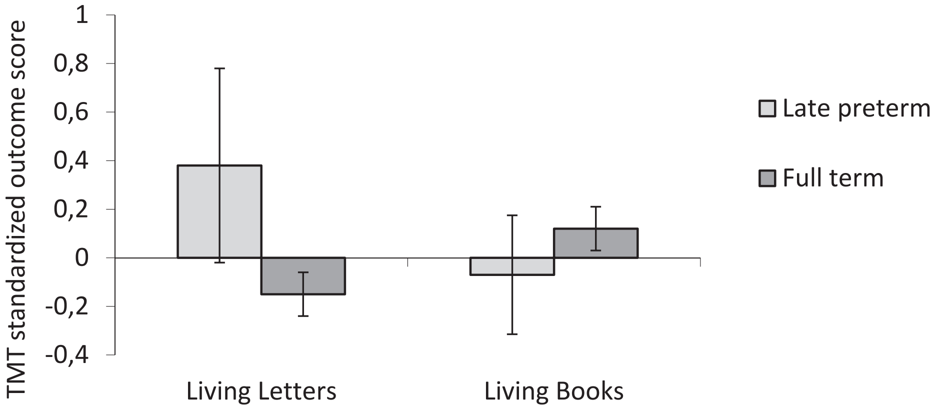


Figure 3. Interaction between late preterm and condition with word recognition (long term) as outcome measure

Repetition of the analysis with a MI approach yielded highly similar results and hence the same conclusions. Estimates and standard errors were highly comparable across parameters (Supplementary Table 2), including the interaction between late preterm and condition. Estimates for complete cases were: 14.80 (*6.78*); for MI: 12.71 (*5.53*).

If we used late preterm as marker for susceptibility, the effect size of *Living Letters* vs. control program was significantly larger in the susceptible group than in the non-susceptible group, .37 and -.28, respectively (Table 3). The full-term group profited even more from the control program (*Living Books*) than from *Living Letters*, as is indicated by an effect size of -.31 (Table 3).

Discussion

The main aim of this study was to test the reproducibility of the finding that children with mild perinatal adversities were not vulnerable, but in fact were more susceptible, to the learning context than were children without perinatal adversities. Previous research had demonstrated that children with perinatal adversities benefited from a computer-based remedial intervention with an adaptive feedback regime (*Living Letters*), and that effects remained well into Grade 1 (Van der Kooy-Hofland et al., 2012). Effects found in the previous small-scale study were large: 1.5 standard deviations (84% CI, .74, 2.15) on short term measures and 1.17 standard deviations (84% CI, .44, 1.8) on long-term measures. In the replication, we were unable to reproduce these effects despite the fact that the current study included a larger number of children with perinatal adversities. In the current study, effects for both the long and short term were small (.09 and .10, respectively) and non-significant.

In the Van der Kooy-Hofland et al. (2012) study, the small sample size precluded looking into the effects of *Living Letters* for children born late preterm and children small for gestational age separately, however, in the current study we were able to examine effects for these subsamples. We found significant effects for the children born later preterm, although the effects were notably smaller than in the Van der Kooy-Hofland et al. study (2012). Directly after receiving *Living Letters,* children born later preterm outperformed their peers, and they preserved this advantage well into Grade 1, without any further additional support in the period between the post-test and post-posttest. *Cohen's ds* were close to .40, both directly after the intervention and a year later, indicating that 65.5% of the treatment group would score above the mean of the control group (Cohen's U_3 index), and that there was a 61% chance that a person picked at random from the treatment group would have a higher score than a person picked at random from the control group (probability of superiority) (http://rpsychologist.com/d3/cohend/).

In sum, preterm children outperformed other children when they received the instruction program *Living Letters*, a program that provided instruction and guided practice in naming letters and phonemic awareness. However preterm children who did not receive *Living Letters* lagged behind their peers on the short term measure; they did not receive systematic instruction and guided practice in naming letters and phonemic awareness. These children were expected to learn through accidental events, such as writing their names or a parent naming letters. In sum, we found evidence for the theory that children born late preterm are more susceptible to the qualities of instructional environment, for better and for worse. Thus, the previous finding that mild perinatal adversities are not a vulnerability but a susceptibility factor was reproduced only for children born later preterm, but not for children small for gestational age.

Limitations through non-replication pathways

Findings of the current study only partially replicated previous findings. We cannot know whether the original experiment, the subsequent experiment, both, or none are correct or wrong (Nosek & Errington, 2017); a number of pathways to non-replication could potentially have influenced the findings presented. In search of an understanding *why* results were only partially replicated, we distinguish issues pertaining to a) methods, b) results and c) transferability.

a) Non-replication through methods

Compared to the previous study (Van der Kooy-Hofland et al., 2012), it is possible that there were small modifications in the experimental setup related to scaling up the research (Ioannidis, 2017). We had, for instance, less control over the distribution of sessions over time. While teachers were advised to do the programs twice a week, not all teachers followed up on this suggestion and some even compressed the intervention into a brief period of a few weeks. Even though this occurred for only a small proportion of the group, it may have caused a negative effect on learning outcomes. According to Hattie's meta-analysis (2015), spaced practice is much more effective than massed practice.

Furthermore, teachers may not have been as motivated in the current study as in the Van der Kooy-Hofland et al. (2012) study. In response to an open question in an online questionnaire that teachers completed after the intervention, teachers complained that '*For some children Living Letters took too long'*, and that '*Children did not understand why they had to keep playing the same game over and over again´*. In the Van der Kooy-Hofland et al. (2012) experiment, the researchers heard similar complaints, but the researchers maintained close contact with the teachers while the experiment was carried out and explained the importance of repetition each time teachers complained. Teachers may thus have been more motivated to encourage and challenge their pupils.

b) Non-replication through results

It is also possible that the differences in sensitivity and quality of the instruments used in the Van der Kooy-Hofland et al. study (2012) resulted in more robust detection of results compared to the instruments used in this study. Test administered by the researchers as in the Van der Kooy-Hofland et al. study (2012) may be more sensitive compared to tests administered by teachers, as was done in the current study.

The large-scale study also had limitations related to its size, one of which was the relatively large proportions of missing data. However, as indicated by analyses based on sets including data imputed with the help of innovative statistical techniques, results were robust.

c) *Non-replication through transferability*

Another reason for the non-reproducibility of prior findings may be differences in participant groups. For instance, in the original Van der Kooy-Hofland et al. (2012) study, the group with small perinatal adversities included a larger proportion of late preterm children (48%) than was included in the current study (39%), a difference that aligns the larger overall effect of perinatal adversities in the previous study. However, testing effects of *Living Letters* in randomly composed groups with perinatal adversities that were similar in composition to the Van der Kooy-Hofland et al. (2012) sample, and drawing such samples 50 times, did not produce evidence supporting this post-hoc explanation for the nonreproducibility of the effect of *Living Letters* in the group with mild perinatal adversities.

A more plausible hypothesis is that correlates of perinatal adversities are more important than the perinatal adversities themselves in shaping responses in experimental systems. For instance, a strong candidate for biological susceptibility to programs that instruct and guide, may be stress reactivity. Children with perinatal adversities are known to experience more stress than other children, however the correlation between stress and perinatal adversities is at most moderate. Stress scores of children with perinatal adversities thus may vary quite a bit across samples, which would mean that the susceptibility to stress-reducing programs like *Living Letters* would vary across studies. Perhaps children's stress levels were, by chance, high in the Van der Kooy-Hofland et al. (2012) sample with mild perinatal adversities.

Future directions

Results found in a subsample that included late preterm children supported the differential susceptibility hypothesis, suggesting that being born preterm was not a vulnerability but a susceptibility factor. However, it should be noted that these results were not the outcome of confirmatory analyses, and thus need further examination via new RCTs. A series of RCT designs, each targeting one of the three pathways, could provide insight into the reason for non-reproducibility. To test for the influence of method on non-reproducibility, the experiment of Van der Kooy-Hofland et al. (2012) could be replicated exactly, with, as the only difference, a larger proportion of children with mild perinatal adversities in the sample. This would make it possible to, even with a smaller sample, test for possible differences between children small for gestational age and children born late preterm. To examine if the sensitivity and quality of the posttest (i.e. non-reproducibility through results) might have accounted for the non-reproducibility of findings, a planned missing data design could be used. Such a design makes is possible to improve validity, while maintaining the large power associated with large samples (Graham, Taylor, Olchowski, & Cumsille, 2006). Lastly, the influence of transferability could be concretized by focusing on the suggested biological mechanism underlying increased susceptibility, instead of focusing on its phenotypical correlate. In this, case children's stress reactivity seems a plausible explanation for the positive effects of *Living Letters,* it may thus be promising to include this characteristic as a marker for biological susceptibility in follow-up RCTs rather than perinatal adversities.

References

Aisa, B., Tordera, R., Lasheras, B., Del Río, J., & Ramírez, M. (2007). Cognitive impairment associated to HPA axis hyperactivity after maternal seperation in rats. *Psychoneuroendocrinology*, 256-266.

Bolt, R., Van Weissenbruch, M., Lafeber, H., & Delemarre-Van de Waal, H. (2001). Glucocorticoids and lung development in the fetus and the preterm infant. *Pediatric Pulmonology*, 76-91.

Boyce, T., Chesney, M., Alkon, A., Tschann, J., Adams, S., Chesterman , B., . . . Wara, D. (1995). Psychobiologic reactivity to stress in childhood respiratory illnesses: Results of two prospective studies. *Psychosomatic Medicine*, 411-422.

Buske-Kirschbaum, A., Krieger, S., Wilkes, C., Rauh, W., Weiss, S., & Hellhammer, D. (2007). Hypothalamic-pituitary-adrenal axis function and the cellular immune response in former preterm children. *The Journal of Clinical Endocrinology & Metabolism*, 3429-3435.

Economides, D., Nicolaides, K., Linton, E., Perry, L., & Chard, T. (1988). Plasma cortisol and adrenocorticotropin in appropriate and small for gestational age fetuses. *Fetal Diagnosis and Therapy*, 158-164.

Goodman, S., Fanelli, D., & Ioannidis, J. (2016). What does research reproducibility mean? . *Science Translational Medicine*, 341-341ps12.

Gotlib, I., Joormann, J., Minor, K., & Hallmayer, J. (2008). HPA axis reactivity: A mechanism underlying the associations among 5-HTTLPR, stress, and depression. *Biological Psychiatry*, 847-851.

Graham, J., Taylor, B., Olchowski, A., & Cumsille, P. (2006). Planned missing data designs in psychological research. *Psychological Methods*, 323-343.

Graham, J. W., Olchowski, A., & Gilreath, T. (2007). How many imputations are really needed? Some practical clarifications of multiple imputation theory. *Prevention Science*, 206-213.

Guarini, A., Sansavini, A., Fabbri, C., & Savini, S. (2010). Long-term effects of preterm birth on language and literacy at eight years. *Journal of Child Language*, 865-885.

Gutbrod, T., Wolke, D., Soehne, B., Ohrt, B., & Riegel, K. (2000). Effects of gestation and birth weight on the growth and development of very low birthweight small for gestational age infants: A matched group comparison. *Archives of Childhood Disease: Fetal and Neonatal Edition*, 208-214.

Hattie, J. (2015). The applicability of visible learning to higher education. *Scholarship of Teaching and Learning in Psychology*, 79-91.

Hutton, J., Pharoah, P., Cooke, R., & Stevenson, R. (1997). Differential effects of preterm birth and small for gestational age on cognitive and motor development. *Archives of Disease in Childhood: Fetal and Neonatal Edition*, 75-81.

Kirkegaard, I., Obel, C., Hedegaard, M., & Henriksen, T. (2006). Gestational age and birth weight in relation to school performance of 10-year-old children: A follow-up study of children born after 32 completed weeks . *Pediatrics*, 1600-1606.

Kolb, B., & Whishaw, I. (2009). *Fundamentals of Human Neuropsychology.* New York : Worth Publishers.

Lansink, N., & Hemker, B. (2010). *Wetenschappelijke verantwoording van de toetsen Taal voor kleuters groep 1 en 2 uit het Cito Volgsysteem primair onderwijs.* Arnhem: Cito.

Levitt, N., Lambert, E., Woods, D., Hales, C., Andrew, R., & Seckl, J. (2000). Impaired glucose tolerance and elevated blood pressure in low birth weight, nonobese, young South African adults: Early programming of cortisol axis. *The Journal of Clinical Endocrinology & Metabolism*, 4611-4618.

Little, R.(1988). A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association*, 1198-1202.

Luke, D. (2004). *Multilevel Modeling.* London: SAGE Publications.

McCarton, C., Wallace, I., Divon, M., & Vaughan, H. (1996). Cognitive and neurologic development of premature, small for gestational age infant hrtough age 6: Comparison by birth weight and gestational age. *Pediatrics*, 1167-1178.

Morse, S., Zheng, H., Tang, Y., & Roth, J. (2009). Early school-age outcomes of late preterm infants. *Pediatrics*, 622-629.

Nosek, B., & Errington, T. (2017). Reproducibility in cancer biology: Making sense of replications. *eLife*.

Parker, R. (1990). Power, control, and validity in research. *Journal of Learning Disabilities*, 613-620.

Phillips, D., & Jones, A. (2006). Fetal programming of autonomic and HPA functioning: Do people who were small babies have enhanced stress responses? . *The Journal of Physiology*, 45-50.

Phillips, D., Barker, D., Fall, C., Seckl, J., Whorwood, C., Wood, P., & Walker, B. (1998). Elevated plasma cortisol concentrations: A link between low birth weight and the insulin resistance syndrome. *The Journal of Clinical Endocrinology & Metabolism*, 757-760.

Quigley, M., Poulsen, G., Boyle, E., Wolke, D., Field, D., Alfirevic, Z., & Kurinczuk, J. (2012). Early term and late preterm birth are associated with poorer school performance at age 5: A cohort study. *Archives of Childhood Disease: Fetal and Neonatal Edition*, 1-7.

Ready, D., & Chu, E. (2015). Sociodemographic inequality in early literacy development: The role of teacher perceptual accuracy. *Early Education and Development,* 970-987.

Reynolds, R., Walker, B., Syddall, H. Andrew, R., Wood, P., Whorwood, C., & Phillips, D. (2001). Altered control of cortisol secretion in adult men with low birth weight and cardiovascular risk factors. *The Journal of Clinical Endocrinology & Metabolism*, 245-250.

Sommerfelt , K., Andersson, H., Sonnander, K., Ahlsten, G., Ellertsen, B., Markestad, T., . . . Bakketeig, L. (2000). Cognitive development of term small for gestational age children at five years of age. *Archives of Disease in Childhood*, 25-30.

Stichting Perinatale Registratie Nederland. (2011). *Grote lijnen 10 jaar perinatale registratie Nederland.* Utrecht: Stichting Perinatale Registratie Nederland.

Van Baar, A., Vermaas, J., Knots, E., De Kleine, M., & Soons, P. (2009). Functioning at school age of moderately preterm children born at 32 to 36 weeks' gestational age. *Pediatrics*, 251-257.

Van der Kooy-Hofland, V., Van der Kooy, J., Bus, A., Van IJzendoorn, M., & Bonsel, G. (2012). Differential susceptibility to early literacy intervention in children with mild perinatal adversities: Short- and long-term effects of a randomized controlled trial. *Journal of Educational Psychology*, 337-349.

Supplementary Table 1. *Estimates and Standard Errors in short- and long-term analyses using a complete case (CC) and a multiple imputation approach (MI) for analyses with MPA as susceptibility factor*

	Short term				Long term			
	CC	*p*	MI	*p*	CC	*p*	MI	*p*
Intercept	-.28 (*.21*)	.176	-.24 (*.11*)	.048	37.50 (*6.55*)	<.001	30.34 (*2.45*)	<.001
Cohort	.05 (*.11*)	.657	.07 (*.07*)	.304	12.18 (*3.41*)	<.001	.44 (*2.68*)	.875
Pretest	.63 (*.09*)	<.001	.63 (*.08*)	<.001	-2.65 (*2.57*)	.303	6.13 (*2.56*)	.063
MPA*	-.22 (*.14*)	.111	-.19 (*.11*)	.089	1.02 (*3.96*)	.796	-2.52 (*3.11*)	.450
Condition	-.19 (*.11*)	.068	-.18 (*.10*)	.082	-3.14 (*2.87*)	.275	-7.21 (*1.30*)	<.001
MPA X Condition	.35 (*.19*)	.066	.34 (*.15*)	.036	5.82 (*5.34*)	.277	7.52 (*4.88*)	.182

*MPA= mild perinatal adversities

Supplementary Table 2. *Estimates and Standard Errors in short- and long-term analyses using a complete case (CC) and a multiple imputation approach (MI) for analyses with LP and SGA as susceptibility factors*

	Short term				Long term			
	CC	*p*	MI	*p*	CC	*p*	MI	*p*
Intercept	-.26 (*.21*)	.207	-.44 (*.21*)	.058	32.68 (*4.54*)	<.001	30.50 (*1.60*)	<.001
Cohort	.03 (*.11*)	.756	.10 (*.10*)	.318	-2.44 (*2.51*)	.331	-.27 (*.99*)	.788
Pretest	.62 (*.09*)	<.001	.69 (*.08*)	<.001	7.00 (*2.26*)	.002	8.57 (*1.01*)	<.001
LP*	-.29 (*.20*)	.140	-.25 (*.28*)	.379	-4.86 (*4.71*)	.304	-2.66 (*3.47*)	.454
SGA*	-.10 (*.15*)	.524	-.07 (*.11*)	.515	3.49 (*4.06*)	.391	1.21 (*2.63*)	.652
Condition	-.18 (*.10*)	.095	-.16 (*.10*)	.140	-6.35 (*2.56*)	.014	-3.22 (*2.05*)	.139
LP X Condition	.54 (*.27*)	.048	.44 (*.37*)	.263	14.80 (*6.87*)	.032	12.71 (*5.53*)	.037
SGA X Condition	.11 (*.21*)	.612	.14 (*.16*)	.391	4.47 (*5.55*)	.422	-.07 (*2.80*)	.979

*LP = late preterm, SGA= small for gestational age

2

Chapter 3

Planned missing data in early literacy interventions:

A replication study with an additional gold standard

Introduction: In a previous large scale RCT into the effects of a digital early literacy program, we found that children born late preterm were susceptible for the qualities of the learning environment: They fell behind peers when in a control condition, but outperformed them when assigned to the early literacy program. Results of the study, however, deviated in some respects from previous research, demonstrating the need for replication. Replication, however, often is complicated by a range of obstacles such as the resources needed to carry out an additional large-scale study, especially if that study requires administration of high-quality but time-intensive (and thus costly) reference measures. Use of a planned missing data approach where these reference measures are incomplete can help to address these limitations. *Methods:* In the current study, we use a planned missing data approach to examine whether results of the original RCT replicated when using additional, higher-quality, outcome measures. The high-quality measures were more closely aligned with the measured construct of early literacy and language performance, and thus were potentially more sensitive to changes in performance. Because the high-quality measures were more costly and time-intensive to administer, they were administered to a randomly selected subsample of children. We refer to the scores from these measures as "gold standard data." Three gold standard models were fitted, varying in how much gold standard data were included and in how closely the extra measurements approached the skill targeted by the intervention (i.e. construct validity). *Results:* Two out of three gold standard models showed improved model efficiency as compared to the model without gold standard data. Only the model with the broadest gold standard data did not lead to improvement: in this model efficiency even diminished. In one of two efficient gold standard models main results could be replicated, in both models estimates were comparable to the model without gold standard data. *Conclusion:* Results could be replicated using a gold standard approach. Estimates remained comparable to those found without using gold standard data. Previous results were thus not further approached. Additionally, gold standard data can only be used to improve model efficiency in RCT-designs, when gold standard data show sufficient convergent validity. Planned missing data designs can thus be used to replicate experimental results, but only when only gold standard testing closely approximating the trained skills at hand are included.

Developing and validating effective interventions is a central goal of educational research. Experimental designs (random-controlled trials; RCTs) are, by definition, the most powerful designs for testing the effectiveness of interventions, but they may be hard to realize. Factors such as small sample sizes or the use of outcome measures with suboptimal validity (Curtis, et al., 2015) can compromise the trustworthiness of the experimental results. To examine the influence of such factors on the results of a study, one can replicate the study with a new, larger sample, and with higher-quality measures; however, such replications are, by their very nature, time and resource intensive as they require recruitment and data collection from a new sample. An alternative approach is to use a planned missing data approach and administer high-quality, time intensive measures to a randomly selected subset of participants in the original study. Although not yet widely deployed in RCT-studies (Kegel & Rippe, under review), a planned missing data approach holds promise for increasing internal validity in experimental studies (Rhemtulla & Little, 2012).

The current study employs a planned missing data approach to "replicate" the results of a large-scale experiment that examined the differential susceptibility of kindergarten children to an educational intervention in early literacy (see Merkelbach, Plak & Rippe, 2018). For a randomly selected subsample of the original participants, additional high-quality, time-intensive assessments were administered. Data from these assessments allowed for more sensitive and precise examination of intervention effects.

Differential Susceptibility

Over the past decade, the differential susceptibility model (Belsky & Pluess, 2009) has become universally applied in behavioral sciences. Central to this model is the notion that individuals carrying certain genetic or neurobiological markers may be more susceptible to the quality of their environment (e.g. various types of interventions), both for better and for worse. In contrast to the common diathesis stress model (Zuckerman, 1999), which postulates that subgroups with certain biological vulnerabilities will fall behind when conditions are adverse, the differential susceptibility postulates that subgroups will fall behind when conditions are adverse, but will succeed – and even surpass less susceptible peers – when conditions are favorable.

