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

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.

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.

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.