Can we predict non-response in developmental tasks? Assessing the longitudinal relation between toddlers' non-response and early academic skills

Spit, S.; Mulder, H.; van Houdt, C.; Verhagen, J.

DOI
10.1002/icd.2376

Publication date
2023

Document Version
Final published version

Published in
Infant and Child Development

Received: 8 June 2022 | Revised: 29 August 2022 | Accepted: 7 September 2022

DOI: 10.1002/icd.2376

REPORT

WILEY

Can we predict non-response in developmental tasks? Assessing the longitudinal relation between toddlers' non-response and early academic skills

Sybren Spit[1,2] | Hanna Mulder[3] | Carolien van Houdt[3] | Josje Verhagen[2]

[1]Knowledge and Strategy Department, Ministry of Education, Culture and Science, The Hague, The Netherlands

[2]Amsterdam Center for Language and Communication, University of Amsterdam, Amsterdam, The Netherlands

[3]Department of Development & Education of Youth in Diverse Societies, Utrecht University, Utrecht, The Netherlands

Correspondence
Sybren Spit, Ministry of Education, Culture and Science, Rijnstraat 50, 2512 XP, Den Haag, The Netherlands.
Email: s.b.spit@minocw.nl

Funding information
Nederlandse Organisatie voor Wetenschappelijk Onderzoek, Grant/Award Numbers: 411-20-442, 411-20-452

Abstract

To date, virtually no studies have examined toddlers' non-response in developmental tasks. This study investigates data from 3667 toddlers to address (1) whether two aspects of non-response (completion and engagement) are separable, (2) how stable these aspects are from ages two to three, (3) how non-response relates to background characteristics, and (4) whether non-response at ages two and three predicts early academic skills at age six. Structural equation modelling shows that completion and engagement are separable constructs, relatively stable across age, and related to several background characteristics. Especially engagement predicts later academic performance. Results show that non-response in behavioural tasks in toddlers is not random, increasing the likelihood of sampling bias and lack of generalizability in developmental studies.

KEYWORDS
early academic skills, non-response, representativity, task completion, task engagement, toddlers

1 | INTRODUCTION

Representativity of study samples has been a topic of debate for decades, with one of the main conclusions being that participant samples are often biased (Kruskal & Mosteller, 1979; Kukull & Ganguli, 2012). Sample bias is problematic, as it makes it difficult to determine whether research outcomes generalize to the population of interest, which, in turn, can hamper theory construction (Andringa & Godfroid, 2020). Recently, this problem has received renewed attention, as it surfaced that many samples in psychological research are WEIRD (i.e., Western, Educated, Industrialized, Rich, and Democratic). In particular university students tend to be over-represented in the field of psychology, making it difficult to assess whether theories hold for the population at large (Arnett, 2008; Henrich et al., 2010; Nielsen et al., 2017).

An additional and related problem arises in research with young children, where a subset of the children investigated typically does not comply with the tasks given to them. They do not complete the tasks, deliberately give wrong answers, or do not respond at all. Since children who do not comply with a task are usually excluded from analyses, conclusions in the field are based on children who are able to complete tasks at a young age. However, earlier research on young children has shown that non-responders differ from responders in major ways (Bathurst & Gottfried, 1987; Bell & Slater, 2002), which does not only make samples unrepresentative but also likely leads to biased conclusions.

Most of the previous studies investigating non-response in young children have concentrated on infants (Slaughter & Suddendorf, 2007; Stets et al., 2012; van der Velde & Junge, 2020). Hence, very little is known about non-response in slightly older children, who are mature enough to carry out simple instructions in behavioural tasks, but may have problems understanding instructions and sustain their attention. In the current study, we aimed to fill this gap by investigating non-response on behavioural tasks in two- and three-year-old children. These children performed a large battery of tasks measuring language and executive functions at toddler age (ages two and three). Their scholastic achievements were tracked during primary school, as part of a large-scale longitudinal study on children's linguistic, cognitive, and (pre-)academic abilities (Mulder et al., 2014, 2017). The overall aims of our study were to investigate (1) which factors related to non-response in toddlers and (2) whether children's non-response in toddlerhood would predict early academic skills 3 to 4 years later.

2 | NON-RESPONSE IN DEVELOPMENTAL STUDIES

Non-response is common in studies in which young children perform behavioural, visual or neuro-linguistic tasks (Frank et al., 2017; ManyBabies Consortium, 2020; van der Velde & Junge, 2020). Reasons for uncooperative or unresponsive behaviours are diverse and may relate to children's inability to sit still, follow up or understand instructions, or problems with selectively and continuously attending to the task. In infant studies, non-response is often referred to as fussiness, which represents a broad concept that covers not responding to a task, low attention, crying, as well as many other behaviours that do not enable optimal gathering of data. This problem is not trivial: it has been estimated that, in visual paradigms, an average of 14% of the participating infants do not complete the experimental procedure, due to fussiness (Slaughter & Suddendorf, 2007).

Several studies have examined which factors are associated with non-response (Slaughter & Suddendorf, 2007; Stets et al., 2012). Although Slaughter and Suddendorf (2007) found no relation between differential exclusion rates across visual paradigm studies and the researchers' chances of finding effects, they did observe that younger children were more likely to become fussy during the experiment than older children. A meta-analysis by Stets et al. (2012) on ERP studies with infants showed, furthermore, that experimental conditions might affect whether drop out is likely, such that attrition rates were higher when children were exposed to visual stimuli as compared to auditory stimuli. Klein-Radukic and Zmyj (2015) found a relationship between infants' temperament and drop-out rates in looking time experiments, such that non-completers were more likely to fuss and cry in situations in which they were

restricted (e.g., in an infant seat, during caretaking), slower to recover from distress or excitement, more rapid in speed of approach and excitement towards certain stimuli and less likely to experience enjoyment when being held by their caregiver than completers. These authors also observed that drop-out at 6 months predicted drop-out at 9 months, suggesting that the characteristics implicated in non-response are relatively stable over time, at least over a three-month time span.