Many studies of differential susceptibility target the dopamine system, specifically the DRD4-gene. A number of these studies have focused on the differential effects of the home environment and childhood upbringing experiences (e.g. Beach, Brody, Lei, & Philibert, 2010; Bakermans-Kranenburg & Van IJzendoorn, 2011). These studies have demonstrated that children carrying a certain allele of the DRD4-gene, namely the 7-repeat allele, which is associated with decreased efficiency of dopamine production in the prefrontal cortex, are more susceptible to the qualities of the home environment and to their upbringing. Other studies of the DRD4-gene have focused on the differential effects of the educational environment. These studies have demonstrated that four-year-old children carrying the 7-repeat allele are more susceptible to a digital intervention promoting alphabetic knowledge and phonemic awareness than children not carrying the 7-repeat allele (Kegel, Bus, & Van IJzendoorn, 2011), and that kindergarten children carrying the 7-repeat allele are more susceptibility to digital animated storybooks than children not carrying the 7-repeat allele (Plak, Kegel, & Bus, 2015; Plak, Merkelbach, Kegel, Van IJzendoorn, & Bus, 2016).

Differential susceptibility in children with perinatal adversities

In recent years, studies of differential susceptibility have targeted factors other than the genetic make-up of the child, such as the differential susceptibility of children with mild perinatal adversities to the educational environment. In a small-scale experimental study focusing on kindergartners (N = 100) with and without perinatal adversities, the effects of a digital program stimulating letter knowledge and phonemic awareness (*Living Letters*) were compared to the effects of a control program with digital storybooks (*Living Books* (Van der Kooy-Hofland, Van der Kooy, Bus, Van IJzendoorn, & Bonsel, 2012). Children with mild perinatal adversities in the study were children who were small for gestational age at birth and/or were born late preterm (born between the 34th and 38th week of pregnancy). Results revealed that children with perinatal adversities were differentially susceptible to the *Living Letter* intervention: In the control condition (*Living Books*), the children performed significantly less well than their peers, but in the experimental (*Living Letters*) condition, they significantly outperformed their peers. Impressively, the effects for the perinatal adversities group were large not just in the short term (*Cohen's d* = 1.24) but also one year after the intervention (*Cohen's d* = 1.11).

In a large-scale replication of the van der Kooy-Hofland et al. (2012) study, differential susceptibility of children with mild perinatal adversities to *Living Letters* was once again examined (Merkelbach, et al., 2018). Participants in the study were 439 kindergartners, 142 of whom were children with perinatal adversities. Within the perinatal adversities group, 49 were children born late preterm and 102 were children born small for gestational age. Unlike the van der Kooy-Hofland et al. (2012) study, results did not reveal differential susceptibility for the perinatal diversities group as a whole, nor for the children born small for gestational age. However, differential susceptibility was found for the children born late preterm. Although the effect sizes in the replications study were substantial, they were considerably more modest than those found in the Van der Kooy-Hofland et al.(2012) study (2012): *Cohen's d* = .38 in the short term, and *Cohen's d* = .37 in the long term.

Planned missing data approach with gold-standard measures

The discrepancies between the van der Kooy-Hofland et al. (2012) and Merkelbach et al. (2018) studies might be due to differences in study design. In the Van der Kooy-Hofland et al. (2012) study, researchers supervised the implementation of the intervention, ensuring that digital sessions took place twice a week. In the replication study, as a practical consequence of the large sample size, teachers scheduled interventions sessions, which resulted in a less consistent dispersion across time, which in turn might have resulted in lower learning gains and thus reduced effect sizes. Additionally, in the Van der Kooy-Hofland et al. (2012) study, researchers administered posttests, whereas in the Merkelbach et al. (2018) study, teachers administered the posttest. Finally, in the Van der Kooy-Hofland et al. (2012) study, the posttests consisted of a large number of items (k = 40), while in the Merkelbach et al. (2018) study, the posttests consisted of a much smaller number of items (k = 23). Fewer items on an assessment is associated with lower reliabilities, higher bias in scores, and less differentiation in skill levels (Cronbach, 1951), under the assumption that items are of equivalent quality. The administration of posttests by the teachers rather than the researchers, and the smaller number of posttest items might have resulted in more noise in the Merkelbach et al. (2018) replication study and might have influenced the size of effects. One way to address the potential limitations of the Merkelbach et al. (2018) study is to use a planned missing data approach with use of gold-standard measures.

3

Planned missingness is a way to improve validity of results while maintaining the large power associated with larger sample sizes (Graham, Taylor, Olchowski, & Cumsille, 2006). A planned missing data approach with gold-standard measures involves the administration of an additional set of high-quality, 'gold-standard,' measures to a randomly selected subgroup of participants (Little & Rhemtulla, 2013). 'Gold standard measures' are measures that are typically more expensive and time consuming to collect than other measures, but that are likely to provide more sensitive and valid information on the construct of interest. In a planned missing data approach, the selection of participants who are administered the gold-standard measures is determined by the researchers ahead of time in a random fashion. Planned missingness, thus, relies on the presumption that gold-standard measurement data meet the criteria of being missing completely at random (MCAR), and hence that missingness is not associated with any bias (Garnier-Villarreal, Rhemtulla, & Little, 2014). Using scores from the less expensive (but possibly biased) measures as an auxiliary to the scores from the reliable, non-biased, gold standard, measures, a shared variance factor between the measures can be identified (Little & Rhemtulla, 2013). This shared factor is assumed to be a valid estimate of performance and is thus be expected to result in the best-fitting model (Garnier-Villarreal et al., 2014), and, consequentially, in the most accurate descriptions of individual effects. A recent study has suggested that using a gold standard in a large-scale experimental GxE design into the effects of digital interventions improved model fit and offered the best basis for individual assessment (Rippe & Merkelbach, under review).

Current study

In the current study, we use a planned missing data approach with gold-standard measures to reexamine the data collected in Merkelbach et al. (2018). Specifically, for a randomly selected subsample of children from the Merkelbach et al. (2018) study, trained research assistants administered an additional set of gold standard early literacy measures in the areas of letter knowledge, phonemic awareness, and writing. We expect that a missing data approach with gold-standard measures would offer a clearer and less biased picture of effects.

Three research questions were addressed in the study:

1. Can we replicate interactions between intervention (i.e. *Living Letters*) and susceptibility factor (i.e. late preterm), as found in the Merkelbach et al. (2018) replication study, utilizing a planned missing data approach?
2. Does the planned missing data approach reveal interactions between *Living Letters* and other mild perinatal adversities, specifically being small for gestational age at?
3. Do effect sizes in the replication study, now based on an extended set of tests, approach effect sizes found in the van der Kooy-Hofland et al. (2012) study?

Methods

In this section, we describe the methods for the larger study (see also the Merkelbach et al. (2018) study), but also include the information specific to the current study, that is, to the missing data / gold-standard replication study.

Participants

The initial sample as used in the Merkelbach et al (2018) study consisted of 981 five-year-old children. Subjects were excluded from analysis mainly due to incomplete perinatal information. The final sample, for which complete data on the predictive variables and the immediate post-test were available, consisted of 439 children from 147 different schools (55.5% boys; mean age: 66.78 months (SD = 3.88)). On average, there were 1 to 2 children per classroom in the study ($Mean$ = 1.66 children per classroom, SD = .89). Gold standard measures were administered to a randomly selected (32.6%, n= 143) subsample of children (57.3% boys; mean age: 66.42 months (SD = 3.88)). Children in the subsample were from 54 different schools.

Design

The study employed an experimental design. Children were randomly assigned to either the experimental condition (i.e. *Living Letters*) or the control condition (i.e. *Living Books*). For the larger study, teachers coordinated sessions and administered post-testing. Teachers were not informed about which program was considered to be the target condition or control condition, but were aware of the condition to which children were assigned. The research assistants who administered the gold standard measures for the selected subsample of children were blind to the condition to which the child had been assigned.

Procedure

Data collection took place in two consecutive school years (2012/2013 and 2013/2014). From August to February, schools were recruited by sending out flyers and letters containing information about the content and purpose of the study through both email and mail. Participating schools were offered three months of free access to all intervention programs, which normally require a paid subscription (http://www.bereslim.nl). When teachers agreed to participate, they were asked to select pupils from their classroom with poor language/literacy skills, for instance pupils who were not yet able to write their proper name, to rhyme, to name a few letters, and to identify sounds in words. Teachers were told that it was preferable that these children scored below the 40th percentile (between 0 and 59) on a standardized Cito language/literacy test (CLT) that was administered in January in the schools (Lansink & Hemker, 2010). If there were not enough children scoring below the 40th percentile, teachers were asked to include other children who they believed were in need of additional help with early literacy skills. Parents provided informed written consent for the child's participation in the study. In year 1, near the end of the study, parents also were asked for consent for retrieving perinatal information. Only 43% of parents provided consent for receiving perinatal information – perhaps due to the fact that the request was made at the end of the study. In the second year, parental consent for the child's participation and for retrieving perinatal information both were requested at the beginning of the study. Most parents (94%) provided consent for retrieving perinatal information in the second year of the study.

Similar to the Van der Kooy-Hofland et al. (2012) study, the current contrasted *Living Letters* with *Living Books*. Other conditions included in the larger study are not discussed in this manuscript (see Merkelbach et al., 2018 for details). Children were randomly assigned to condition by the researchers. The intervention sessions took place once a week, and were spread over a period of approximately eight to twelve weeks. Except for logging in, children worked on their own without adult assistance. During the sessions, children wore headphones in order to prevent being disturbed by other children. Children worked with the mouse and did not have to make use of the keyboard.

At the end of the eight to twelve week intervention period, teachers administered three digital tests measuring alphabetic knowledge and phonemic awareness (i.e. phonological skills, word recognition, and decoding) to participating children on an individual basis. Testing took approximately ten minutes. Teachers were not allowed to help children, but were expected merely to mark the child's responses as either correct or incorrect.

In order to carry out the planned missingness design, we collected additional data from a battery of gold standard measures from a randomly selected subsample of children. In the total cohort just over 40% of children were randomly selected and received additional testing. By chance in the subsample included in the current study (only those children meeting criteria to answer the raised research questions – e.g. those assigned to the right conditions, *n* = 439) this percentage was somewhat lower: around 33%. These additional tests were administered by highly trained research assistants. The gold-standard measures consisted of three early literacy tests that targeted alphabetic knowledge and phonemic awareness. These measures are described in in the Measures section.

Intervention programs

The target program, *Living Letters*, was designed to promote knowledge of the alphabetic principle and phonemic awareness in kindergartners. Two main characters, a boy and a girl, explain the assignments and an online tutor, the boy's teddy bear, provides adaptive feedback after each assignment. Feedback is also given when the assignment is completed correctly. After the child provides the correct response, or the correct response is modeled, the teddy bear confirms that the answer is correct and explains why. If children provide incorrect responses in the games, the online tutor (the teddy bear) immediately provides feedback. In case of an incorrect response, three levels of feedback are provided: (1) first, repeating instructions; (2) second, providing cues to answer the question; (3) third, modeling the correct response. Feedback is provided in all games of *Living Letters.* In the first 22 games of *Living Letters*, children practice recognizing their own written name (or 'mamma') among other symbol strings or scribbles. The subsequent six games focus on the sound of the first letter of the child's name. In the last twelve games, children select pictures of words that start or end with the first letter of their own name.

Control children received *Living Books* during the same period of time. *Living Books* includes eight digital, animated, age-appropriate stories based on high-quality children's books. Each story is 'read' twice to the child by a computerized voice while children watch animations and listen to background sounds and music that support comprehension of the story content. The text is not presented as print on screen but only orally. Each reading session is interrupted four times so that children can answer two questions about the story events and two about difficult words in the text. After answering each

question, children receive immediate feedback, as well as positive feedback in the form of compliments, regardless of their individual performance.

Measures

Pretest

At pretest, the Cito Literacy Test for Kindergarten Pupils (CLT) was used. The CLT is a group-administered test applied in January/February in the schools. The test consists of 60 paper-pencil questions measuring a range of language and literacy skills: vocabulary, critical listening, rhyming, hearing the first or last word in a sentence, sound blending, writing conventions, and prediction of book content based on book cover (Lansink & Hemker, 2012). Children's pretest score was coded as scoring among the lowest 25% (score of 59 or below) or average (score of 59 and beyond).

Posttest: Entire sample

As mentioned earlier, a battery of early literacy measures was administered by teachers to all participating children in the study. The battery included a phonemic awareness task, a letter knowledge task, and a word recognition task.

Phonemic awareness. The Phonemic Awareness Task included five items. Children identified the first sound of five words (e.g. muis [mouse]) while pictures of the words were shown on the computer screen. *Cronbach's α* was .76.

Letter knowledge. Children identified ten letters presented on screen (i.e. *s*, *k*, *a*, *p*, *r*, *o*, *v*, *m*, *t*, & *n*). *Cronbach's α* was .83.

Word recognition. Children were asked to match a printed word with picture. For each of six words (e.g. dak [roof]) there were four options (one correct, three incorrect) from which they could choose. The incorrect options varied in systematic way: no letter correct (lom), first letter correct (dor), first and last letter correct (dek). *Cronbach's α* was .83.

Aggregate measure. Principal component analysis (PCA) applied to the three tests resulted in one component explaining 67.59% of the variance. Component loadings ranged from .74 to .86. Scores were transformed into standardized weighted averages, in which a higher score indicating better alphabetic skills.

Posttest: Gold standard measures for randomly selected subsample

In addition to the measures described above, three gold-standard measures were administered by research assistants to a randomly selected subset of the sample. These three measures included a vocabulary-, a word recognition-, and a writing measure.

Vocabulary. The vocabulary test consisted of 25 items in which a sentence derived from a digitally animated storybook, was read to the child, after which a target word was repeated, and children were asked to give a definition of the word (e.g. ' 'Are you lost little one?' the bear asked kindly. What does *lost* mean?'). Answers were scored as correct (1), partly correct (.5), or incorrect (0). *Cronbach's α* was .72. For no item did deleting the item result in a higher *Cronbach's α*.

Word recognition. Ten word recognition items were administered to the students for the gold-standard word recognition test, including the six items used in the teacher administered test and four new items. As with the task administered by the teachers, children were asked to match a printed word with a picture. For each word there were four options, and the incorrect options varied systematically. *Cronbach's α* was .74.

Writing. The writing test, developed by Bus and Levin (2003), consisted of six items asking children to write their own names and five other short words. Items were scored on a seven-point scale with a score of 0 indicated drawing and a score of 6 indicating a completely correctly written word. *Cronbach's α* was .80)

Aggregate measures. A total of three planned missing data models was fitted. In the first model, a two-factor approach was used, in which the word knowledge task was considered as one factor (measuring vocabulary), and word recognition and writing were combined into another factor (targeting alphabetic knowledge and phonemic awareness). In the second model, only the factor measuring alphabetic knowledge and phonemic awareness was used. In the third model, only the writing score was used because it most closely approached the skills trained by *Living Letters*.

Statistical analyses

Basic analysis

As with the previous studies, to test effects of *Living Letters,* a multilevel approach using mixed models was applied to account for variance attributable to school-level characteristics (Luke, 2004). We employed a likelihood ratio test to examine model improvement when intercepts or intercepts and slopes were allowed to vary across schools. The following variables were included in the analyses: pretest score, condition, small for gestational age, late preterm, and two two-way interactions (small for gestational age*condition, late preterm*condition).

Applying the gold standard

For model estimations, we used the *lavaan* package, Beta version 5.20 in R version 3.3.1. The number of EM iterations was set to a maximum of 5000. Full information maximum likelihood was used to account for missing data. To obtain stable and robust estimates of the parameter Standard Errors, the proportional bootstrap was used with 1000 runs.

To evaluate replicability when using the gold standard, three model variants were fitted to the data. In the first model a very general approach was considered. We

explored if adding broad gold standard literacy measurements, not directly related to the intervention, could improve the model fit. Therefore, all measurements administered during the gold standard test sessions were considered. Test scores were split into two factors describing different components of early literacy development (i.e. one factor focussing on vocabulary (Cronbach's α = .78), and another factor focussing on word recognition and writing (Cronbach's α = .89), which are both skills relying on alphabetic knowledge and phonemic awareness). In the second model we applied a more focused approach: only the second factor, focussing on same skills as were assessed by the teacher, was entered. In the third model only writing was considered, because this measure is known to be a strong indicator of alphabetic knowledge.

Comparing model fit

The fit of the model without gold standard data, and the fit of the three gold standard models were compared. To determine the absolute fit of each model the Comparative Fit Index (CFI), Normed Fit Index (NFI), standardized Root Mean Residual (sRMR), and Root Mean Square Error of Approximation (RMSEA) were inspected. The CFI and NFI should be as high as possible (ideally above .90), while the sRMR and the RMSEA should be as low as possible (ideally below .06). To evaluate the relative efficiency of the models, the Akaike's Information Criterion (AIC) and the Bayesian Information Criterion (BIC) are inspected. Models with lower values on these statistics are more efficient.

Comparing results. After determining if the planned missing data models were able to improve the fit of the model without the gold standard, and selecting the best fitting gold standard model, individual parameters obtained from the model without the gold standard and from the best fitting model were compared to determine whether results could be replicated using a gold standard approach.

Results

Sample characteristics

Because the same set of data is used, sample characteristics of participants in the original analysis (N = 439) are similar to those reported in the Merkelbach et al. study (2018). These characteristics are reported in Table 1.

Table 1. *Sample characteristics for the complete group, and compared per condition*

	Complete group (*N* = 439)	Experimental: *Living Letters* (*n* = 230)	Control: *Living Books* (*n* = 209)	*p*
Male	55.4%	53.9%	56.9%	.524
Age (in months)	66.81 (*4.23*)	59.53 (*7.80*)	66.86 (*4.30*)	.793
Father's education (max = 6)	3.71 (*1.38*)	3.74 (*1.42*)	3.69 (*1.35*)	.721
Late preterm	12.5%	12.6%	12.4%	.958
Small for gestational age	23.2%	22.6%	23.9%	.745
CLT pretest (raw score)	59.85 (*8.06*)	59.53 (*7.80*)	60.22 (*8.35*)	.372
CLT pretest (percentage low)	49.7%	50.4%	48.8%	.733
Alphabetic knowledge posttest (z-score)	.00 (*1.00*)	-.04 (*1.00*)	.04 (*1.00*)	.389

From this group of participants, we randomly selected a subsample consisting of 144 children (32.8%). In this subsample gold standard testing was administered. Sample characteristics of this gold standard subsample are reported in Table 2, and are, as is expected when missingness is at random, comparable to characteristic found in the complete sample. Within the gold standard sample, no differences between conditions regarding background characteristics (e.g. educational level of the father and age of the child) were found. However, on two of the three gold standard measures (i.e. word knowledge and word recognition) children in the *Living Books* (i.e. control) condition had higher scores than children in the *Living Letters* condition. This might suggest that in general, *Living Books* might have been better in stimulating these skills.

Table 2. *Sample characteristics in the gold standard sample, for the complete group, and compared per condition*

	Complete group (*N* = 144)	Experimental: *Living Letters* (*n* = 75)	Control: *Living Books* (*n* = 69)	*P*
Male	56.9%	54.7%	59.4%	.565
Age (in months)	66.49 (*3.97*)	66.67 (*4.27*)	66.31 (*3.64*)	.549
Father's education (max = 6)	3.76 (*1.38*)	3.80 (*1.43*)	3.71 (*1.32*)	.706
Late preterm	9.7%	10.7%	8.7%	.690
Small for gestational age	20.8%	18.7%	23.2%	.504
CLT pretest (raw score)	59.78 (*6.42*)	59.20 (*6.09*)	60.42 (*6.74*)	.256
CLT pretest (percentage low)	48.6%	54.7%	42.0%	.130
Alphabetic knowledge posttest (z-score)	.05 (*.93*)	.05 (*.93*)	.04 (*.94*)	.946
Word knowledge (gold standard)	.68 (*.12*)	.66 (*.12*)	.71 (*.12*)	.024
Word recognition (gold standard)	2.28 (*.51*)	2.17 (*.51*)	2.39 (*.50*)	.009
Writing (gold standard)	4.11 (*.88*)	4.05 (*.91*)	4.17 (*.85*)	.354

Comparing model fit

The gold standard models, as well as the model without the gold standard, show relatively good absolute fit. With the exception of the first, broad, gold standard model (including both factors), CFI and NFI-values are above .90 in all models, while sRMR and RMSEA are below .60 for all four models (Table 3). From this we might consider the fit of the model without the gold standard, as well as the fit of gold standard model 2 (one factor including word recognition and writing), and gold standard model 3 (writing only), as satisfactory. However, when looking at the relative efficiency of the models, gold standard model 2 and gold standard model 3 are superior to the original model and to the first gold standard model, showing lower values on both the AIC and the BIC-index (Table 3). Gold standard model 1, including two factors, even diminished the fit of the original model, because values on both the AIC- and the BIC-index were higher for this model than for the model without the gold standard.

Table 3. *Comparing model fit*

Model	*CFI*	*NFI*	*sRMR*	*RMSEA*	*AIC*	*BIC*
Main model	1	1	.00	0	16172.29	16391.78
Model 1; two factors*	.89	.85	.04	.04	16673.24	16927.65
Model 2; one factor**	1	.98	.01	0	15469.81	15709.26
Model 3; writing	1	.99	.01	0	15473.55	15712.99

*factor 1 = word knowledge, factor 2 = word recognition and writing, **factor = word recognition and writing

Overall, we can conclude that both gold standard model 2 and gold standard model 3 are an improvement, compared to the model without the gold standard, while gold standard model 1 shows a deterioration of model efficiency.

