Protective teaching mechanisms in case of mild perinatal adversity

Merkelbach, I.

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

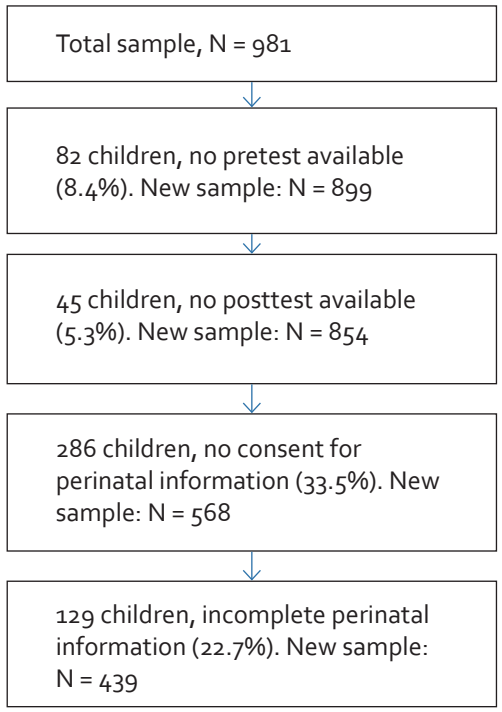

Figure 1. Participant selection scheme

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

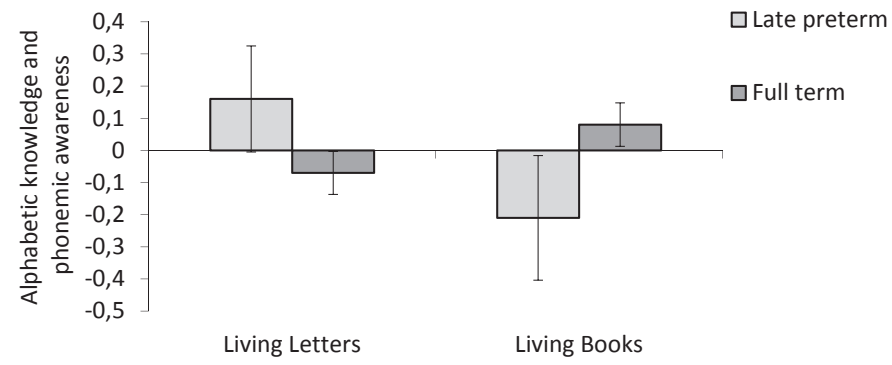

Figure 2. Interaction between late preterm and condition with alphabetic knowledge and phonemic awareness (short term) as outcome measure

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.

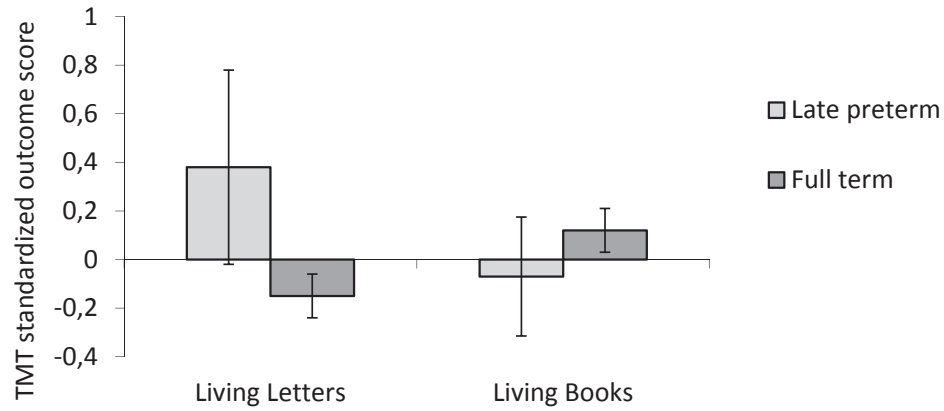

Figure 3. Interaction between late preterm and condition with word recognition (long term) as outcome measure

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