Replication of results

Because only gold standard model 2 and 3 showed comparable absolute fit with the model without the gold standard, and showed an increase in relative efficiency as compared to the model without the gold standard, only results of these models are compared to the results of the model without the gold standard (Table 4).

Table 4. *Comparing results of analysis in model without gold standard, and gold standard models 2 and 3*

	Model without gold standard		Model 2 (1 factor)		Model 3 (writing)	
	Est (*SE*)	*p-value*	Est (*SE*)	*p-value*	Est (*SE*)	*p-value*
Pretest	.32 (*.04*)	<.001	.46 (*.04*)	<.001	.48 (*.04*)	<.001
Late preterm	-.02 (*.07*)	.731	-.08 (*.07*)	.237	-.07 (*.07*)	.333
Small for gestational age	-.07 (*.08*)	.333	-.05 (*.07*)	.455	-.06 (*.08*)	.429
Condition	-.02 (*.08*)	.748	.02 (*.08*)	.781	-.01 (*.08*)	.869
LP * condition	.18 (*.08*)	.027	-.15 (*.10*)	.138	-.19 (*.09*)	.031
SGA * condition	.06 (*.07*)	.431	.04 (*.09*)	.733	-.03 (*.08*)	.707

Estimates (and standard errors) are highly comparable across all three models, showing that in general the analysis yielded similar results. In all models, pre-test was a significant predictor. However, the interaction between late preterm and condition (*Living Letters* vs. *Control program*), which was significant in the model without the gold standard (p = .027), was also significant in gold standard model 3 (p = .031), but failed to reach significance in gold standard model 2 (p = .138). The interaction between small for gestational age and condition was not significant in either model 2 (p = .733) or model 3 (p = .707), consistent with the results of the model without the gold standard (p = .431).

Discussion

The aim of the current study was to examine whether results of a large-scale intervention could be replicated, specifically whether effects of the intervention would be moderated by child characteristics. Specifically, we tested whether we could replicate the interaction effect between *Living Letters*, a digital intervention program promoting alphabetic knowledge and phonemic awareness, and late preterm birth, using a planned missing data approach. Results were replicated, however not in all planned missing data models fitted to the data.

To test if replication was possible, three planned missing data models, differing in the amount and accuracy of gold standard data included, were fitted to the data. All models showed relatively good absolute fit, however none better than the model without gold standard data. Only in two of the fitted models the relative efficiency of the model improved when compared to the model without the gold standard. In one of the planned missing data models – the model in which the broadest range of gold standard data was included – efficiency even diminished when compared to the model without the gold standard. This reduction implies that the gold standard data did not approximate the skill-set stimulated by *Living Letters* (i.e. word knowledge). These findings demonstrate that obtaining more, but possibly less relevant, information does not always lead to model improvement, and thus that selection of tests to serve as gold standard measurements should take place with caution. Because gold standard data are assumed to be measured without bias (Garnier-Villareal et al., 2014), high quantities of information with limited validity are not preferable to using less information with higher levels of construct validity. In the two models that showed improvement of model efficiency, estimates and thus effect sizes were comparable to those in the model without the gold standard (described in Merkelbach et al., 2018). The main finding, a significant interaction between condition and late preterm birth, was replicated in only one of the planned missing data models, that is, the model including only the measurement with the highest level of convergent validity – writing. However, this effect disappeared in the other, somewhat broader, model that included both word recognition and writing.

Additionally, we explored whether using a planned missing data design would reveal interactions between *Living Letters* and being small for gestational age. However, as in the analysis without missing data, in all three planned missing data models, this interaction remained non-significant. Because p-values are very large (around .80), further improvement of power is not expected to result in the manifestation of this interaction.
Lastly, we tested whether effect sizes would approach effects found in the Van der Kooy-Hofland et al. (2012) study if a planned missing data approach was used to improve design validity. We would expect clearer effects if bias, and thus measurement error, might possibly explain the reduced effect sizes (Gerhart, Wright, McMahan, & Snell, 2000) of the Merkelbach et al. (2018) replication study when compared to the original experiment Van der Kooy-Hofland et al. (2012) study. However using a planned missing data approach did not result in the emergence of clearer effects. We might thus conclude that bias and measurement error cannot explain the discrepancy between the Merkelbach et al. (2018) and Van der Kooy-Hofland et al., (2012) studies. This suggests that neither the way teachers administered tests, nor the validity of the original posttests, were factors that likely influenced the results. It is possible that the discrepancies in the results between these two studies might thus be explained by other factors, such as the quality of implementation of the intervention (i.e. less consistent dispersion of sessions when teachers coordinate the intervention).

Conclusion

In the current study we tested if, using a planned missing data approach, we could replicate results of a large scale RCT examining the differential effects of a digital early literacy intervention focused on alphabetic skills and phonemic awareness. Three planned missing data models were fitted to the data of the large scale RCT. In only one of the models did model fit improve, whilst results could also be replicated. Adding gold standard data did not result in effects sizes approaching those found in the previous small scale study, suggesting that bias and measurement error did not account for the differences in effect sizes found between the original and the replication study.

In the model in which replication was possible, only gold standard data with high convergent validity were included (i.e. writing), while gold standard measures approaching the skill trained by the intervention less closely (e.g. word knowledge) were not included. Planned missing data approaches in replicating RCT-studies can thus be useful, but only when used with care: Previous findings might be replicated using a planned missing data approach, however, only when only gold standard testing closely approximating the trained skills at hand are included

References

Bakermans-Kranenburg, M., & Van IJzendoorn, M. (2011). Differential susceptibility to the rearing environment depending on dopamine related genes: New evidence and a meta-analysis. *Development and Psychopathology*, 39-52.

Beach, S., Brody, G., Lei, M.-K., & Philibert, R. (2010). Differential susceptibility to parenting among African American youths: Testing the DRD4 hypothesis. *Journal of Family Psychology*, 513-521.

Belsky, J., & Pluess, M. (2009). Beyond diathesis stress: Differential susceptibility to environmental influences. *Psychological Bulletin*, 885-908.

Bus, A., & Levin, I. (2003). How is emergent writing based on drawing? Analyses of children's products and their sorting by chilren and mothers. *Developmental Psychology*, 891-905.

Curtis, M., Bond, R., Spina, D., Ahluwalia, A., Alexander, S. & Giembycz, M. (2015). Experimental design and analysis and their reporting: new guidance for publication in BJP. *British Journal* of Pharmacology, 3461-3471.

Cronbach, L. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 297-344*.

Garnier-Villarreal, M., Rhemtulla, M., & Little, T. (2014). Two-method planned missing designs for longitudinal research. *International Journal of Behavioral Development*, 411-422.

Gerhart, B., Wright, P. McMahan, G., & Snell, S. (2000). Measurement error in research on human resources and firm performance: how much error is there and how does it influence effect size estimates? *Personnel Psychology*, 803-834.

Graham, J., Taylor, B., Olchowski, A., & Cumsille, P. (2006). Planned missing data designs in psychological research. *Psychological Methods*, 323-343.

Kegel, C., Bus, A., & Van IJzendoorn, M. (2011). Differential susceptibility in early literacy instruction through computer games: The role of the dopamine D4 receptor gene (DRD4). *Mind, Brain, and Education*, 71-78.

Lansink, N., & Hemker, B. (2010). *Wetenschappelijke verantwoording van de toetsen Taal voor kleuters groep 1 en 2 uit het Cito Volgsysteem primair onderwijs.* Arnhem: Cito.

Little, T., & Rhemtulla, M. (2013). Planned missing data designs for developmental researchers. *Child Development Perspectives, 199-204*.

Merkelbach, I., Plak, R., & Rippe, R. (2018). Reproducibility of young learners' susceptibility to the learning context. *Learning and Individual Differences,* 167-175.

Plak, R., Kegel, C., & Bus, A. (2015). Genetic differential susceptibilty in literacy-delayed children: A randomized controlled trial on emergent literacy in kindergarten. *Development and Psychopathology*, 69-79.

Plak, R., Merkelbach, I., Kegel, C., Van IJzendoorn, M., & Bus, A. (2016). Brief computer interventions enhance emergent academic skills in susceptible children: A gene-by-environment experiment. *Learning and Instruction*, 1-8.

3

Rhemtulla, M., & Little, T. (2012). Planned missing data designs for research in cognitive development. *Journal of Cognition and Development*, 425-438.

Van der Kooy-Hofland, V., Van der Kooy, J., Bus, A., Van IJzendoorn, M., & Bonsel, G. (2012). Differential susceptibility to early literacy intervention in children with mild perinatal adversities: Short- and long-term effects of a randomized controlled trial. *Journal of Educational Psychology*, 337-349.

Zuckerman, M. (1999). *Vulnarability to psychopathology: a biosocial model.* Washington: American Psychological Association.

Chapter 4

Teacher opinions and beliefs on the effectivity of digital learning material

Introduction: Teacher 'beliefs and expectations' regarding digital interventions predict implementation of such interventions. We explore how these constructs might relate to digital skills of the teacher, and intervention results perceived by the teacher. *Method:* Kindergarten teachers (*N* = 106) filled in the Beliefs and Attitudes towards Digital Educational Material (BADEM) questionnaire. In a series of analyses the four identified core concepts (beliefs and expectations, implementation and involvement, digital use and skill of the teacher, and (perceived) intervention results).were interrelated. *Results:* Significant associations were found between 'implementation and involvement' and '(perceived) results', and between 'digital use and skill' and 'beliefs and expectations'. Associations between 'implementation and involvement' and 'use and skill', and between 'beliefs and expectations' and 'implementation and involvement' were marginally significant. *Conclusion:* Core concepts were strongly associated. Use and skill predicted strongest for other concepts in the model. To promote implementation of digital interventions digital skills of the teacher should be trained.

Digital educational environments offer unique opportunities to adapt the learning environment to the individual needs of students. Recent studies in kindergarten samples show for example that students vulnerable to experiencing learning problems, can thrive when the digital educational environment meets their specific needs (e.g. Van der Kooy-Hofland, Van der Kooy, Bus, Van IJzendoorn, & Bonsel, 2012, Plak, Merkelbach, Kegel, Van IJzendoorn, & Bus, 2016, Merkelbach, Plak, & Rippe, 2018). Under such circumstances, these vulnerable students (can) even outperform their non-vulnerable peers, while in regular learning environments, these children tend to lag behind. Although many questions pertaining to digital educational interventions for kindergartners still need to be addressed, as it currently stands it can be argued that deployment of digital educational programs holds great potential for a considerable number of young children whom are currently falling behind in the regular classroom environment.

In order to effectively implement digital learning material in classrooms, teachers not only need to be capable and equipped to work with digital educational programs, and able to select the right material and effectively integrate it into their curriculum, but they should also be intrinsically supportive of using these materials (Güllbahar, 2007). While teachers seem to generally appreciate the benefits of using technology in their instruction, effective integration of computer programs and other digital material, for which both high levels of involvement and an ICT encouraging school environment are vital (Schiller, 2002), has proven to be often unsuccessful (e.g. Güllbahar, 2007; Vannatta & Nancy, 2014).

This lack of effective use of digital material might be explained by challenges faced by teachers when implementing digital interventions. Such challenges, both internal and external to the teacher, could prevent teachers from (effectively) implementing digital programs in their curriculum, consequentially denying vulnerable children valuable opportunities. External challenges which teachers might encounter include for example a lack of access (e.g. sufficient availability of computers, internet access, etc.), and digital support (e.g. technological support) (Johnson, Jacovina, Russell, & Soto, 2017). However, in primary schools in countries like the Netherlands typically one computer is available for each five pupils and high speed internet access is common (Nationale Onderwijsgids, 2015). Here, it thus seems likely that internal challenges encountered by the teacher are a greater impediment for the implementation of digital learning material than external challenges. Potential internal challenges are negative beliefs and expectations concerning digital programs, lack of digital skills and limited knowledge of computers and digital educational programs, and little involvement and interest in digital programs (Johnson, et al., 2017; Hu, Clark, & Ma, 2003).

The *Technology Acceptance Model* (TAM) (Davis, 1989) describes how internal challenges may hamper effective implementation, by stating that expectations and beliefs regarding technology are of great predictive value for (effective) implementation. TAM focuses on two constructs that describe these beliefs and expectations: *perceived usefulness* and *perceived ease of use*. Perceived usefulness depends on results of an intervention as perceived by the teacher. If a teacher feels a digital intervention helps pupils to learn, perceived usefulness will be high. Perceived ease of use (from the point of view of the teacher) on the other hand, depends on the complexity of the technology, but also largely on the skill level of the individual working with the technology, in this case thus the teacher. According to TAM, in order to facilitate effective implementation of technology, scores should be relatively high on both perceived usefulness and perceived ease of use.

In line with this reasoning, we might thus expect to predict the teacher's level of implementation of, and involvement in a digital intervention via the model depicted in Figure 1. Because we focus on components internal to the teacher, we did not include complexity of the digital material as a predictor of beliefs and expectations in this model. Because in the current study all teachers worked with the same digital educational material, and thus with the same level of complexity, leaving out this component did not lead to the exclusion of valuable information. The current study aims at testing this mediation model, thereby taking a first step in identifying which (teacher) constructs could serve as an anchor for intervention in order to further promote effective use of digital material in classrooms.

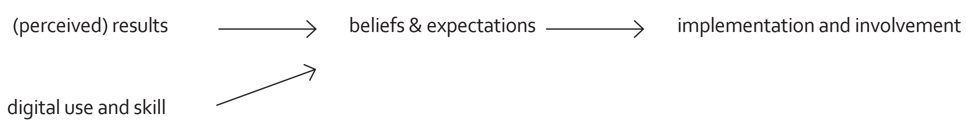


Figure 1. Expected model predicting active deployment of digital material in the classroom, based on the TAM.

Method

Design

In the school years 2013/2014 and 2014/2015 a total of 183 schools participated in the research project *What Works for Whom*, a large scale study on the (differential) effects of digital educational programs in kindergarten (e.g. Plak et al., 2016). In two consecutive research waves kindergarten teachers participating in this study implemented digital literacy- and numeracy interventions over the course of two to three months. After completing the intervention, teachers of the 139 schools participating in the second research wave (N = 106) were asked to complete a survey about their beliefs and

expectations about digital educational material and the use of such material in their curriculum, their personal computer use and digital skills, their current use of computers and digital material in the classroom, and their opinions on participation in the *What Works for Whom*-project and on the interventions used in this study. The *What Works for Whom*- project is described in more detail in Plak et al. (2016) and Merkelbach et al. (2018).

Participants

We surveyed only those teachers whom had participated in the second wave of the *What Works for Whom*-project, in which 139 schools participated. Questionnaires were send out to all participating schools, of 95 schools at least one teacher responded. From most schools, one teacher filled in the questionnaire. Of five schools two teachers answered the questions, while of three schools three teachers filled in the questionnaire. Schools were located across The Netherlands, in both rural and urban areas. The mean age of teachers was 43.90 (SD = 11.94) years. 91.5% of teachers was female, one teacher was male (.9%) and for eight teachers their gender was not reported. Part-time work was indicated by 67.0% of teachers and they thus share their teaching responsibilities with another teacher.

Measures

Questionnaire

The questionnaire on Beliefs and Attitudes towards Digital Educational Material (BADEM) was developed by the researchers and digitally distributed. Questions regard background information of teachers, beliefs and expectations towards the use of computer programs in the curriculum, computer use and skills, appreciation of participation in the *What Works for Whom*-project, and use of the computer in the classroom (both as part of, and outside of the research project). Filling out took between 10 and 15 minutes. For a complete overview of questions and respective answers, see Supplementary Table 1 (gray boxes identify the answers given most frequently by teachers). Overall, teachers displayed a moderately positive attitude towards working with digital material. It must be noted however, that this questionnaire was sent out only to teachers who had already agreed to participate in a study into the effects of a digital learning intervention, most likely resulting in some bias. We probably reached teachers with a more than average positive attitude towards digital material, while missing teachers with a critical stance. However also in most other studies, teacher opinions towards digital material turn out to be nuanced and predominantly positive (e.g. Lam, 2000, Chen, 2008). Also in this study, answers were sufficiently spread, and teachers did not exclusively feel positive towards digital programs: for example, only 19.1% of teachers did not believe that children with literacy delays would benefit more from individual teacher attention than from working with computer programs.

Data analysis

Items were grouped per construct of the mediation model based on TAM, as depicted in Figure 1: ‘use and skill’ (item 7 and item 8), ‘(perceived) results’(item 9, item 14, item 15), ‘beliefs and expectations’ (item 1 to item 6), and ‘implementation and involvement’(item 9 and item 21). Items describing the same construct were combined into one empirically weighted score by performing a principal component analysis (PCA). We chose to this approach, instead of deploying a SEM with latent constructs, because we wanted those constructs included in the analysis to match constructs described in the model depicted in Figure 1 as closely as possible, and because this approach would minimize the number of parameters included in the model. After defining these constructs, the mediation analysis depicted in Figure 1 was tested using structural equation modelling (SEM).

Results

Components

The component describing digital ‘use and skill’ explained 77.66% of variance, both items showed loadings of .88. A high score indicated a high level of computer skills and use. The component describing ‘(perceived) results’ explained 78.09% of variance, both items showed loadings of .88. A high score indicated positive (perceived) results. The component describing ‘beliefs and expectations’ of teachers explained 42.15% of variance, loadings varied from -.48 (item 1) to .83 (item 6). A high score indicated negative beliefs and expectations towards digital material. Since the other component identified so far describe positive feelings and associations, this scale was reversed, so that a higher score now indicated more positive beliefs. Lastly, the component describing ‘implementation and involvement’ explained 65.91% of variance, both items showed loadings of .81. A high score indicated a high level of implementation by and involvement of the teacher.

Mediation analysis

The overall mediation model proved significant (*F* (2, 91) = 10.52, *p* <.001), and explained 19% of variance (R^2= .19) of the ‘implementation and involvement’ construct. Not all of the expected paths proved significant. Figure 2 shows which paths could be confirmed and which could not, dashed lines indicate non-significant relations.

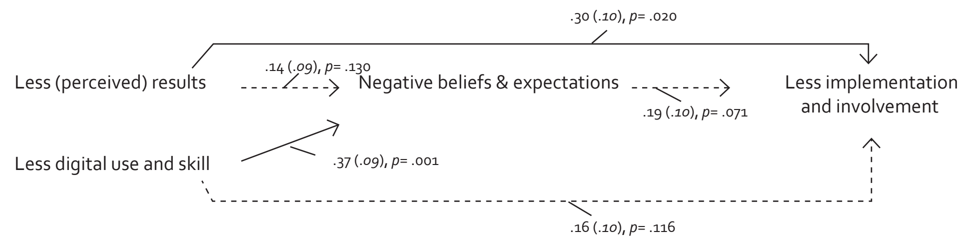


Figure 2. Results mediation model including estimates (*SE*) and *p*-values (continued lines = significant, dashed lines = not significant)

‘(Perceived) results’ was directly associated with ‘implementation and involvement’ (*p* = .020), however no association was found between ‘(perceived) results’ and ‘beliefs and expectations’ (*p* = .130). Digital ‘use and skill’ was also not associated with ‘implementation and involvement’ (*p* = .116), however a direct effect on ‘beliefs and expectations’ was found for digital ‘use and skill’ (*p* = .001). There was however no mediation, since the association between ‘beliefs and expectations’ and ‘implementation and involvement’ just failed to reach significance (*p*= .071). In conclusion, this set of associations between these four constructs, does not indicate mediation of the association between ‘(perceived) results’ and/or digital ‘use and skill’, and ‘implementation and involvement’ through ‘beliefs and expectations’ of the teacher.

These results do not align with our expectations. Surprisingly absent is an association between teacher ‘beliefs and expectations’ regarding digital material and level of ‘implementation and involvement’, a link which was however firmly established in TAM-based literature.

Secondary explorative analysis

One possible explanation for these unexpected results is that the tested mediation model is an oversimplification of true connections between the concepts in the model, because the model suggests one-directional causality. For example, it might not be reasonable to assume that ‘implementation and involvement’ is the final stage in this model, consequently not influencing the other constructs. Instead, teachers who do not make use of digital material (in their teaching; i.e. thus scoring low on implementation and involvement) could be expected to not train their digital skills sufficiently. This suggested additional association is supported by the finding that under promoted digital implementation and involvement of kindergarten teachers (by providing them with laptops and digital material to work with), teacher skill level with educational technology increases (Donovan, Green, & Hansen, 2011). Additionally, since digital material has been shown to have the possibility to stimulate a broad range of (academic) skills in kindergartners (e.g. Lieberman, Bates, & So, 2009), less (effective) implementation of such materials would evidently lead to less (perceived) results. Because 'implementation and involvement' thus can influence both digital 'use and skill', and level of '(perceived) results', we extended the initial mediation model with associations A and B, as depicted in Figure 3. Additionally, digital 'use and skill' of the teachers might not only be predictive for 'beliefs and expectations', but also for '(perceived) results' since technology will most likely have little effect if teachers are not adequately trained to use this technology or have a too low digital skill level to do so (Savage , Erten, Abrami, Hipps, Comaskey, & Van Lierop, 2010). Teachers with low levels of digital 'use and skill' will have more trouble selecting appropriate interventions and will encounter more (technical) difficulties while implementing interventions, which in turn could lead to lower '(perceived) results. To account for this, association C (as depicted in Figure 3) was added to the model.

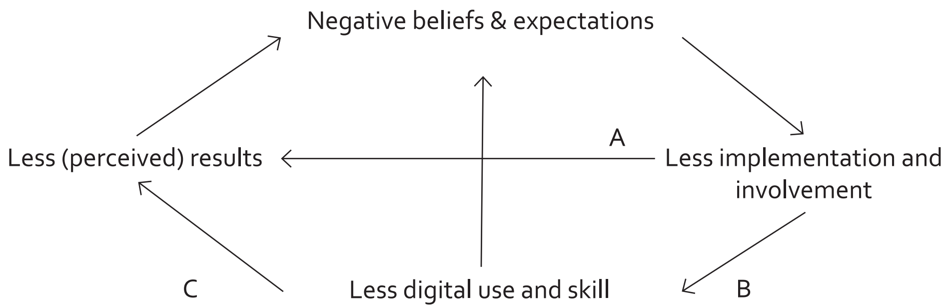


Figure 3. Proposed extended model of teacher influences on use of digital material in the classroom.

Results of this analysis are depicted in Figure 4. Connections between '(perceived) results' and 'digital use and skill' (p = .543) and between '(perceived) results' and 'beliefs and expectations' (p = .880) did not reach significance. However the association between 'implementation and involvement' and '(perceived) results' reached significance (p <.001), as did the association between digital 'use and skill' and 'beliefs and expectations' (p <.001). The associations between 'implementation and involvement' and 'use and skill' (p = .068), and between 'beliefs and expectations' and 'implementation and involvement' (p = .054) were marginally significant.

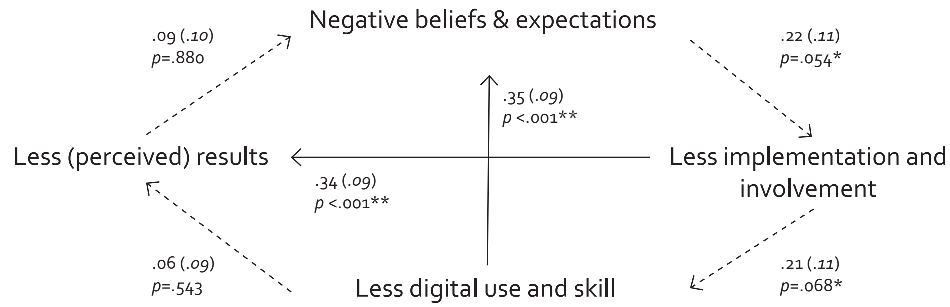


Figure 4. Results of the expanded model circular model (solid lines = significant, dashed lines = marginally significant, grey dashed lines = not significant)

As was seen in the original mediation results, 'digital use and skill' is a significant predictor of 'beliefs and expectations', while '(perceived) results' is not predictive. Also, as was the case in the mediation model, the expected association between 'beliefs and expectations' and 'implementation and involvement' was not convincingly detected, however the association was now marginally significant (p= .054). The association between 'implementation and involvement' and '(perceived) results' was significant, as it was in the original model, however in the current model the direction is converted. The added association between 'implementation and involvement' and 'digital use and skill' reached a marginal significance level (p = .068).

To assess sensitivity and stability of the just-described results, we repeated the analysis under exclusion of the one male teacher, since he might be considered conceptually different from the other respondents. Results (depicted in Supplementary Figure 1) were highly comparable to those presented in Figure 4.

Discussion

Digital educational material can offer children unique learning opportunities (e.g. Van der Kooy-Hofland et al., 2012, Plak et al., 2016, Merkelbach et al., 2018). However, both external and internal challenges can prevent teachers from effectively implementing digital interventions in the classroom (Johnson et al., 2017). In this study, we focused on internal teacher challenges, assuming that influence of internal challenges outweighs that of external challenges (such as availability of digital material and hardware) in a modern western society. Based on the TAM (Davis, 1989) we expected 'beliefs and expectations' of the teacher about the utility of digital material (i.e. perceived ease of use and usefulness of such material) to mediate the relation between teacher's digital 'use and skill' of and 'implementation of and involvement' in such digital programs, as well as the relation between '(perceived) results' of digital intervention by the teacher and 'implementation and involvement' (Figure 1). We were however not able to confirm the proposed mediation (Figure 2). While some of the associations reached significance, most associations could not be confirmed.

Additional, exploratory analyses were run, in which we tested a circular model of the implementation of digital material by the teacher (Figure 3). Here, too, not all proposed associations could be confirmed. However, four (66.67%) of the proposed associations proved (marginally) significant, all depicted in Figure 5.

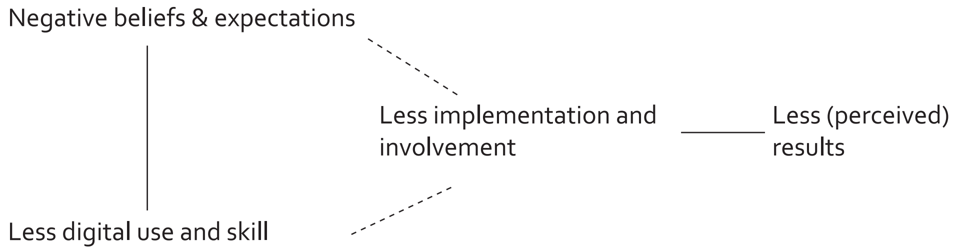


Figure 5. *Confirmed associations of the circular model of the implementation of digital educational by the teacher (solid lines = significant, dashed lines = marginally significant)*

This model suggests that 'implementation and involvement' are key ingredients for achieving and perceiving intervention results. Additionally, whether or not digital material will be implemented can be predicted by both the 'beliefs and expectations' of the teacher towards digital material and the level of digital 'use and skill' of the teacher, while digital 'use and skill' of the teacher also predicts for 'beliefs and expectations'. The level of digital 'use and skill' of the teacher thus seem to be the base of the current model. Additionally, recent case studies show that promoting digital 'use and skill' of kindergarten teachers by offering ICT training can positively influence teachers' perceptions and practices, as well as reduce obstacles teachers encounter while implementing digital material (Ihmeideh & Al-Maadadi, 2018). Therefore digital 'use and skill' is expected to be the aspect that could best be intervened on in order to promote the integration of digital material in standard curricula. Developers of digital material should thus not only focus on developing effective digital interventions and programs, they should also secure ease of use of such material for the teacher, and offer clear instructions on how to use and implement digital material. Additionally teacher training should include courses in which using digital material in the classroom is discussed and trained in a professional manner (Instefjord & Munthe, 2017).

In the current model no causal relations can be revealed and we cannot be sure of the direction of associations. Additionally, concepts are broadly defined, future research should

focus on carrying out dedicated experiments to establish a predictive association between promoting digital skills of teachers and a) changes in their 'beliefs and expectations', b) in their level of 'implementation and involvement', and eventually and most importantly c) in achieved results. Additionally, further specification of teacher training requirements is needed. For example, should teachers be trained to use specific digital interventions, should we focus on improving the general use of digital technology, or are both concepts equally important? By both broadening and specifying our understanding of how internal characteristics of the teacher might influence the implementation of digital interventions and thereby eventually the results of such programs, we can more precisely identify which steps could be taken in order to stimulate the use of digital material in the classroom.

In summary, digital interventions have the possibility to help children learn, but internal challenges of the teacher might prevent the effective implementation of such digital programs as standard part of the curriculum. The current study reveals that teacher 'implementation of and involvement' in digital material might form the link between the relation between teacher digital 'use and skill', and the teacher's '(perceived) results' of digital learning material. Additionally, the association between teacher 'beliefs and expectations' regarding digital material and '(perceived) results' is also connected by 'implementation and involvement' of the teacher. Lastly, 'beliefs and expectations' can be predicted from the level of digital 'use and skill'. These findings suggest that improving digital dexterity of teachers might be the most effective way to promote the implementation of digital material in classrooms, eventually leading to better academic results. The current study is however explorative and non-experimental, therefore causality and direction of associations between core concepts cannot be established. More research, for example RCT's which intervene on digital skills of the teacher or by promoting 'beliefs and expectations', are needed to specify current findings and to establish if found relations are causal.

References

Chen, C.-H. (2008). Why do teachers not practice what they believe regarding technology integration? *The Journal of Educational Research*, 65-75.

Davis, F. (1989). Perceived usefulness, perceives ease of use, and use acceptance of information technology. *Management Information Systems Quarterly*, 319-340.

Donovan, L., Green, T., & Hansen, L. (2011). One-to-one laptop teacher education: does involvement affect candidate technology skills and dispositions. *Journal of Research on Technology in Education*, 121-139.

Güllbahar, Y. (2007). Technology planning: A roadmap to succesful technology integration in schools. *Computers & Education*, 943-956.

Hu, P.-H., Clark, T., & Ma, W. (2003). Examining technology acceptance by school teachers: A longitudinal study. *Information & Management*, 227-241.

Ihmeideh, F., & Al-Maadadi, F. (2018). Towards improving kindergarten teachers' practices regarding the intergration of ICT into early years settings. *The Asia-Pacific Education Researcher*, 65-78.

Instefjord, E., & Munthe, E. (2017). Educating digitally competent teachers: A study of integration of professional digital competence in teacher education. *Teaching and Teacher Education*, 37-45.

Johnson, A., Jacovina, M., Russell, D., & Soto, C. (2017). Challenges and solutions when using technologies in the classroom. In S. A. Crossley, & D. S. McNamara, *Adaptive Educational Technologies for Literacy Instruction* (pp. 13-30). New York: Routledge.

Lam, Y. (2000). Technophilia vs. technophobia: a preliminary look at why second-language teachers do or do not use technology in their classrooms. *The Canadian Modern Language Review*, 389-420.

Lieberman, D., Bates, C., & So, J. (2009). Young children's learning with digital media. *Computers in the Schools*, 271-283.

Merkelbach, I., Plak, R., & Rippe, R. (2018). Reproducibility of young learners' susceptibility to the learning context. *Learning and Individual Differences,* 167-175.

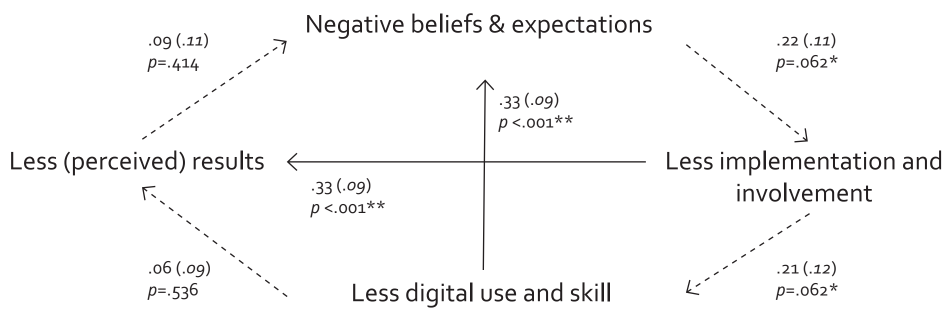
Plak, R., Merkelbach, I., Kegel, C., Van IJzendoorn, M., & Bus, A. (2016). Brief computer interventions enhance emergent academic skills in susceptible children: A gene-by-environment experiment. *Learning and Instruction*, 1-8.

Savage , R., Erten, O., Abrami, P., Hipps, G., Comaskey, E., & Van Lierop, D. (2010). ABRACADABRA in the hands of teachers: the effectiveness of a web-based literacy intervention in grade 1 language arts programs. *Computers & Education*, 911-922.

Schiller, J. (2002). Interventions by school leaders in effective implementation of information and communications technology: perceptions of Australian principles. *Journal of Information Technology for Teacher Education*, 289-302.

Van der Kooy-Hofland, V., Van der Kooy, J., Bus, A., Van IJzendoorn, M., & Bonsel, G. (2012). Differential susceptibility to early literacy intervention in children with mild perinatal adversities: Short- and long-term effects of a randomized controlled trial. *Journal of Educational Psychology*, 337-349.

Vannatta, R., & Nancy, F. (2014). Teacher dispositions as predictors of classroom technology use. *Journal of Research on Technology in Education*, 253-271.



Supplementary Figure 1. Results expanded circular model tested in a sample without the male teacher

Supplementary Table 1. *Kindergarten teacher survey on computer use*

Teacher report on use of digital educational programs

Teacher opinions and beliefs regarding the use of digital material in instruction

	Totally disagree	**Disagree**	**Neutral**	**Agree**	**Totally agree**
1.Computer programs can help children to catch up literacy delays	n = 0 (0%)	n = 0 (0%)	n = 20 (21.3%)	n = 60 (63.8%)	n = 14 (14.9%)
2.Children with literacy delays benefit more from individual teacher attention than from computer programs	n = 0 (0%)	n = 18 (19.1%)	n = 43 (43.6%)	n = 31 (33.0%)	n = 4 (4.3%)
3.Computers offer a valuable contribution to language education	n = 0 (0%)	n = 1 (1.0%)	n = 5 (4.8%)	n = 51 (48.6%)	n = 48 (45.7%)
4.Working with computer programs takes the teacher a lot of effort	n = 15 (14.4%)	n = 61 (58.7%)	n = 22 (21.2%)	n = 5 (4.8%)	n = 5 (4.8%)
5.Working with computer programs takes the teacher a lot of time	n = 11 (10.6%)	n = 47 (45.2%)	n = 29 (27.9%)	n = 16 (15.4%)	n = 1 (1.0%)
6.The time/effort it takes the teacher to work with computer programs is better spend on other activities	n = 21 (20.2%)	n = 62 (59.6%)	n = 19 (18.3%)	n = 2 (1.9%)	n = 0 (0%)

Computer use and -skills

	Mean (*SD*)	**Range**			
7.How do you judge your computer skills on a scale from zero to ten? Zero meaning you have no computer skills, ten meaning you have excellent computer skills	7.40 (*1.21*)	2 - 10			
	Never	Sometimes	Regularly	Frequently	Always
8.I often make private use of computers	n = 0 (0%)	n = 4 (3.8%)	n = 5 (4.7%)	n = 30 (28.6%)	n = 66 (62.9%)
9.Do you - outside of the What Works for Whom project - use computer programs in your instructions?	n = 8 (8.5%)	n = 6 (6.4%)	n = 27 (28.7%)	n = 29 (30.9%)	n = 24 (25.5%)

Opinions on participation in What Works for Whom

	Totally disagree	**Disagree**	**Neutral**	**Agree**	**Totally agree**
10.The instructions provided by the researchers during What Works for Whom were satisfactory	n = 1 (1.0%)	n = 18 (17.8%)	n = 28 (27.7%)	n = 52 (51.5%)	n = 2 (2%)

Continued Supplementary Table 1. *Kindergarten teacher survey on computer use*

11.The logging in procedure during What Works for Whom was satisfactory	n = 8 (7.9%)	n = 24 (23.8%)	n = 25 (24.8%)	n = 42 (41.6%)	n = 2 (2%)
12.The accessibility of the programs used during What Works for Whom was satisfactory	n = 7 (6.9%)	n = 18 (17.8%)	n = 28 (27.7%)	n = 46 (45.5%)	n = 2 (2%)
13.Participating in What Works for Whom took little time	n = 6 (6.4%)	n = 26 (27.7%)	n = 15 (16%)	n = 39 (41.5%)	n = 8 (8.5%)
14.Children benefitted from participating in What Works for Whom	n = 0 (0%)	n = 3 (3.2%)	n = 40 (42.6%)	n = 45 (47.9%)	n = 6 (6.4%)
15.What Works for Whom offers valuable support for delayed children	n = 0 (0%)	n = 2 (2.1%)	n = 41 (43.6%)	n = 45 (47.9%)	n = 6 (6.4%)
16.What Works for Whom offers a good addition to the regular curriculum	n = 0 (0%)	n = 6 (6.4%)	n = 37 (39.4%)	n = 43 (45.7%)	n = 8 (8.5%)
	No	**Maybe**	**Yes**		
17.Would you have participated in What Works for Whom if you would have known what the research project encompasses in advance?	n = 12 (12.8%)	n = 28 (29.8%)	n = 54 (57.4%)		

Digital classroom practices (in- and outside the research period)

	No	**Yes**			
18.Did children work with the computer inside de classroom?	n = 22 (21.2%)	n = 82 (78.8%)			
19.Did children wear headphones while working with the computer programs ?	n = 7 (6.7%)	n = 97 (93.2%)			
20.Did you sit beside the child the first time he/she worked with What Works for Whom?	n = 31 (30.7%)	n = 70 (69.3%)			
	Totally disagree	**Disagree**	**Neutral**	**Agree**	**Totally agree**
21.I have a clear image of what the programs of What Works for Whom encompass	n = 6 (6.4%)	n = 19 (20.2%)	n = 23 (24.5%)	n = 42 (44.7%)	n = 4 (4.3%)
22.I provided children with help while interacting with the programs	n = 25 (24.8%)	n = 50 (49.5%)	n = 22 (21.8%)	n = 3 (3.0%)	n = 1 (1.0%)

Chapter 5

Differential efficacy to a digital numeracy program in kindergarten

Introduction: Children born late preterm have been shown to be highly susceptible to the quality of the educational environment. Because numeracy and math problems are most firmly established in this group of children, this study tests the effects of a digital early numeracy intervention on delayed kindergartners, especially in late preterm children. *Methods:* In a large random controlled trial, preschool pupils (*N*=375, mean age = 67 months, *SD* = 4.50) were assigned to either a digital, guiding and structured numeracy program or a digital control program. Children worked in a classroom environment for a period of approximately two to three months. *Results:* The group as a whole did not benefit from the intervention, but the program benefited a late preterm subsample (n=40, (*Cohen's d* = .71, *CI* = .07 / 1.36). While these pupils fell behind when working with the control program. *Conclusion*: Digital numeracy interventions can support the early numeracy skills of late preterm children. It might be that highly structured and guiding programs sooth this group of children, whom are generally prone to experience high levels of stress reactivity.

In early childhood mathematical abilities develop long before the formal education of such skills starts. Infants can already discriminate between different small numbers of items and can determine numerical equivalence across perceptual modalities (Wynn, 1992). This sense of numbers keeps developing throughout childhood (Griffin, 2004), and before kindergarten starts most children have learned how to count (Wynn, 1990). Such quantity related skills are predictive for later math performance, as is also the case for other early skills, such as logical thinking (Bryant & Nunes, 2002) and visuo(spatial) skills such as recognizing, comparing, and classifying items by shape and size (Aunio & Niemivirta, 2010). This collection of precursors for later math performance is generally referred to as *early numeracy*. Delays in the development of these early numeracy skills can sort long lasting negative effects on the development of mathematical abilities (Desoete, Ceulemans, De Weerdt, & Pieters, 2010). Research shows however that mathematical performance is particularly susceptible to the effects of interventions (e.g. Gervasoni, 2001), especially when implemented at an early age. If we could identify vulnerable children in time, early interventions could thus possibly prevent long term delays.

One such group that has been found particularly vulnerable for experiencing problems with numeracy and mathematics is the group of children born preterm: although these children generally experience more problems in all cognitive domains (e.g. Chyi, Lee, Hintz, Gould, & Sutcliffe, 2008, Woythaler, McCormick, Mao, & Smith, 2015), problems in the field of numeracy and mathematics are most pronounced (e.g. Poulsen, et al., 2013), resulting in adverse outcomes persisting far into adulthood (Basten, Jaekel, Johnson, Gilmore, & Wolke, 2015). Cognitive problems are not only consistently associated with very preterm birth (born < 32 weeks of pregnancy), but also with late preterm birth (between the 34^{th} and 38^{th} week of pregnancy). Even though late preterm birth is 'only' considered a mild perinatal adversity (Van der Kooy-Hofland, Van der Kooy, Bus, Van IJzendoorn, & Bonsel, 2012), these children thus still consistently show higher levels of cognitive problems (Shah, Kaciroti, Richards, Oh, & Lumeng, 2016, Searle, Smithers, Chittleborough, Gregory, & Lynch, 2017) and for example lower IQ-scores (De Jong, Verhoeven, & Van Baar, 2012) than their peers.

Similar outcomes are found in a another group subject to mild perinatal adversity, those born small for gestational age (below the 10^{th} percentile). In childhood and adolosence this group, too, is at risk for experiencing a range of cognitive problems (e.g. Sommerfelt, et al., 2000, Paz, Gale, Laor, Danon, Stevenson, & Seidman, 1995), among which more frequent as well as higher levels of learning disabilities (O'Keeffe, O'Callaghan, Williams, Najman, & Bor, 2003) and poorer school performance (Larroque, Bertrais, Czernichow, & Léger, 2001). For those born small for gestational age, negative associations between mild perinatal adversity and cognitive outcomes persist into adulthood: people who were small for gestational age at birth tend to show lower levels of academic achievement and professional attainment (e.g. Strauss, 2000; Larroque et al., 2001).

Although these mild perinatal adversities are associated with increased chances of negative educational and cognitive outcomes, considering mild perinatal adversities as just a vulnarability factor might be short-sighted. Research shows that people who have experienced mild perinatal adversities might be more susceptible to qualities of the environment, for better and for worse as is described in the *differential susceptibility model* (Belsky & Pluess, 2009). This pattern has been found in rearing settings, in which children with mild perinatal adversities have been found to be more susceptible to the influence of maternal harsh parenting (Windhorst, et al., 2017), but also in educational settings. A small scale experiment shows that kindergartners who have experienced mild perinatal adversiteis are more susceptible to a digital early literacy intervention: *Living Letters*, a program training phonological awareness and alphabetical knowledge (Van der Kooy-Hofland et al., 2012). In this experiment, children with literacy delays were exposed to either *Living Letters* or a control program, with which they worked once a week for ten minutes, over the course of two to three months. For children without mild perinatal adversities *Living Letters* had no effect, neither negative nor positive. While on the other hand, children with mild perinatal adversities fell behind even further when they had worked with the control program, but outperformed their peers when they had worked with *Living Letters.*

In this small-scale study (Van der Kooy-Hofland et al., 2012) children with mild perinatal adversities were treated as a homogenous group,supposedly responding similarly to the qualities of an intervention like *Living Letters.* A large scale replication study showed however that this might not be completely correct (Merkelbach, Plak & Rippe, 2018). In this study, as in the Van der Kooy-Hofland et al. study (2012), children worked with either *Living Letters* or a control program over the course of about two to three months. Because the sample size was large, a distinction could be made between children born late preterm and children who were small for gestational age at birth. As children without mild perinatal adversities, children who were small for gestational age were found not to be susceptible to the qualities of *Living Letters*. However, for late preterm children a differential susceptibility pattern was found, as in the Van der Kooy-Hofland et al. (2012) study, these children fell behind when assigned to the control condition, but outperformed their peers when assigned to *Living Letters.*

Main features of *Living Letters* are structure, repetition of assignments, continous and adaptive feedback, and guidance (Merkelbach et al., 2018), which seem to meet the needs of children born late preterm particularly well. In contrast, they might not fully fit the needs of those born small for gestational age. Because numeracy and mathetical problems are more prevalent in children born late preterm than in their full term peers (Nepomnyaschy, Hegyi, Ostfeld, & Reichman, 2012), and such problems are generally fairly susceptible to the effects of interventions (e.g. Gervasoni, 2001), especially when carried out at an early age, we wondered if late preterm kindergartners would benefit from a program similar in design to *Living Letters*, but targeting early numeracy instead of early literacy skills. In addition, we wondered if children born small for gestational age would respond differently to this intervention, as is suggested in Merkelbach et al. (2018).

We therefore tested the results of the digital early numeracy program *Clever Together* in children born late preterm, children small for gestational age, and children without mild perinatal adversities. Like *Living Letters*, *Clever Together* consists out of short games which train several early numeracy skills and which are repeated several times. In addition, *Clever Together* (just as *Living Letters*) also includes digital tutors that offer the child continuous and adaptive feedback, and high levels of guidance and explanation. *Clever Together* highly resembles *Living Letters* in terms of substantive features, as well as in design (e.g. the same digital tutors) and duration and dosage (ten minutes, once a week for two to three months).

Building on earlier findings, we thus expected children born late preterm to benefit from working with *Clever Together* while no results would be expected for children born small for gestational age and children without mild perinatal adversities. To test this hypothesis, we used standardized numeracy tests, that are administered as a standard part of the student tracking system used in the vast majority of Dutch kindergarten and primary schools (Cito, 2017). Pupils were randomly assigned to *Clever Together* and to the control program consisting of digital animated storybooks, to compare their performance.

We hypothesize that:

1. *Clever Together* stimulates early numeracy skills in kindergarten children lagging behind in such skills, however, effects are only present in vulnerable children;
2. Children who are born late preterm will benefit from working with *Clever Together;*
3. Children without mild perinatal adversities and children small for gestational age will not benefit from working with the program.

Method

Design

In the current study, we tested the benefits of a numeracy program, *Clever Together*, in a large group in which delays in numeracy skills were common(N = 375). In a large-scale experiment consisting of two waves (2012/2013 and 2013/2014), delayed kindergarten children were randomly assigned to the control condition, a book-reading program (*Living Books*), or the experimental condition, *Clever Together.* Because children worked with the programs on an individual basis, children from the same classrooms could be assigned randomly to control and treatment conditions. Given the large sample (N = 375) and the stability of the percentage of perinatal adversities, it was reasonable to assume that random assignment by the researcher would result in a comparable number of children born late preterm between conditions (Experimental = 10.1%, Control = 11.2%, Table 1). Teachers could not influence the assignment procedure since they merely logged children in. From that moment on children worked individually with their assigned program. Pre-testing and post-testing of early numeracy skills took place as part of the regular monitoring system applied in Dutch kindergarten classrooms, with a standardized numeracy test administered group wise by the teacher, blind for the hypotheses of the study, in January/February of the second kindergarten year and in May/June, just preceding first grade of primary school. Testing in January/February preceded the intervention while the test in May/June was administered directly after the intervention.

Participants

Kindergarten classrooms of 140 elementary schools, both urban and rural, located across the Netherlands, were included in the trial. Kindergarten teachers selected children from their classroom who were delayed in literacy skills for participation; children had to be five or six years old (Mean age: 67.12 months; SD = 4.50). After receiving informed consent from a parent, children were randomly assigned to one of the computer programs (in the overarching research three different programs were used). The target group (N = 879) in the current trial was randomly assigned to either *Clever Together* or the control program. Children were excluded due to missing data on the numeracy pretest (n =50) or numeracy posttest (n = 88). Children were also excluded when there was no consent from the mother to retrieve perinatal information from the national perinatal registry (n = 268) or if consent was given but the information provided by parents (home address and date of birth of the mother) was incomplete and we were therefore unable to retrieve perinatal information from the registry (n = 96). Lastly, children were excluded when information about the educational level of the father (n = 4) was lacking. The final sample therefore consisted of 375 children (Figure 2).

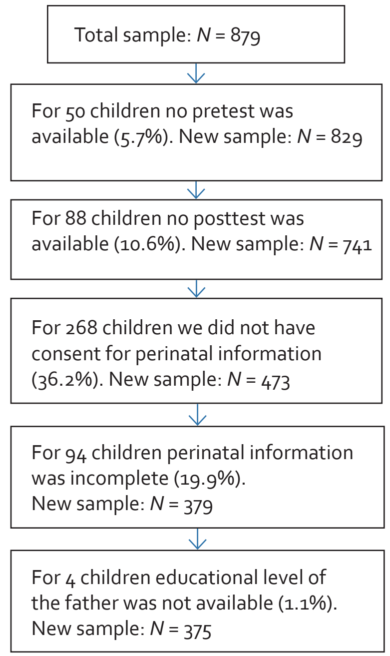


Figure 2. Flowchart of final participant inclusion

Procedure

About 1750 randomly selected schools throughout the Netherlands had received information by mail through brochures and on social media about educational computer programs that might provide extra guidance to kindergarten children who were delayed in academic skills. As a result, 140 schools signed up to participate. Schools were offered three months of free access to educational computer programs that normally require a paid subscription (http://www.bereslim.nl), after completion of the intervention.

Parents provided informed written consent and their email address. Thereafter, parents received a link to a website with frequently asked questions about the project. If they had any further questions, they could also contact the researchers personally (via phone or email). In the first wave, parents' consent for retrieving perinatal information was not a condition for participating in the study. In this wave, parents were asked to provide consent for retrieving perinatal information after the intervention was already completed, because hypotheses regarding differential susceptibility to the educational environment in children with mild perinatal adversities were formulated shortly after collecting general consent. This procedure might explain the high rate of missing perinatal data, which was mainly due to non-consent (67.7%), in that year. In the second year, in reaction to the high rates of missing data, consent for perinatal information was included as a condition for participation in the study and thus for receiving the intervention. This resulted in a much lower rate of missing data (31.7%), largely due to a normal proportion (20%) of matching errors between the registry and our research database.

After assignment to one of the intervention programs, children received access to one of the programs once a week during two to three months, as this was also the dosage used in previous studies (e.g. Van der Kooy-Hofland et al., 2012; Plak, Merkelbach, Kegel, Van IJzendoorn, & Bus, 2016). Children played the games in a classroom setting only receiving adult assistance for logging in. They wore headphones to prevent that the program would attract and distract other children.

Intervention programs

CleverTogether. The first 45 tasks (Figure 3) were hide-and-seek games to practice with orienting and locating (e.g. locating objects during hide-and-seek games), as well as with prepositions (e.g. 'in', 'behind'). Sim, one of the characters in the game, asked for help in finding Sanne who is hiding behind one of the objects in the illustration (e.g. "*I am going to hide behind the blue tree*"). In the other 30 games (Figure 3), children had to assemble objects (e.g. a bike) from their parts (e.g. tires, frame, steering wheel), select attributes for an activity (e.g. taking a shower), thereby practicing with spatial prepositions (e.g., 'in', 'behind').

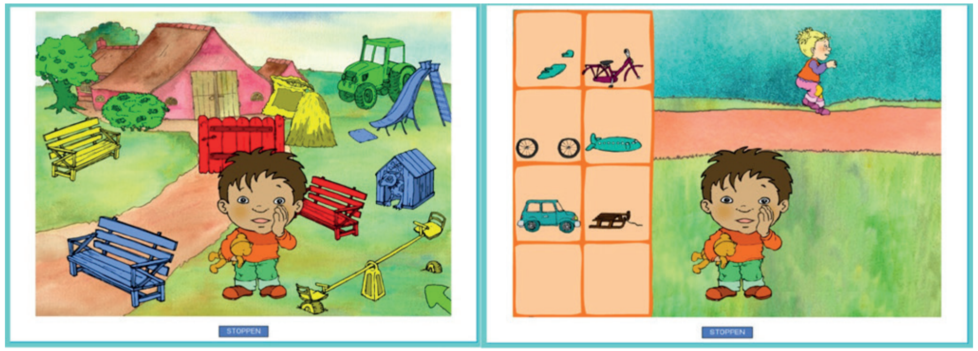


Figure 3. *CleverTogether* games: Find Sanne who is hiding behind one of the objects (left) or assemble an object from different parts (right)

In the program, a teddy bear provided responsive replies in a supportive tone. In case of errors a hierarchical set of replies dependent on the child's response was provided (Figure 4). Moreover, assignments that were not answered correctly at a first try were repeated in later sessions followed by similar adaptive feedback to create more opportunities for practicing difficult assignments.

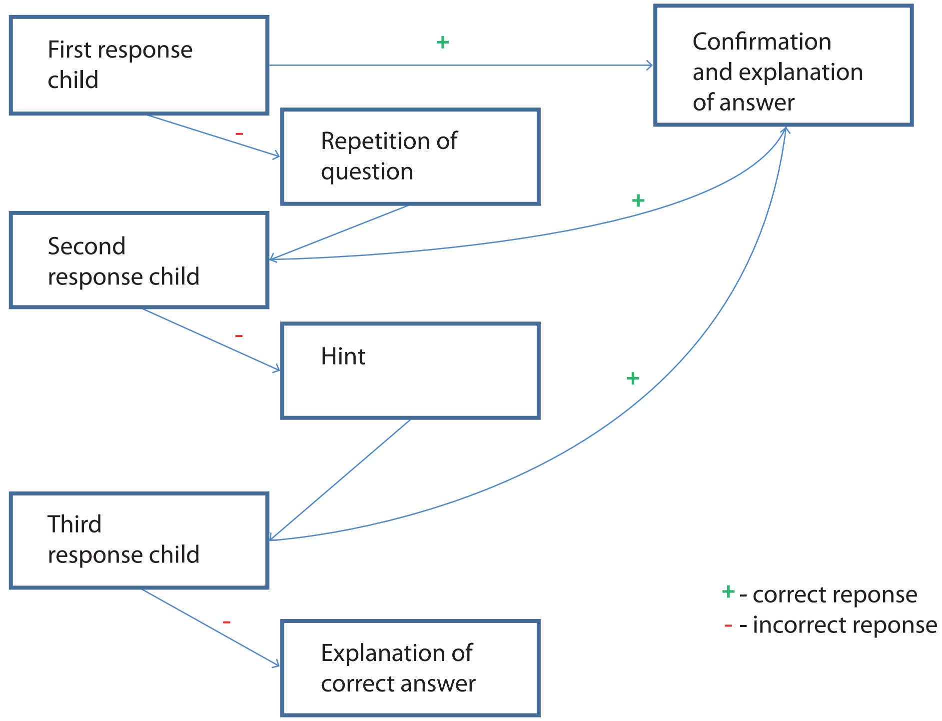


Figure 4. Feedback circle *CleverTogether*

Control condition. The control program consisted out of eight digital, age-appropriate, multimedia storybooks with oral text, each read twice. The story text matched the nonverbal, film-like information including animated pictures, music, and sounds. Each story was interrupted four times by digital tutors for a question about difficult words that appeared in the text or about story events, followed by a similar set of hierarchical replies as is offered in *CleverTogether*. However, in this program the questions and answers only occupied a small part of the session, about 10% of the total duration.

Measures

Background variables

Children's age (in months) and sex, and the educational level of the father were assessed. Following the rationale of Van der Kooy et al. (2012) that educational level of the father is more strongly associated with mild perinatal adversities than educational level of the mother (as was the case in the current study), we made use of father's educational level instead of that of the mother. The sex and the date of birth of the child were reported by the teacher of the child. The educational level of the father was reported by the parent(s) on a 7-point scale (ranging from no education to university or higher).

Cito Numeracy Skills

The Cito Numeracy Test for Kindergarten Pupils (CNT) is a group-administered standardized numeracy test for kindergarten children administered by the teachers in January/February and May/June in the senior year of kindergarten[1] when children were five to six years of age (Koerhuis & Keuning, 2011). The test consists of 48 questions that focus on number sense (e.g. '*Where do you see three rabbits?*'), classification (e.g. '*Which dog is the biggest?*'), and geometry (e.g. '*Which shadow matches this picture?*'). Based on normative scores, the pretest score of the CNT January/February was dichotomized and coded into scoring at the 40^{th} percentile or lower (0, score lower than 78) or scoring average and above (1, score of 78 or higher). As the posttest, the full range of scores on the CNT May/June was used. Versions of the CNT administered in January/February and the CNT administered in May/June were similar in content and design but included different items.

Perinatal Data

The Netherlands Perinatal Registry (Centraal Bureau voor de Statistiek, 2013) combines data about duration of pregnancy and weight of the child at birth from three registries: the national obstetric database by midwives, the national obstetric database by gynecologists, and the national neonatal/pediatric database (Méray, Reitsma, Ravelli, & Bonsel, 2007) and covers about 96% of all pregnancies in the Netherlands.

Duration of pregnancy was dichotomized into being born full term (0) or being born late preterm (1) which was defined as a gestational age at birth of 34 weeks – 37 weeks + 6 days. Small for gestational age was dichotomized into 'not small for gestational age at birth' (0) and 'small for gestational age at birth' (1), which was defined as lower than the 10^{th} percentile of birth weight for gestational age, taking into account gender and parity.

Statistical analyses

Dependent measures were regressed on the intervention, late preterm birth and small for gestational age, and the interactions between late preterm birth and intervention, and small for gestational age and intervention. For both susceptibility markers a dummy variable was created. Children could thus be in both groups, as was the case for two children. Using the likelihood-ratio test, we tested whether the model fit would improve when we allowed intercepts and slopes to vary across schools (multilevel approach).

[1] In the Netherlands children attend two years of kindergarten. In both years the emphasis is on play instead of formal education.

Because the level of missing data was high (57.3%) we followed Little's (1986) MCAR procedure to see if complete cases analysis would be allowed for. This would be so only when data were missing completely at random (MCAR). In addition, other procedures were applied to account for missing data and models were compared. Models were estimated with a full information maximum likelihood (ML) approach and a multiple imputation (MI) approach. Using ML, models were estimated on the basis of all information available, both from complete and incomplete cases (n=879), yielding the most likely association parameters. Using MI, missing values were imputed (m=100) via chained equations. Applying predictive mean matching, linear and logistic regression prediction where appropriate. The imputation scheme includes all model variables, interactions as well as exogenous variables. Estimates of parameters and standard errors were pooled over imputed datasets, yielding very precise parameter estimates, however potentially showing slightly increased standard errors to account for multiple estimation of missing information. To assess robustness of results, estimates and standard errors were compared between the three approaches (complete case, ML, and MI). Similarity of estimates indicates robustness, while considerable differences signal that results derived from complete case analysis might be strongly biased.

Results

Missing data

Based on Little's MCAR test (1986), we could reject the null hypothesis that data were not missing completely at random (χ^2 = .08, p = .777); therefore it also made therefore sense to apply complete case analysis including only individuals with complete data.

Sample characteristics

Sample characteristics are presented in Table 1. A small majority of children was male (54.9%), in accordance with the general finding that more boys than girls are delayed in the early years of schooling (Gurian, 2010). In this sample of 375 children, 40 children (10.7%) were born late preterm and 94 children were born small for gestational age (25.1%). Sample characteristics did not differ between groups.

Table 1: *Percentages or means (standard deviations) for the complete group of children with complete cases and for the experimental (Clever Together) and control conditions (Living Books) separately (N = 375); p-values for Chi-square or t-test.*

	Complete group (*n* =375)	Experimental group (*n* =179)	Control group (*n* =196)	*p*
Male	54.9%	55.3%	54.6%	.889
Age (in months)	67.12 (*4.50*)	67.58 (*4.64*)	66.70 (*4.33*)	.060
Father's education (max = 6)	3.74 (*1.50*)	3.72 (*1.50*)	3.77 (*1.51*)	.774
Mild perinatal adversities	32.5%	35.2%	30.1%	.293
Late preterm	10.7%	10.1%	11.2%	.714
Small for gestational age	25.1%	27.4%	23.0%	.324
CNT pretest (max = 137)	78.67 (*10.30*)	79.54 (*11.44*)	77.88 (*9.09*)	.119
Percentage delayed children[1]	45.3%	40.8%	49.5%	.091

1.Below 40th percentile on CNT pretest math

Effects of Clever Together

Table 2 shows the mean standardized outcome scores on the standardized numeracy test for late versus full term children and for the group as a whole.

Table 2. *Means and Standard Deviations for standardized numeracy post-tests by condition and mild perinatal adversities*

	CNT posttest (standardized)			
	Clever Together	*n*	*Living Books*	*n*
Full term	.04 (*1.05*)	161	-.04 (*.93*)	174
Late Preterm	.43 (*1.26*)	18	-.33 (*.88*)	22
Not SGA[1]	.05 (*.97*)	130	-.11 (*.87*)	151
SGA	.16 (*1.30*)	49	.04 (*1.11*)	45
Total	.08 (*1.07*)	179	-.07 (*.93*)	196

1. SGA = small for gestational age

Effects of Clever Together

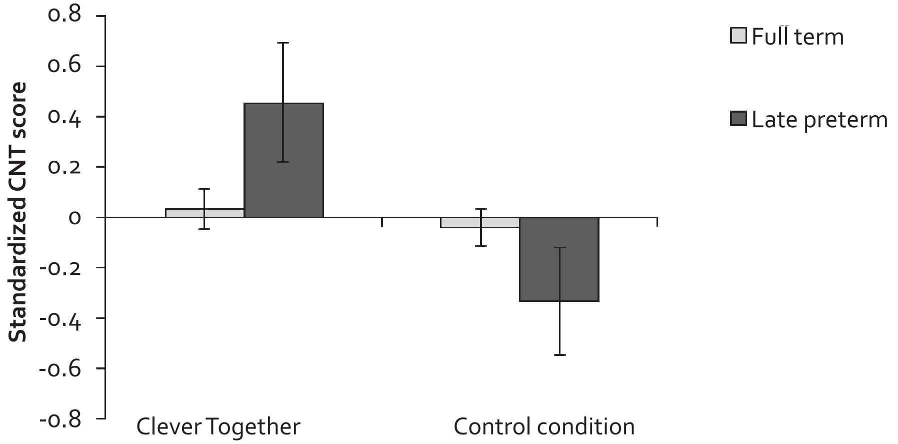
The CNT (June) was regressed on dichotomized numeracy pretest score, late preterm versus full term, small for gestational age vs. normal for gestational age, and the two two-way interaction: late preterm X condition, and small for gestational age X condition. First, we tested whether or not it was necessary to allow the intercept and slope to differ between schools in the regression model (Bickel, 2007). The difference between the -2log likelihood of the model with a random intercept and the -2log likelihood of the model without a random intercept equaled .94. Following a chi-square distribution with one degree of freedom, this difference was not significant (p > .10). Likewise, the difference between the -2log likelihood of the model with a random intercept and slope and the -2log likelihood of the model with only a random intercept (.00) was not significant (p > .10). This indicates that variability in scores on the numeracy test administered after the intervention was similar across schools, therefore a non-hierarchical ordinary least squares (OLS) was applied.

Results of the OLS are presented in Table 3. The CNT pretest (F (1, 373) = 171.78, p <.001) showed a main effect, children with an average or above score on the pretest scored higher (*Mean* = 93.22 (*11.03*)) than children with a below average score on the pretest (*Mean* = 79.69 (*8.54*)). No main effects were found for late preterm birth (F (1, 373) = 2.55, p = .111) and small for gestational age (F (1, 373) = .31, p = .577). There was no significant interaction between small for gestational age X condition (F (1, 373) = .38, p = .537), however the interaction, born late preterm X condition was significant (F (1, 373) = 5.63, p=.018). Children born late preterm scored higher on the posttest than their peers when working with *Clever Together* but lagged further behind with *Living Books*, the control condition (see Figure 5). Four CLT scores were outliers (more than three *SD*s above the sample mean).

Repetition of the analysis using MI and ML yielded highly similar results and thus similar substantive conclusions indicating that results derived from complete case analysis were not biased. Estimates and standard errors were highly comparable across all parameters (Supplementary Table 1), including the interaction between late preterm and condition. Estimates for complete cases were: 7.88 (*3.32*); for MI: 7.84 (*3.32*), and for ML: 6.90 (*2.94*).

Table 3. *Numeracy skills regressed on CNT pretest, Clever Together, born late preterm, small for gestational age, and interactions between conditions and mild perinatal adversities (N = 375)*

Measure	Beta (*SE*)	*F (df=373)*	Mean Square	*p*-value
Intercept	82.41 (*2.13*)	1499.96	147995.72	<.001
Main effects				
Cohort	-1.67 (*1.12*)	2.55	251.63	.138
CNT pretest	13.55 (*1.03*)	171.78	16949.26	<.001
Clever Together (vs. *Living Books*)	.20 (*1.23*)	.03	2.58	.872
Late preterm	-3.55 (*2.23*)	2.55	251.63	.111
Small for gestational age	.94 (*1.68*)	.31	30.80	.577
Two-way interaction				
Late preterm * *Clever Together*	7.88 (*3.32*)	5.63	555.01	.018
Small for gestational age * *Clever Together*	-1.46 (*2.36*)	.38	37.64	.537



$N_{full\ term/Clever\ Together}$= 161; $N_{preterm/Clever\ Together}$= 18; $N_{full\ term/Control}$ = 174; $N_{preterm/Control}$= 22.

Figure 5. Adjusted, standardized mean scores on the standardized CNT posttest for children born late preterm versus full term assigned to *Clever Together* or the control condition

Effect sizes of the intervention were calculated for the group as a whole and separately for children born late preterm and children born full term (Table 4). For the group as a whole, a small, non-significant, positive effect of *Clever Together* on numeracy skills at the end of senior kindergarten year was found (*Cohen's d* = .15, *CI* = -.05 / .35). In the group born full term, the effect size was close to zero (*Cohen's d* = .08, *CI* = -.13 /.30). However, *Clever Together* produced a large effect in the late preterm group (*Cohen's d* = .71, *CI* = .07 / 1.36).

Table 4. *Effect sizes of Clever Together for the complete group, children born late preterm and children born full term separately.*

		CNT posttest		
Dataset	**group**	***n***	***Cohen's d***	***95% CI***
Complete sample *(*N* = 375)	Full term	335	.08	-.13 / .30
	Late preterm	40	.71	.07 /1.36
	Total group	375	.15	-.05 /.35

Discussion

Previous studies showed that late preterm children attending kindergarten, although generally at risk for developing academic delays (Chyi et al., 2008), are highly susceptible to a digital early literacy intervention (*Living Letters*); after working with this program they even tend to outperform their peers. This result was typical for late preterm children, while both children with another mild perinatal adversity (i.e. small for gestational age) and children without mild perinatal adversities did not benefit from working with the program (Merkelbach et al., 2018), suggesting that when it comes to susceptibility to the (educational) environment, children with mild perinatal adversities should not be treated as a homogenous group. In line with these results we expected late preterm children also to benefit from *Clever Together*, a digital intervention highly similar to *Living Letters* in approach and design, but targeting a different academic skill: early numeracy. Children born small for gestational age and children without mild perinatal adversities were however not expected to benefit.

Results offer support for the hypothesis that *Clever Together* can boost early numeracy skills, but only in subgroups susceptible to its particular features. No interaction between small for gestational age and condition was found, and when considering the group as a whole, not making a distinction between those with or without perinatal adversities, also no positive effects of *Clever Together* could be distinguished (*Cohen's d* = .15, *CI* = -.05/.35), while late preterm children clearly benefitted from working with the program (*Cohen's d* = .71, *CI* = .07 / 1.36). Consistent with the differential susceptibility model (Belsky & Pluess, 2009), when assigned to the control condition, late preterm children lagged behind as compared to their peers, while they outperformed their peers after having worked with *Clever Together*.

Features central to both *Clever Together* and *Living Letters* thus seem to meet the needs of late preterm children particularly well. Although both programs also show strong resemblance in design, substantive features are most likely to have led to the learning gains experienced by late preterm children. We hypothesize that the high levels of repetition, structure, guidance, and feedback, central to both *Clever Together* and *Living Letters*, might facilitate learning in late preterm children. A positive effect of these features on especially late preterm children is plausible, because preterm birth is associated with increased levels of maternal stress during pregnancy (Dole, Savitz, Hertz-Picciotto, Siega-Riz, McMahon, & Buekens, 2003; Mulder, Robles de Medina, Huizink, Van den Bergh, Buitelaar , & Visser, 2002), which in turn is predictive for increased levels of fearfulness (Pike, 2004) and stress reactivity (Meaney, 2001) in offspring. In an educational environment these features are likely to be expressed as performance- and test anxiety, which are known to have detrimental effects on performance (McDonald, 2001). In normal school settings increased levels of stress reactivity might thus cause children to shut themselves from learning experiences (Van der Kooy-Hofland et al., 2012). In *Clever Together* however the repetition, structure, feedback, and guidance central to the program help clarify the task and hand. Because task clarity is associated with lowering of levels of cardiovascular reactivity to stress (Richter & Gendolla, 2006), we could assume that the features of *Clever Together* (and *Living Letters*) thus result in lower levels of stress through providing high levels of clarity and predictability, thereby facilitating learning in children normally possibly too stressed to effectively process all information presented. This proposed biological mechanism underlying the increased susceptibility to the digital learning environment found in late preterm children, is however still highly speculative. More research is needed to identify which exact features support the learning of late preterm children and why.

Conclusion

The digital early numeracy intervention *Clever Together* can boost the early numeracy performance of kindergartners born late preterm, while other children do not benefit from this intervention. On the other hand, late preterm children fall behind when assigned to a control condition, following the pattern as described by the differential susceptibility model. This pattern does not hold for children born small for gestational age. As a possible explanation for the effectivity of *Clever Together* in preterm children we expect that structure, guidance, and feedback provided by this program might have a soothing effect on children born late preterm, a group expected to experience higher levels of stress reactivity. However, more research is needed to found this speculative hypothesis.

Unavoidably this study has some limitations. We tested the effect of late preterm birth on the results of a digital program in the field of numeracy, expecting a similar result as was found for a literacy program. However, it should be noted that the studies looking into effects of *Living Letters* (Merkelbach et al., 2018) and *Clever Together* (current study) are not completely independent. In both studies the same control condition was used, thus including largely the same sample of children. Additionally, the current overarching study (in which thus three digital programs were included: *Clever Together*, *Living Letters*, and the control condition) has been designed with the prime purpose of exploring and stimulating early literacy skills. Therefore, teachers selected children on the basis of possible problems in the field of early literacy, instead of in the field of early numeracy. Children in this study who experienced problems in the field of numeracy, are thus those children who experience problems in both domains. These children might, even though literacy- and numeracy skills are highly correlated (in total sample: $r = .589$, $p < .001$), differ from children who only experience problems in the field of numeracy. Another limitation is that we can only speculate about effective functionalities in *Clever Together* and the biological mechanisms explaining this effectivity.

Although details of the *Clever Together* numeracy intervention need further study, we can conclude that those children having experienced mild perinatal adversity, but only when born preterm, benefit from numeracy training via *Clever Together*.

References

Aunio, P., & Niemivirta, M. (2010). Predicting children's mathematical performance in grade one by early numeracy. *Learning and Individual Differences*, 427-435.

Basten, M., Jaekel, J., Johnson, S., Gilmore, C., & Wolke, D. (2015). Preterm birth and adult health. *Psychological Science*, 1608-1619.

Belsky, J., & Pluess, M. (2009). Beyond diathesis stress: differential susceptibility to environmental influences. *Psychological Bulletin*, 885-908.

Bickel, R. (2007). *Multilevel Analysis for Applied Research It's just regression!* New York: The Guilford Press.

Centraal Bureau voor de Statistiek. (2013). *Documentatierapport Perinatale Registratie (PRN).*

Chyi, L., Lee, H., Hintz, S., Gould, J., & Sutcliffe, T. (2008). School outcomes of late preterm infants: Special needs and challenges for infants born at 32 to 36 weeks gestation. *The Journal of Pediatrics*, 25-31.

De Jong, M., Verhoeven, M., & Van Baar, A. (2012). School outcome, cognitive functioning, and behaviour problems in moderate and late preterm children and adults: a review. *Seminars in Fetal and Neonatal*, 163-169.

Desoete, A., Ceulemans, A., De Weerdt, F., & Pieters, S. (2010). Can we predict mathematical learning disabilities from symbolic and non-symbolic comparison tasks in kindergarten? Findings from a longitudinal study. *British Journal of Educational Psychology*, 64-81.

Dole, N., Savitz, D., Hertz-Picciotto, I., Siega-Riz, A., McMahon, M., & Buekens, P. (2003). Maternal stress and preterm birth. *American Journal of Epidemiology*, 14-24.

Gervasoni, A. (2001). Special programs for children who are low attaining in mathematics: do they help? *24th Annual MERGA conference*.

Griffin, S. (2004). Building number sense with Number Worlds: a mathematics program for young children. *Early Childhood Research Quarterly*, 173-180.

Gurian, M. (2010). *Boys and girls learn differently! A guide for teachers and parents.* San Fransisco: Jossey-Bass.

Koerhuis, I., & Keuning, J. (2011). *Wetenschappelijke verantwoording van de toetsen rekenen voor kleuters.* Arnhem: Cito.

Larroque, B., Bertrais, S., Czernichow, P., & Léger, J. (2001). School difficulties in 20-year-olds who were born small for gestational age at term in a regional cohort study. *Pediatrics*, 111-115.

Little, R. (1986). Survey nonresponse adjustments for estimates of means. *International Statistique Review*, 139-157.

McDonald, A. (2001). The prevalence and effects of test anxiety in school children. *Educational Psychology*, 89-101.

Meaney, M. (2001). Maternal care, gene expression, and the transmission of individual differences in stress reactivity across generations. *Annual Review of Neuroscience*, 1161-1192.

Méray, N., Reitsma, J., Ravelli, A., & Bonsel, G. (2007). Probabilistic record linkage is a valid and transparent tool to combine databases without patient identification number. *Journal of Clinical Epidemiology*, 883e1-883e11.

Merkelbach, I., Plak, R., & Rippe, R. (2018). Reproducibility of young learners' susceptibility to the learning context. *Learning and Individual Differences*, 167-175.

Mulder, E., Robles de Medina, P., Huizink, A., Van den Bergh, B., Buitelaar , J., & Visser, G. (2002). Prenatal maternal stress: effects on pregnancy and the (unborn) child. *Early Human Development*, 3-14.

Nepomnyaschy, L., Hegyi, T., Ostfeld, B., & Reichman, N. (2012). Developmental outcomes of late-preterm infants at 2 and 4 years. *Maternal and Childen Health Journal*, 1612-1624.

O'Keeffe, M., O'Callaghan, M., Williams, G., Najman, J., & Bor, W. (2003). Learning, cognitive and attentional problems in adolescents born small for gestational age. *Pediatrics*, 301-307.

Paz, I., Gale, R., Laor, A., Danon, Y., Stevenson, D., & Seidman, D. (1995). The cognitive outcome of full-term small for gestational age infants at late adolosence. *Obstetrics & Gynecology*, 452-456.

Pike, I. (2004). Maternal stress and fetal responses: Evolutionary perspectives on preterm delivery. *American Journal of Human Biology*, 55-65.

Plak, R., Merkelbach, I., Kegel, C., Van IJzendoorn, M., & Bus, A. (2016). Brief computer interventions enhance emergent academic skills in susceptible children: a gene-by-environment experiment. *Learning and Instruction*, 1-8.

Poulsen, G., Wolke, D., Kurinczuk, J. J., Boyle, E. M., Field, D., Alfirevic, Z., & Quigley, M. A. (2013). Gestational age and cognitive ability in early childhood: A population based cohort study. *Paediatric and Perinatal Epidemiology*, 371-379.

Richter, M., & Gendolla, G. (2006). Incentive effects on cardiovascular reactivity in active coping with unclear task difficulty. *International Journal of Psychophysiology*, 215-225.

Searle, A., Smithers, L., Chittleborough, C., Gregory, T., & Lynch, J. (2017). Gestational age and school achievement: a population study. *Archives of Disease in Childhood: Fetal & Neonatal Edition*, 409-415.

Shah, P., Kaciroti, N., Richards, B., Oh, W., & Lumeng, J. (2016). Developmental outcomes of preterm infants from infancy to kindergarten. *Pediatrics*, 1-10.

Sommerfelt, K., Andersson, H., Sonnander, K., Ahlsten, G., Ellersten, B., Markestad, T., . . . Bakketeig, L. (2000). Cognitive development of term small for gestational age children at five years of age. *Archives of Disease in Childhood*, 25-30.

Strauss, R. (2000). Adult functional outcomes of those born small for gestational age: twenty-six-year follow-up of the 1970 British Birth Cohort . *Journal of the American Medical Association*, 625-632.

Van der Kooy-Hofland, V., Van der Kooy, J., Bus, A., Van IJzendoorn, M., & Bonsel, G. (2012). Differential susceptibility to early literacy intervention in children with mild perinatal adversities: short- and long-term effects of a randomized controlled trial. *Journal of Educational Psychology*, 337-349.

Windhorst, D., Rippe, R., Mileve-Seitz, V., Verhulst, F., Jaddoe, V., Noppe, G., . . . Bakermans-Kranenburg, M. (2017). Mild perinatal adversities moderate the association between maternal harsh parenting and hair cortisol: Evidence for differential susceptibility. *Developmental Psychobiology*, 324-337.

Woythaler, M., McCormick, M., Mao, W., & Smith, V. (2015). Late preterm infants and neurodevelopmental outcomes at kindergarten. *Pediatrics*, 424-431.

Wynn, K. (1990). Children's understanding of counting. *Cognition*, 155-193.

Wynn, K. (1992). Addition and subtraction by human infants. *Nature*, 749-750.

Chapter 6

Stress reactivity during challenging educational situations in children born late preterm

Introduction: In kindergarten, late preterm children show increased susceptibility to structured and guided learning environments. One explanation for this observation might be found in increased levels of stress reactivity in late preterm children. A direct link between late preterm birth and increased levels of stress reactivity has however not been established. *Methods:* Kindergarten children (*N* = 250) took part in an educational test session administered by the researchers. During this session for each child a set of three saliva samples was collected, from which levels of the stress hormone cortisol were determined. Patterns of stress reactivity were compared between children born late preterm and their full term peers. *Results:* Children born late preterm show increased levels of stress reactivity. In anticipation of the unfamiliar test situation, cortisol levels are high. However, cortisol levels drop quickly when the test situation turns out to be highly structured and guided. In children born full term no fluctuations in cortisol levels between measurements were found. *Conclusion:* Children born late preterm show increased levels of stress reactivity. In unfamiliar situations they seems to show an increased stress response, while in guided and structured situations reduced levels of stress are experienced. Offering late preterm children an educational environment with high(er) levels of guidance and structure could thus facilitate learning in this group.

Children born late preterm (between 34th and 38th week of pregnancy) are at increased risk for experiencing a range of difficulties later in life. These children are more likely to encounter health problems, socio-emotional problems, and cognitive problems (Morse, Zheng, Tang, & Roth, 2009) which can result in diverse academic delays. At ages five and seven late preterm children experience more problems in both numeracy skills and word reading than their full term peers (Poulsen et al., 2013). Although predictive for academic delays, recent research suggests however that late preterm birth cannot be straightforwardly considered a vulnerability factor for such problems, but might be considered a plasticity factor (Van der Kooy-Hofland, Van der Kooy, Bus, Van IJzendoorn, & Bonsel, 2012; Merkelbach et al., 2018). These recent studies show that children born late preterm show increased susceptibility to the qualities of their educational environment, both for better and for worse.

Digital interventions offering five year old children structure, repetition of assignments, continuous and adaptive feedback, and guidance, resulted in late preterm children outperforming their peers (Merkelbach et al., 2018). These results were achieved with low intensity interventions: children only worked with the digital programs once a week for ten minutes, over a period of just two to three months. Results were comparable for both an intervention promoting early literacy skills (i.e. *Living Letters,* Van der Kooy-Hofland et al., 2012; Merkelbach, et al., 2018) and an intervention promoting early numeracy skills (i.e. *Clever Together,* (Merkelbach, Plak, & Rippe, under review). When assigned to a digital control condition which did not contain the same essential features, late preterm children fell behind as compared to their peers. This pattern of experiencing exceptionally positive outcomes in positive environments and exceptionally poor outcomes in negative environments is described by the *differential susceptibility model* (Belsky & Pluess, 2009): children born late preterm can thus be considered highly susceptible to the qualities of the interventions. Full term children however were not affected: they did not benefit nor experience disadvantage from working with *Living Letters* or *Clever Together*. Which biological mechanism explains why late preterm children are especially susceptible to some or all of the qualities central to *Living Letters* and *Clever Together*, remains however unknown.

A potential explanation for why late preterm children responded especially well to interventions containing structure, repetition of assignments, continuous and adaptive feedback, and guidance, might be found in dysregulation of the stress response in these children. A recent study reported on increased dispersion of the level of the stress hormone cortisol in hair samples, a reliable indicator of cortisol release over an extended period of time (D'Anna-Hernandez, Ross , Natvig, & Laudenslager, 2011), in late preterm children and children born small for gestation age (i.e. children with mild perinatal adversities) as compared to children who did not experience such adversity (Windhorst, et al., 2017). Children who had experienced mild perinatal adversities showed decreased levels of hair cortisol in case of high levels of maternal harsh parenting. However, when maternal harsh parenting was low, high levels of hair cortisol were found in this group. For children without mild perinatal adversities parenting style was not predictive of hair cortisol levels (Windhorst, et al., 2017). Since chronic stress can result in down regulation of the *Hypothalamus-Pituitary-Adrenal axis* (HPA-axis) (Fries, Hesse, Hellhammer, & Hellhammer, 2005), the bodily mechanism controlling the secretion of cortisol (Kolb & Wishaw, 2009), low levels of hair cortisol could indicate high levels of chronic stress. It could thus be assumed that the susceptible group experienced higher levels of stress in negative environments (i.e. high levels of harsh parenting) and lower levels of stress in positive environments (i.e. low levels of harsh parenting) than their non-susceptible peers.

Late preterm children may experience high levels of stress in both negative home environments and negative educational environments. On the other hand, a positive educational environment, as offered by programs like *Living Letters* and *Clever Together*, could result in a decrease of experienced stress in this group. High levels of stress could cause children to shut themselves from learning experiences, suggesting that lowering these stress levels would facilitate learning. However, for these programs to influence acute stress levels, directly facilitating learning, children late preterm would have to be able to show rapid and substantial changes in experienced stress, and thus experience increased levels of *stress reactivity.*

From an evolutionary perspective, increased levels of stress reactivity in late preterm children would be plausible, too. Talge, Neal, & Glover (2007) state that having high levels of stress reactivity can be evolutionary adaptive, because this trait can be useful when living in harsh and unpredictable environments. High stress reactivity facilitates rapid bio behavioral changes (e.g. increased heart rate) that prepare the body for direct action (Sapolsky, 2015). Already while still in the uterus, the fetus can receive cues about the quality of the environment (s)he will grow up in. High levels of maternal stress for example have a signaling function to the fetus, predicting an unsafe environment (e.g. Pike, 2004) in which high stress reactivity could thus be desirable. Through the process of *early programming* changes in the uterus due to maternal stress can evoke changes in the HPA-axis of the fetus (Matthews, 2002), resulting in these higher levels of stress reactivity. At the same time, high levels of maternal stress can also increase chances of late preterm birth (Dole, et al., 2002). Because high levels of maternal stress predict for both increased stress reactivity of the child and late preterm birth, considerable overlap between these factors would be likely. To our knowledge however, this association has not been evaluated.

In this study we compared acute stress responses between late preterm children and their full term peers in an educational environment. We hypothesize that children born late preterm will show elevated cortisol levels in anticipation of a potentially stressful event (i.e. being taken out of the classroom by an unfamiliar test assistant in order to complete performance tests) and a pronounced drop in cortisol levels when the child is consequently placed in an environment comparable to *Living Letters* and *CleverTogether,* in this case a testing environment offering structure, repetition of assignments, continuous and adaptive feedback, and guidance. We expected no changes in stress levels of full term children. Since acute changes in stress cannot be measured with the use of hair cortisol, we used successive salivary cortisol levels as an indicator of stress reactivity (e.g. Blair, Granger, & Peters Razza, 2005). To this end, salivary cortisol levels were determined at the start, halfway, and at the end of the test session.

Method

Design

The current study is embedded in a large scale study on the effects of digital educational programs in five-year-olds: *What Works for Whom* (e.g. Plak, Merkelbach, Kegel, Van IJzendoorn, & Bus, 2016). The project consisted out of two successive research waves (2012/2013 and 2013/2014) in which a total of 1972 children participated. The current sample (N = 440) was randomly selected from the group of children participating in the second research wave (N = 1083). As a critical test of stress reactivity, saliva samples were collected during a one-on-one test situation lasting on average almost 35 minutes (M = 34.81, SD = 7.14).

Participants

The subsample of 440 kindergarten children was subjected to an additional test session that took place after the intervention period, during which a series of three saliva samples was collected. Children from 63 different schools across the Netherlands were randomly selected for additional testing. Complete data for testing stress reactivity patterns were available for 250 children (see flow diagram of data attrition in Figure 1).

Procedure

After the intervention period (in which efficacy of e.g. *Living Letters* and *CleverTogether* was explored) conducted in the context of the overarching research study, trained test assistants, instructed to be guiding, supportive, and to offer continuous feedback, visited a selection of participating preschool classrooms for additional individual testing. Testing took place in a quiet, separate room in the school. During test sessions, children completed several tasks related to literacy skills and impulse control. Saliva samples were collected before, halfway through, and at the end of this test session that lasted about 35 minutes.

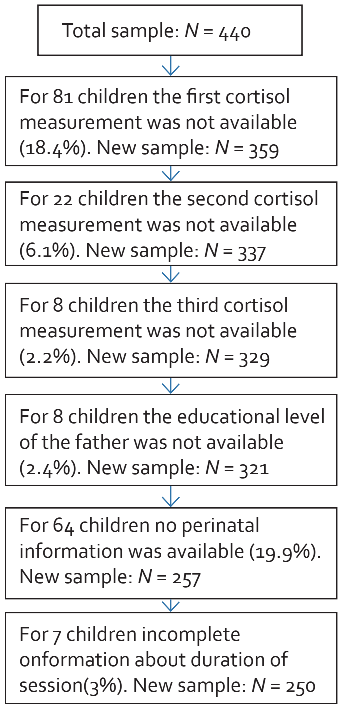


Figure 1. Flow diagram data attrition

Measurements

Perinatal information – The Netherlands Perinatal Registry (Stichting Perinatale Registratie Nederland, 2011) contains comprehensive data on pregnancy, pregnancy care (interventions, referrals), and pregnancy outcomes. These variables are recorded by the health care provider during prenatal care, delivery, and the neonatal and lying-in period. The register covers approximately 96% of all deliveries in the Netherlands. The data from three registers (the National Obstetric Database by midwives, the National Obstetric Database by gynecologists, and the National Neonatal/Pediatric Database) are annually sent to the national registry office, where a number of range and consistency checks are conducted. The perinatal registry can be accessed by researchers, provided that they have the written permission of the mother. Missing values in our sample were largely due to non-consent for retrieving data (61%). Attrition was also due to failure of connecting

registry data to the research database (39%). Criterium for being classified as born late preterm was to have a gestational age at birth of 34 - 37 weeks and 6 days.

Stress reactivity

*S*tress reactivity was derived from the cortisol levels in saliva samples. Since cortisol levels in saliva tend to lag behind five to twenty minutes, and to peak ten to thirty minutes after exposure to a stressor (Kirschbaum & Hellhammer, 2000), the first saliva sample provides an indication of the level of stress experienced while anticipating being taken from the classroom setting by a stranger in order to complete tests. The second sample indicates the stress level after the first few minutes of testing, while the third sample indicates the level of stress halfway through the test session.

Salimetrics saliva swabs (rimless non-sterile polypropylene-round base) were used for saliva collection. To collect saliva samples, test assistants held the swab under the tongue of the child for at least 60 seconds. If they asked children were allowed to hold the swab themselves. The majority of swabs absorbed enough saliva for analysis. Samples were analyzed at the Salimetrics Saliva Lab (Cambridge, UK), which is HTA (Human Tissue Authority) licensed. This Salimetrics Centre of Excellence uses GLP and level 2 containment protocols and equipment. The analyses are performed on an 8 fixed tips, fast wash system, robotic manipulator, two incubator units (MIOs), Tecan Sunrise and Tecan Hydroflex controlled using TecanEvoWare and TecanEvoWarePLUS software with Liebert UPS power backup. Units are enclosed in a Bigneat HEPA filter cabinet. Samples were stored at -80°C. When assaying samples were kept at 4°C on ice at all times and returned to -80°C directly after use.

Statistical analyses

Assessment of the missingness mechanism was first performed (Little, 1986) because nonrandom missingness would lead to biased complete case analysis. To further account for missing data, analyses using imputed data were repeated twice. To assess sensitivity of missing data handling, once imputed data were generated under strict assumptions concerning the distribution of data (m = 100, N = 440) and empirical range, and once under less strict assumptions (m = 100, N = 440).

To assess differences in patterns of stress reactivity in children born late preterm and their full-term peers mixed-model ANCOVAs were applied, with repeated measures for salivary cortisol levels (μg/dL) at the three different time points. 'Late preterm birth vs. full term birth' was considered as the between-subjects' factor. As covariates in the analysis were entered: time (in minutes) between the different saliva measurements and time of day at the start of the test session. Time of day was included since cortisol levels do not only fluctuate in response to stress but also follow a diurnal pattern (Van den Bergh, Van Calster, Smits, Van Huffel, & Lagae, 2008).

6

Results

Missing data

Based on the procedure as described by Little (1986) we were able to reject the null hypothesis of data being not missing completely at random ($\chi^2 = 20.78$, $p = .187$), therefore complete case analysis could be applied.

Sample characteristics

Children were on average 67.13 (*SD* = 3.84) months old and a small majority was male (57.2%). Generally, average cortisol levels did not differ largely between time points (start of the session, halfway through the sessions, and at the end of the session) at which saliva collection took place (range between .08 µg/dL and .09 µg/dL) (Table 1).

Table 1. *Descriptives (means and standard deviations / percentages) of background variables (i.e. age, sex, father's educational level), late preterm birth, and cortisol assessments)*

***Descriptives* (*N* = 250)**	**Mean (*SD*) / percentage**
Age, months	67.16 (*3.85*)
Sex, male	57.8%
Educational level father (max = 6)	3.85 (*1.40*)
Cortisol level 1: beginning test session (µg/dL)	.09 (*.05*)
Cortisol level 2: halfway test session (µg/dL)	.09 (*.07*)
Cortisol level 3: end test session (µg/dL)	.08 (*.05*)
Time (min.) between collection sample 1 and sample 2	16.65 (*3.23*)
Time (min.) between collection sample 2 and sample 3	17.85 (*4.35*)
Time of day at start test session	10.53 (*1.35*)
Born late preterm	10.0%

Patterns in cortisol trajectory

A mixed design was applied to test differences between late preterm vs. full term children, differences within participants (i.e. the course of cortisol levels in the three successive measurements), and the interaction between these between- and within-subject factors.

The assumption of sphericity was violated (*Mauchly's W* = .94, $p < .001$) and the Greenhouse-Geisser coefficient exceeded .75 (*GG* = .94). The Huyn-Feldt correction (*HF* = .96) was therefore applied (Field, 2009). Of all between-subject variables and covariates, only the time of day at which the saliva sample was collected, showed a significant association with the average level of salivary cortisol (F (1, 245) = 8.54, $p = .004$) over the three time points. Considering the main and interaction effects within subjects, only the interaction between the trajectory of salivary cortisol levels and late preterm birth was present (F (1.93, 471.76) = 4.78, $p = .010$) (Table 2).

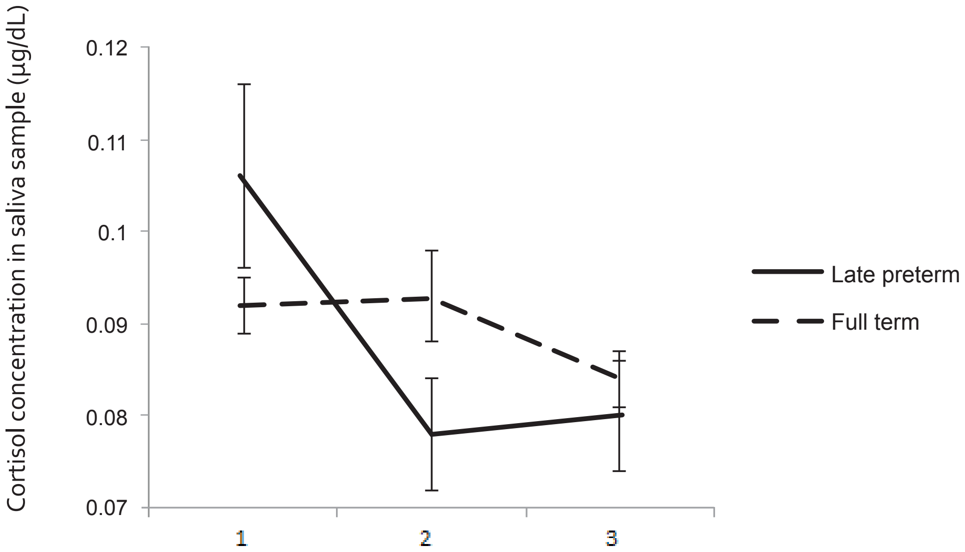
Table 2. *Results of mixed design ANCOVA with salivary cortisol levels as repeated outcome measure and late preterm and time measurements as between-subjects variables*

Predictor	***F(df* = 471.76)**	***p***
Between		
Preterm birth	.04	.846
Time between session 1 and 2, minutes	.37	.546
Time between session 2 and 3, minutes	.83	.365
Time of day at start test session	8.54	.004
Within		
Cortisol trajectory[1]	.41	.654
Cortisol trajectory * time of day at start test session	.33	.709
Cortisol trajectory * time between session 1 and 2	.09	.909
Cortisol trajectory * time between session 2 and 3	.34	.706
Cortisol trajectory * born late preterm	4.78	.010

1. Course of the cortisol levels found in the three successive measurements

The interaction between late preterm birth and cortisol trajectory is depicted in Figure 2, based on mean salivary cortisol levels corrected for background variables (i.e. time between the measurements and time of day).

For full term children cortisol (and thus stress-) levels remained constant over the three measurement points. For late preterm children cortisol levels dropped between the first and second measurement. Additionally, the first measurement was elevated, whereas in children with mild perinatal adversities, this was not the case. Between the second and the third point stress levels remained constant.



$N_{\text{no mild perinatal adversities}}$ = 173; $N_{\text{mild perinatal adversities}}$ = 77 ; $N_{\text{full term}}$ = 225; $N_{\text{late preterm}}$ = 25

Figure 2. Cortisol trajectory (µg/dL) late preterm vs. full term children

Post-hoc analysis confirmed that the significant differences found in the course of salivary cortisol between late preterm children and full term children were attributable to differences in the gradient of slopes between the first and the second measurement (F (1, 245) = 7.91, p = .005), and not to differences between the second and third measurement (F (1, 245) =1.28, p = .260).

Repetition of the analysis in imputed datasets yielded similar results (Supplementary Table 1). All results remained constant, including the significant finding regarding the interaction between late preterm and cortisol trajectory (complete cases: F (1, 242) = 4.61, p = .010, multiple imputation under strict assumptions: F (1, 432) = 6.60, p = .001), multiple imputation without strict assumptions: F (1, 432) = 8.03, p <.001).

Discussion

We expected five-year-old children born late preterm to show increased levels of stress reactivity as compared to their full term peers, and thus to show elevated levels of stress in anticipation of a possibly stressful event (i.e. testing by an unfamiliar adult). However, since late preterm children have been shown to react well to guided, structured environments (Merkelbach et al., 2018), a decrease in experienced stress when the test

setting turned out to be highly structured and guided was expected in this group. On the other hand, we expected the stress levels of full term children to remain relatively constant over the course of these events since they are not expected to experience increased levels of stress reactivity. Results confirm that patterns in salivary cortisol, which were used as a proxy of experienced stress, differ between late preterm children and their full term peers (F (2, 471.76) = 4.78, p = .010). In anticipation of being tested, salivary cortisol levels were elevated in children born late preterm, but as soon as testing had started, a clear decrease took place. In full term children stress levels remained constant (F (1, 245) = 7.91, p = .005).

Children born late preterm do thus show a different neurobiological reaction pattern during stressful events, like anticipating being tested by a stranger, as compared to their peers, suggesting an elevated level of stress reactivity (Kolb & Whishaw, 2009) in this group. Although increased stress reactivity can be an asset in unpredictable situations in which (physical) danger is a considerable possibility, since it offers the opportunity to react fast (Sapolsky, 2015), in low-risk situations like an educational setting, increased stress can stand in the way of effective learning. In line with the finding of Blair, Granger and Peters Razza (2005) that stress reactivity is associated with executive function, self-regulation, and letter knowledge, we might expect that deviations in stress reactivity patterns might prevent children from absorbing information from the environment.

Although elevated at the start of the session, not during the entire session did cortisol levels remain high in children born late preterm. When placed in the testing situation, stress levels almost immediately seemed to decrease. The testing environment employed in the current study shows great similarity to the environment offered by the digital educational interventions *Living Letters* (stimulating early literacy) and *Clever Together* (stimulating early numeracy) that have been found particularly useful to support the learning of children born late preterm (Merkelbach et al., 2018). Both the test environment and these digital programs offered structure, continuous and adaptive feedback, and guidance. Possibly these features provided children born late preterm with much needed support, clarifying the test at hand. As was the case during the testing situation, we might thus expect that while interacting with *Living Letters* and/or *Clever Together* also a decrease in the level of stress experienced by late preterm children took place. The possibly stress reducing features of these programs (and the testing situation) might explain why they are very helpful for children born late preterm, while they do not sort effects for their full term peers.

These results thus show that children born late preterm experience higher levels of stress reactivity in educational settings than their full term peers. Additionally, our findings suggest that structure, continuous and adaptive feedback, and guidance could be helpful to support the learning of highly stress reactive children, such as children late preterm, since they might have stress reducing effects. However, a direct link between the effectivity of such features in promoting learning and stress reactivity levels has not been established yet. Additionally, more research is needed to identify which specific features of the learning environment are essential to facilitate the learning of children showing high levels of stress reactivity.

Conclusion

Five-year-olds born late preterm do show a different neurobiological reaction pattern to stressful events than their full term peers, suggesting higher levels of stress reactivity in this group. For children who were born late preterm, the prospect of being placed in an unfamiliar, unpredictable situation resulted in increased salivary cortisol levels. However, when placed in a highly structured, guiding situation, stress levels decreased almost immediately and dropped even below the cortisol levels of their full term peers. For children born full term no fluctuations in cortisol levels have been detected. These results suggest that structured and guiding intervention programs might be highly effective for late preterm children especially, since they might have stress reducing qualities that facilitate learning in this group.

References

Belsky, J., & Pluess, M. (2009). Beyond diathesis stress: differential susceptibility to environmental influences. *Psychologial Bulletin*, 885-908.

Blair, C., Granger, D., & Peters Razza, R. (2005). Cortisol reactivity is positively related to executive function in preschool children attending Head Start. *Child Development*, 554-567.

Centraal Bureau voor de Statistiek. (2013). *Documentatierapport Perinatale Registratie (PRN).*

D'Anna-Hernandez, K., Ross , R., Natvig, C., & Laudenslager, M. (2011). Hair cortisol levels as a retrospective marker of hypothalamic-pituitary axis activity throughout pregnancy: comparison to salivary cortisol. *Physiology and behavior*, 348-353.

Dole, N., Savitz, D., Hertz-Picciotto, I., Siega-Riz, A., McMahon, M., & Bueskess, P. (2002). Maternal stress and preterm birth. *American Journal of Epidemiology*, 14-24.

Fries, E., Hesse, J., Hellhammer, J., & Hellhammer, D. (2005). A new view of hypocortisolism. *Psychoneuroendocrinology*, 1010-1016.

Kirschbaum, C., & Hellhammer, D.(2000). Salivary cortisol. In G. Fink, *Encyclopedia of Stress* (pp. 379-383). Trier: Academic Press.

Kolb, B., & Whishaw, I. Q. (2009). *Fundamentals of Human Neuropsychology.* New York : Worth Publishers.

Little, R. (1986). A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association* , 1198-1202.

Matthews, S. (2002). Early programming of the hypothalamo-pituitary-adrenal axis. *Trends in Endocrinology & Metabolism*, 373-380.

Merkelbach, I., Plak, R., & Rippe, R. (2018). Reproducibility of young learners' susceptibility to the learning context. *Learning and Individual Differences,* 167-175.

Merkelbach, I., Plak, R., & Rippe, R.[2] (under review). Differential efficacy of a digital numeracy program in kindergarten

Morse, S., Zheng, H., Tang, Y., & Roth, J. (2009). Early school-age outcomes of late preterm infants. *Pediatrics*, e622-e629.

Pike, I. (2004). Maternal stress and fetal responses: evolutionary perspectives on preterm delivery. *American Journal of Human Biology*, 55-65.

Plak, R., Merkelbach, I., Kegel, C., Van IJzendoorn, M., & Bus, A. (2016). Brief computer interventions enhance emergent academic skills in susceptible children: A gene-by-environment experiment. *Learning and Instruction*, 1-8.

Poulsen, G., Wolke, D., Kurinczuk, J., Boyle, E., Field, D., Alfirevic, Z., & Quigley, M. (2013). Gestational age and cognitive ability in early childhood: a population-based cohort study. *Infant and Child Health*, 371-379.

Sapolsky, R. (2015). Stress and the brain: individual variability and the inverted U. *Nature Neuroscience*, 1344-1346.

Stichting Perinatale Registratie Nederland. (2011). *Grote lijnen 10 jaar perinatale registratie Nederland*. Utrecht: Stichting Perinatale Registratie Nederland.

Talge, N., Neal, C., & Glover, V. (2007). Antenatal maternal stress and long-term effects on child neurodevelopment: how and why? . *The Journal of Child Psychology and Psychiatry*, 245-261.

Van den Bergh, B., Van Calster, B., Smits, T., van Huffel, S., & Lagae, L. (2008). Antenatal maternal anxiety is related to HPA-axis dysregulation and self-reported depressive symptoms in adolescence: a prospective study on the fetal origins of depressed mood. *Neuropsychopharmacology*, 536-545.

Van der Kooy-Hofland, V., Van der Kooy, J., Bus, A., Van IJzendoorn, M., & Bonsel, G. (2012). Differential susceptibility to early literacy intervention in children with mild perinatal adversities: short- and long-term effects of a randomized controlled trial. *Journal of educational psychology*, 337-349.

Windhorst, D., Rippe, R., Mileva-Seitz, V., Verhulst, F., Jaddoe, V., Noppe, G., . . . Bakermans-Kranenburg, M. (2017). Mild perinatal adversities moderate the association between maternal harsh parenting and hair cortisol: evidence for differential susceptibility. *Developmental Psychobiology*, 324-337.

Supplementary table 1. *Results of repeated measures ANCOVA with salivary cortisol levels as outcome measure in complete cases, and in two imputed datasets*

	Complete cases analysis		**Imputation with strict assumptions**		**Imputation without strict assumptions**	
Predictor	***F(df = 242)***	***p***	***F(df = 242)***	***p***	***F(df = 242)***	***p***
Cortisol trajectory	.41	.654	.34	.560	.57	.451
Cortisol trajectory * time of day at start test session	.33	.709	.28	.597	.19	.663
Cortisol trajectory * time between session 1 and 2	.09	.909	.11	.740	.08	.777
Cortisol trajectory * time between session 2 and 3	.34	.706	.33	.566	.46	.498
Cortisol trajectory * born late preterm	4.78	.010	6.60	.011	8.03	.005

Chapter 7

General Discussion

Mild perinatal adversities such as being born late preterm or small for gestational age have long been considered vulnerability factors for academic success. However, a small scale study (Van der Kooy-Hofland, Van der Kooy, Bus, Van IJzendoorn, & Bonsel, 2012) established that kindergarten children with such mild perinatal adversities showed increased susceptibility to their literacy environment, both for better and for worse. In this dissertation we aim at both strengthening this body of knowledge through means of replication, and at expanding it by addressing new hypotheses. As a result this dissertation addresses the following issues:

1. Can findings of an earlier study on the effects of a digital program called *Living Letters* on the reading performance of young children with mild perinatal adversities be replicated, by using a larger sample size and a planned missing data approach?
2. If replication of initial program findings (partially) fails, are potential reasons for non-replication found in:
   a) Heterogeneous responses to *Living Letters* by children born late preterm and children small for gestational age?
   b) Sensitivity and quality of post testing?
   c) Fidelity of implementation, possibly explained by teacher opinions on digital material?
3. Does *Clever Together*, a program comparable to *Living Letters* in substantive features, design, and duration and dosage, but targeting early numeracy instead of early literacy skills, yield similar results in children born late preterm, children born small for gestational age, and children without mild perinatal adversities?
4. Can we identify biological mechanisms possibly explaining differential effects?

To strengthen earlier conclusions, we aimed to replicate results of a small scale study (Van der Kooy-Hofland et al., 2012) in which kindergartners, who either were or were not subject to mild perinatal adversities (i.e. being born late preterm or small for gestational age), were exposed to a digital early literacy intervention (i.e. *Living Letters*). Initially, replication was sought through analyzing complete data of a new, larger study with a similar design as was used in the Van der Kooy-Hofland et al. (2012) study. When a partial lack of replicability was observed, we examined the utility of a planned missing data approach (Graham, Taylor, Olchowski, & Cumsille, 2006) to account for potentially diminished validity of outcome measures. As this approach did not fully explain the partial lack of replication, we shifted focus to the potential influence of the teacher on intervention effects, and to this end explored teacher opinions on the possibilities of digital interventions for supporting kindergarten children with learning vulnerabilities.

Van der Kooy-Hofland et al. (2012) found increased susceptibility in kindergartners with mild perinatal adversities to the effects of a digital early literacy intervention, *Living Letters*, offering continuous feedback, guidance and support. Adversity-exposed children fell behind in the control condition, but outperformed their peers when assigned to *Living Letters*. Effects were found both in the short term (directly after the intervention) and in the long term (one year later). In the current study however this increased susceptibility was replicated only for children born late preterm, but not for those born small for gestational age. In the Van der Kooy-Hofland (2012) study the distinction between these two types of mild perinatal adversities was - and could not - be made, due to a small sample size and consequently a lack of power (MacCallum, Browne, & Sugawara, 1996). In addition to only partial replication, effects found in the current study were clearly more modest than earlier findings: in the Van der Kooy-Hofland (2012) study Cohen's *d*s exceeded 1.00 both in the long and the short term, while in the current study the values for Cohen's *d* in the late preterm group ranged between .30 and .40.

The differences between these two studies might originate in decreased quality of the outcome measure in the current study. In the Van der Kooy-Hofland (2012) study, researchers administered both pre- and posttests, while in the current study teachers administered these tests. Differences in level of researcher control might partly explain why effect sizes in the Van der Kooy-Hofland (2012) study and the current study differ so strongly: teacher-administered testing might have introduced a higher amount of error into the scores. To evaluate this assertion a planned missing data approach was used to account for possibly diminished validity of outcome measures without losing power (Graham, Taylor, Olchowski, & Cumsille, 2006): high quality reference (gold standard) measures were administered to a randomly selected subsample of children. Gold standard measures were expected to be of superior sensitivity compared to the general posttest, because these measures both comprised more items and were administered by trained research assistants instead of teachers. By ensuring that the subsample was randomly selected, data were missing completely at random (MCAR) (Garnier-Villarreal, Rhemtulla, & Little, 2014). By using the general posttest as an auxiliary variable to the gold standard, a shared variance factor of these measurements could be identified, which could be considered a valid estimate of performance. We found that results could be replicated using such an approach, however only when only gold standard measures closely approaching the skills targeted by the intervention were included in the model. Addition of gold standard measures that focused on related, but not similar, skills as those targeted by the intervention (e.g. vocabulary as opposed to alphabetic knowledge and phonemic awareness) led to poorer model fit. However, for results that *could* be replicated using such planned missing data models, effects were in the same ballpark as those found without the use of planned missing data and did thus not approach the effects found in the Van der Kooy-Hofland et al. (2012) study. Because suboptimal quality of outcome measures in the current study does not seem to offer a valid explanation for the non-replicability of effect sizes of the Van der Kooy-Hofland et al. (2012) study, we might conclude that the original outcome measures used in the current study – as administered by the teachers - showed satisfactory sensitivity.

By using a planned missing data approach, we could thus conclude that lack of sensitivity of the outcome measure, possibly due to intrinsic qualities of the test or to administration of the test by the teacher instead of the researcher, did not explain why findings of Van der Kooy-Hofland et al.'s (2012) study could only be partly replicated. Teachers could however still be of influence, because they were not only in charge of pre- and post-testing but also managed the implementation of the intervention. To test if teacher characteristics were likely to influence the quality of implementation, teachers were asked to complete a questionnaire on their use of and opinions on digital (educational) material. Digital use and skill of the teacher, (perceived) results of digital material by the teacher, and teachers' expectations and beliefs regarding digital material, were all, directly or indirectly, associated with involvement in and effective implementation of such material. These associations show that teacher characteristics can predict quality of intervention implementation and thus suggest that teachers can influence study results in more ways than just through lower post-test fidelity. Presumably, either teachers have not been as steadfast in implementing the intervention at regular intervals as researchers would have been, or teachers have put children to work under suboptimal circumstances (e.g. noisy environment), which might have led to diminished results in the current study. Teachers who have negative beliefs and expectations about the potential effectivity of interventions might be more inclined to make such undermining choices. Additionally, teachers might have transferred negative expectations or opinions about digital interventions to their pupils, in turn possibly resulting in less investment by or motivation of pupils when working with *Living Letters*. Future studies should focus on the potential influence teachers have on the results of digital interventions. If the influence of the teacher indeed turns out to be substantial, motivating and training teachers for the use of digital material in the curriculum is of great importance.

Our study did not only aim at replication of previous results and explanation of deviations between these previous and current findings, but also at extending knowledge. Aiming to expand on the current body of knowledge, a series of new hypotheses was generated and tested. These addressed new, but related, areas of interest: we explored results of an intervention with an approach similar to *Living Letters*, but targeting a different academic area (i.e. early numeracy) in order to establish whether differential effects were early literacy specific, or if similar effects would also hold for other early academic skills. Furthermore we examined biological mechanisms possibly underlying differential effects found.

We looked into the effects of the digital program *CleverTogether*, highly similar in design to *Living Letters* but targeting another area of academic development in kindergartners, namely early numeracy. We found effects similar to those of *Living Letters*, suggesting that differential effects of supportive digital programs for late preterm kindergartners are not specific for early literacy but can apply to a broader range of academic skills. Additionally, we found support for the notion that children born late preterm show higher levels of stress reactivity as compared to their full term peers. This increased stress reactivity offers a plausible explanation for the good fit between highly supportive and guiding intervention programs, like *Living Letters* and *Clever Together*, and children born late preterm. These findings offer new insights and leads for new studies into the effectiveness of digital programs for children born late preterm and for other groups of stress reactive children. However, replication and further specification of these differential effects and their underlying mechanisms is necessary in order to optimally exploit the practical use of supporting digital programs in the daily classroom environment.

In conclusion, this dissertation shows that children born late preterm are susceptible to digital programs with features comparable to those of *Living Letters* and *Clever Together*, and that such programs can substantively contribute to the learning performance of these children. Additionally, we might assume that high levels of stress reactivity play a central role in explaining this increased susceptibility. However, especially in the light of the replication crisis (Maxwell, Lau, & Howard, 2015), we must be careful to draw firm and final conclusions. In order to fully understand why and under which circumstances digital programs can contribute to the learning gains of kindergartners born late preterm and/or kindergarteners with high levels of stress reactivity, a lot of hypotheses remain to be formulated and repeatedly tested. We would therefore like to conclude with some recommendations and considerations for future researchers studying this topic, or using comparable designs to study the (differential) effects of interventions:

When studying intervention effects, differential effects for subgroups should be taken into account. Considering only main effects might disregard important information and might thus deny vulnerable subgroups valuable learning opportunities: both *Living Letters* and *Clever Together* did for example not show main effects, but proofed supportive for a subgroup of children. When considering such possibly susceptible subgroups, researchers should be aware that subgroups that seem highly similar at first sight (and might even fall under the same denominator: e.g. late preterm and small for gestational age are both mild perinatal adversities) might still react differently to their (learning) environment. Researchers should therefore formulate hypotheses about which (biological) mechanisms are bound to underlie susceptibility, before determining which subgroups of children might show increased susceptibility and should be studied and which interventions should be used. While still in the phase of generating such hypotheses, using a range of small scale (e.g. pilot) studies to explore the effects of making small differences in study design, intervention design, and target group might be more suitable than using one large study (Van Teijlingen & Hundley, 2001) and testing only one specific hypothesis. When considering such small-scale studies, interpreting significance levels should not be the prime focus, because such studies are bound to have limited power. Instead, the direction of effects over a series of related studies should be examined to determine whether patterns can be observed, which could serve as the basis for formulating a well-founded 'final' hypothesis. At this stage large scale replication studies like the current project are of optimal value, because they than can be used to test the ecological validity of the final intervention. Additionally, although researchers should be aware that using planned missing data is only useful when the gold standard measure closely approaches those skills specifically targeted by the intervention, the current dissertation shows that using planned missing data can be a valuable approach to diminish the influence of error in the outcome measure and could improve model fit. Making use of planned missing data designs thereby offers an easily accessible way to improve the quality of future experimental studies.

References

Garnier-Villarreal, M., Rhemtulla, M., & Little, T. (2014). Two-method planned missing designs for longitudinal research. *International Journal of Behavioral Development*, 411-422.

Graham, J. W., Taylor, B. J., Olchowski, A. E., & Cumsille, P. E. (2006). Planned missing data designs in psychological research. *Psychological Methods*, 323-343.

MacCallum, R., Browne, M., & Sugawara, H. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods*, 130-149.

Maxwell, S., Lau, M., & Howard, G. (2015). Is psychology suffering from a replication crisis? What does 'failure to replicate' really mean? . *American Psychologist*, 487-498.

Merkelbach, I., Plak, R., & Rippe, R. (2018). Reproducibility of young learners' susceptibility to the learning context. *Learning and Individual Differences,* 167-175.

Van der Kooy-Hofland, V. A., van der Kooy, J., Bus, A. G., van IJzendoorn, M. H., & Bonsel, G. J. (2012). Differential susceptibility to early literacy intervention in children with mild perinatal adversities: Short- and long-term effects of a randomized controlled trial. *Journal of Educational Psychology*, 337-349.

Van Teijlingen, E., & Hundley, V. (2001). The importance of pilot studies. *Sociology at Surrey*.

Samenvatting

Tot voor kort werden milde perinatale problemen slechts gezien als risicofactoren voor een breed scala aan negatieve uitkomsten. Kinderen die iets te vroeg geboren werden (tussen de 34^{e} en 38^{e} week van de zwangerschap) of bij geboorte klein waren voor de duur van de zwangerschap, hebben immers meer kans op zowel medische, gedragsmatige, als cognitieve problematiek op latere leeftijd. Uit recent onderzoek blijkt echter dat milde perinatale problematiek niet altijd negatieve gevolgen hoeft te hebben; kinderen die dergelijke problemen hebben gehad, lijken niet kwetsbaar voor negatieve ervaringen, maar gevoelig voor de kwaliteit van hun omgeving, ook als deze omgeving positief is. In negatieve omgevingen zijn de uitkomsten van deze kinderen dus over het algemeen negatief, terwijl in positieve omgevingen juist positieve uitkomsten zijn te verwachten. Deze uitkomsten zijn gemiddeld zelfs positiever dan de uitkomsten van leeftijdsgenoten die geen perinatale problematiek hebben meegemaakt: een fenomeen dat bekend staat onder de naam *differentiële ontvankelijkheid*.

In eerder kleinschalig onderzoek werd aangetoond dat kinderen met milde perinatale problematiek op 5-jarige leeftijd ontvankelijker zijn voor een digitale leeromgeving die hun vroege geletterdheid stimuleert dan hun leeftijdsgenoten. Wanneer zij met de interventie hadden gewerkt, deden deze kinderen het zowel direct na de interventie als een jaar later beter dan hun klasgenoten, terwijl zij zonder interventie juist minder goed presteerden. Naar aanleiding van deze studie werd in het project '*Wat Werkt voor Wie'* onderzocht of deze resultaten gerepliceerd konden worden en of resultaten vergelijkbaar zijn voor iets te vroeg geboren kinderen en kinderen die relatief licht waren bij hun geboorte. Daarnaast werd onderzocht of vergelijkbare resultaten worden bereikt wanneer een andere schoolse vaardigheid wordt getraind (i.e. rekenvaardigheden), en werden mogelijke onderliggende mechanismen en verklaringen voor de gevonden resultaten bekeken.

Replicatie

In hoofdstuk 2 wordt een replicatie besproken van het eerder genoemde, kleinschalige onderzoek. In het huidige onderzoek worden iets te vroeg geboren kinderen en kinderen die te licht waren voor de duur van de zwangerschap echter als aparte groepen behandeld, en niet samengevoegd tot een overkoepelende groep met milde perinatale problematiek.

In een grootschalig experiment op 187 scholen werden 5-jarige kinderen willekeurig toegewezen aan een van drie digitale programma's: *Samen Slim* (een digitaal rekenprogramma), *Levende Letters* (een digitaal programma dat foneembewustzijn en alfabetische kennis stimuleert), of *Levende Boeken* (een digitaal controleprogramma dat bestaat uit geanimeerde prentenboeken). Gedurende twee maanden, werkten kinderen eens per week geheel zelfstandig met deze programma's in de klas.

In de eerste studie wordt het contrast tussen *Levende Letters* en *Levende Boeken* onderzocht. Na de interventie namen leerkrachten digitaal een test af die het foneembewustzijn en de alfabetische kennis van leerlingen in kaart te brengt. Uit het vergelijken van de testresultaten bleek dat kinderen zonder perinatale problematiek niet profiteerden van het werken met *Levende Letters*: er was geen verschil in resultaat te zien tussen beide programma's. Hetzelfde gold voor kinderen die te licht waren voor de duur van de zwangerschap. Voor kinderen die iets te vroeg geboren waren, werd echter wel een differentieel effect van *Levende Letters* gevonden. Wanneer deze kinderen hadden gewerkt met het controleprogramma, deden zij het op de eindmeting minder goed dan hun leeftijdsgenoten, terwijl zij het na het werken met *Levende Letters* juist beter deden. Ook in de huidige studie waren effecten na een jaar nog steeds zichtbaar. De grootte van de gevonden effecten was echter wel een stuk kleiner dan die van de effecten die werden gevonden in de eerdere, kleinschalige studie.

Verschillen in effectgroottes

Een mogelijke verklaring voor de gevonden verschillen in effectgroottes tussen de huidige studie en de eerdere, kleinschalige studie, is verminderde betrouwbaarheid van de eindmeting in het huidige onderzoek. In de kleinschalige studie bestond de eindmeting uit meer items en werd deze daarnaast afgenomen door de onderzoekers, terwijl dit in de huidige studie door de leerkracht werd gedaan. Daarom werd onderzocht of resultaten in het huidige onderzoek gerepliceerd konden worden wanneer de betrouwbaarheid van de eindmeting werd vergroot, en of dit zou leiden tot het vergroten van gevonden effecten. Om de betrouwbaarheid van de eindmeting te vergroten, werd een *planned missing data*-aanpak ingezet (Hoofdstuk 3). In een random gekozen subgroep van kinderen werd een additionele, betrouwbare test afgenomen door onderzoekers, ook wel de gouden standaard genoemd. De gedeelde variantie tussen de oorspronkelijke test (bij de gehele groep afgenomen door de leerkracht) en de gouden standaard, is een valide en betrouwbare maat die het mogelijk maakt een betrouwbaarder model te toetsen.

Wanneer er gebruik werd gemaakt van deze zogenaamde *planned missing data*-aanpak, konden resultaten worden gerepliceerd. Replicatie was echter alleen mogelijk wanneer de gouden standaard inhoudelijk dicht bij het getrainde construct lag, in dit geval alfabetische kennis en foneembewustzijn. Wanneer taken aan de gouden standaard werden toegevoegd die verder van dit construct aflagen, zoals woordenschat, konden resultaten niet langer gerepliceerd worden. Hoewel resultaten na het vergroten van de betrouwbaarheid van de eindmeting dus overeind bleven, bleven de effectgroottes nagenoeg gelijk. Gebrek aan betrouwbaarheid van de eindmeting lijkt dus geen verklaring te bieden voor de verschillen tussen effectgroottes gevonden in de huidige studie en die gevonden in het kleinschalige onderzoek dat voorafging aan de huidige studie.

Een andere mogelijke verklaring voor de discrepantie in effectgroottes kan liggen in de implementatie van de programma's door en de houding van de leerkracht. In deze studie waren leerkrachten verantwoordelijk voor het uitvoeren van de interventie, terwijl deze verantwoordelijkheid eerder bij de onderzoekers lag. Hoewel leerlingen zelfstandig werkten, bepaalden leerkrachten wanneer en onder welke omstandigheden kinderen met de programma's werkten. Hoewel we niet direct konden testen wat de invloed van de leerkracht was, konden we wel aannemelijk maken dat de digitale vaardigheden en opvattingen van leerkrachten van belang zijn voor het succes van digitale leerinterventies (Hoofdstuk 4). Leerkrachten met minder digitale vaardigheden en/of negatieve opvattingen en verwachtingen leken minder (effectief) gebruik te maken van digitaal materiaal in de klas. Daarnaast voorspelden beperkte digitale vaardigheden voor negatievere opvattingen over en verwachtingen van digitaal materiaal.

Rekenvaardigheid

Differentiële ontvankelijkheid voor het digitale letterprogramma *Levende Letters* in iets te vroeg geboren kinderen kon dus herhaaldelijk gerepliceerd worden, hoewel effectgroottes verschilden. Eerder onderzoek laat echter zien dat bij iets te vroeg geboren kinderen met name problematiek gerelateerd aan rekenen op de voorgrond staat. Er werd daarom onderzocht of iets te vroeg geboren kinderen ook meer ontvankelijk waren voor de effecten van een digitaal rekenprogramma, *Samen Slim* (Hoofdstuk 5). *Samen Slim* heeft een vergelijkbare aanpak en opbouw als *Levende Letters*: beide programma's zijn erg gestructureerd, bieden veel uitleg, en constante en adaptieve feedback. *Samen Slim* richt zich echter op het trainen van de beginnende rekenvaardigheden.

In deze studie werden kinderen random toegewezen aan *Samen Slim* of het controleprogramma (*Levende Boeken*). Opnieuw werden er geen differentiële effecten gevonden voor kinderen zonder perinatale problematiek of voor kinderen die te licht waren voor de duur van de zwangerschap. In de groep van iets te vroeg geboren kinderen werden wel opnieuw differentiële effecten gevonden. Wanneer deze kinderen werkten met het controle programma deden zij het minder goed dan hun leeftijdsgenootjes, terwijl zij het juist beter deden wanneer zij met *Samen Slim* hadden gewerkt. 5

Stressreactiviteit

Als mogelijk onderliggend mechanisme voor de bevindingen werd toegenomen stressreactiviteit in iets te vroeg geboren kinderen verondersteld. Risicofactoren voor vroeggeboorte, zijn tevens factoren die bijdragen aan verhoogde stressreactiviteit, het lichamelijk sterker reageren op stressoren van buitenaf. Iets te vroeg geboren kinderen werden verondersteld over het algemeen stressgevoeliger te zijn, en daardoor meer baat te hebben bij gestructureerde en duidelijke leeromgevingen, zoals de digitale programma's die in het '*Wat Werkt voor Wie'*-project zijn getoetst. Hoewel een mediatie door stressgevoeligheid tussen iets te vroeg geboren zijn en vergrootte ontvankelijkheid voor gestructureerde leeromgevingen niet kon worden getoetst, kon wel worden aangetoond dat iets te vroeg geboren kinderen inderdaad een verhoogde stress reactiviteit lieten zien (Hoofdstuk 6). Iets te vroeg geboren kinderen lieten in anticipatie op een onbekende taak, afgenomen door een onbekende volwassenen (i.e. een van de onderzoekers) een verhoging in stresshormonen zien. Zodra de testafname, die uiterst gestructureerd verliep en tijdens welke kinderen veel feedback en positieve bevestiging kregen, echter was begonnen, nam de concentratie van dit stresshormoon in deze groep kinderen sterk af. Hun leeftijdsgenoten lieten dit patroon niet zien, zowel voor als tijdens het testen bleef het niveau van stress in deze groep kinderen relatief stabiel.

Aanbevelingen

Digitale educatieve programma's lijken veelbelovende resultaten te kunnen opleveren, hoewel soms alleen voor gevoelige subgroepen van kinderen. Wanneer de effectiviteit van dergelijke interventies wordt geëvalueerd, moeten daarom niet alleen hoofdeffecten in ogenschouw worden genomen. Toekomstig onderzoek zal verder moeten verduidelijken welke interventiecomponenten precies werkzaam zijn voor kwetsbare groepen zoals iets te vroeg geboren kinderen, en welke mechanismen hieraan ten grondslag liggen. Voor het behalen van successen met digitale interventies, lijkt daarnaast niet alleen de match tussen leerling en interventie essentieel, maar ook de kunde en overtuigingen van de leerkracht wat betreft digitaal leermateriaal. Training van docenten in het gebruik en de mogelijkheden van digitaal leermateriaal, kan daarom essentieel zijn voor succesvolle implementatie van dergelijke interventies.

Dankwoord

Mijn dank gaat uit naar Stichting Kennisnet voor het toekennen van subsidie aan Adriana Bus teneinde dit onderzoek mogelijk te maken.

Ook bedank ik graag alle deelnemende scholen, leerkrachten, ouders en leerlingen, en alle studenten die aan *Wat Werkt voor Wie* hebben meegewerkt. Zonder jullie was dit project nooit mogelijk geweest. Dank ook voor Esther en Maarten van Bereslim voor het beschikbaar stellen van de digitale interventieprogramma's en voor de fijne samenwerking.

Eric, bedankt voor je steun, praktische hulp en meedenkende houding.

Chris, bedankt dat je het promotorschap op je hebt willen nemen, voor je vriendelijkheid, en voor je constructieve feedback.

Lieve Ralph, je bent de beste co-promotor die ik me had kunnen wensen. Zonder jou was dit proefschrift er zeker niet geweest. Ik heb altijd het gevoel gehad met alles bij je terecht te kunnen, bedankt voor je constante steun. Bedankt ook voor je positieve houding, opbeurende woorden, statistische hulp, en voor al onze fijne gesprekken en leerzame, uitdagende, motiverende discussies.

Lieve collega's, ook jullie bedankt voor het creëren van een veilige en gezellige werksfeer, ik ben trots op ons team en ben blij dat ik met jullie heb mogen werken! Marga en Marianne, bedankt voor jullie positiviteit en gezelligheid. Thijs, Heleen, en Elise, wat hebben we veel lol gehad! Bedankt voor jullie vrolijkheid, de weekendjes weg, de borrels, en de etentjes, die we er gelukkig, hoewel de meeste van jullie niet meer aan het werk zijn in Leiden, in weten te houden.

Lieve Rachel, bedankt voor het zijn van mijn kamergenootje tijdens de eerste jaren van mijn traject (ik mis je nog steeds heel erg, maar eet nu wel een stuk minder snoep) en voor het zijn van een van mijn beste vriendinnen. Ook zonder jou was dit proefschrift er zeker niet geweest. Wat hebben we veel meegemaakt samen, zowel diepte- als hoogtepunten (misschien moet dat boek er toch maar komen?). Hoe de situatie ook was of is, we konden en kunnen altijd bij elkaar terecht. Ik ben heel blij dat ik aanwezig heb mogen zijn bij de belangrijke momenten in jouw leven en ben er ontzettend trots op dat jij vandaag naast me wil staan als mijn paranimf.

Lieve Dylan, we kennen elkaar al sinds we twaalf jaar zijn en je bent sindsdien altijd enorm belangrijk voor me geweest. Ik ben heel blij dat je vandaag naast me wil staan als mijn paranimf om me te steunen en de boel te relativeren. Ik kijk ernaar uit om straks met je te proosten op de goede afloop.

Lieve vrienden, wat een geluk dat ik jullie allemaal om me heen mag hebben. Jullie maken mijn leven zoveel leuker. Mijn grootste dank gaat uit naar Tim, Mark (van wie ik weet dat hij dit dankwoord, en dan ook alleen het dankwoord, met extreme interesse zal lezen), Floor, Sanne, Chiel, Danko, Daniël, en Rianne (die de liefste is).

Lieve papa en mama, bedankt voor jullie onvoorwaardelijke liefde en voor de deur die altijd open staat. Lieve Suzan, ik ben blij dat jij mijn zus bent. Lieve opa, bedankt voor je voortdurende interesse en voor het van jongs af aan stimuleren van mijn nieuwsgierigheid.

Lieve Joost, bedankt voor je voortdurende steun.