Further evidence for the idea that a multitude of factors relates to non-response comes from van der Velde and Junge (2020). These authors investigated data from a longitudinal study (i.e., the YOUth study, Onland-Moret et al., 2020), in which over 3000 children participated in two EEG-experiments over multiple sessions between the ages of 5 months and 6 years. Using data from a subset of children who participated at 5, 10 months, and 3 years, the authors found that both experiment-related factors (e.g., time of the day, research assistant, season of testing) and child-related factors (e.g., gender, age) predicted data loss. More specifically, early morning testing and testing in spring or summer was related to lower data loss. In addition, the oldest age group (3-year-olds) had much less data loss than the 5- and 10-month-old children, and boys had slightly less data loss than girls. Like Klein-Radukic and Zmyj (2015), moreover, they found that data loss was stable over time and could be predicted over multiple sessions: children who had high data loss in one session, were more likely to have data loss at the later session(s) as well.

Taken together, there is accumulating evidence that non-response in infants is not random, such that non-responding infants share certain characteristics. This raises the question whether similar tendencies can be observed in slightly older children of toddler age (i.e., aged 2 and 3 years), who typically perform tasks that are more demanding in terms of understanding and responding to instructions. Although many studies in the field report attrition analyses examining the effects of data loss (Willoughby et al., 2010, 2012, Willoughby, Wirth, et al., 2012), to the best of our knowledge, virtually no studies have targeted drop-out from experimental tasks in two- and 3-year-olds as their main aim. Yet, non-response in toddlers is common: in earlier work in which 2- and 3-year-olds were administered batteries of language and executive function tasks, percentages of children not completing any of the tasks were 8% or 9% (Hughes & Ensor, 2005; Poulin-Dubois et al., 2011; Willoughby et al., 2012), 13% (Kuhn et al., 2014), or even as high as 26% (Scott et al., 2012).

An exception to this is a study by Bathurst and Gottfried (1987), who did concentrate on non-response in toddlers. Specifically, these authors studied 22 children who had been administered a task battery at three time points during their preschool period (at 30, 36, and 42 months) but had not responded at two successive attempts of administration at one of these time points. Comparing these children to 108 children who did perform the tasks at each of the three time points, Bathurst and Gottfried found that, relative to the latter group, the group of 'untestable' children were less developmentally advanced across a wide range of abilities at all ages from 12 through 72 months. Specifically, untestable children were rated more poorly on a variety of intellectual and developmental skills, academic performance, and social functioning, as based on parental reports. In addition, untestable children were found to be less adaptive to testing situations, temperamentally less sociable, and to have a higher frequency of problem behaviours on adjustment and behaviour checklists, as rated by their parents. However, as sample size was relatively small in this study ($N = 22$ untestable children), it is not clear whether these results can be generalized. Also, the study leaves unclear whether children who were classified as being untestable did not respond to just one task, more than one task, or even all of the tasks administered at a given moment. Therefore, it is not immediately clear how its results should be interpreted.

A further, more general issue that makes it difficult to interpret the results of Bathurst and Gottfried (1987) is that the study did not specify how many items of a task children should fail to respond to in order to be classified as 'untestable'. Across studies, different cut-offs have been applied such that, in some studies, children were excluded if they did not complete a single item of a task, whereas in others, they were excluded if they completed less than half of the items or even all items of a task. In Mulder et al. (2014, 2017), for example, scores of children were only analysed if children had responded to at least half of the items of a task. In other studies, using similar language and executive function tasks, however, less stringent criteria were applied or no information was provided as to whether such 'partial responders' were excluded (Espy et al., 2011; Hughes & Ensor, 2005). Although such decisions can be

well-motivated, for example to avoid drawing conclusions on the basis of very few responses per child, having different cut-offs per study clearly compromises the comparability of results across studies, as well as the generalizability of study results.

3 | THE CURRENT STUDY

As discussed above, non-response in young children is a rather common phenomenon that does not occur completely randomly (Frank et al., 2017; ManyBabies Consortium, 2020; van der Velde & Junge, 2020). Yet, research on the topic is scant, especially on children beyond infant age. The one study available that investigated non-response in toddlers showed that non-responders differed from responders in many socio-cognitive domains (Bathurst & Gottfried, 1987), but due to its small sample size, it is currently unknown whether non-responders form a special group. In the present study, we investigate a large group of children who were administered in a series of language and executive functions tasks at ages 2 and 3 years, and tests measuring early academic skills in grade 1. Furthermore, a set of background variables was recorded, including children's home language and their parents' education level. This data allowed us to examine the characteristics of young children who do not respond optimally in developmental tasks.

Using this longitudinal dataset, we tested whether non-response is stable over time, and whether there is a relation between non-response at toddler age and early academic skills when children are older. Early academic performance is usually highly correlated with parental education, home language background, and executive functioning (Mulder et al., 2017). That is, children who perform well on tasks assessing early academic skills typically have well-developed abilities in language, good memory, and attention skills, and generally come from advantageous family backgrounds. As discussed above, performing a task at all and being engaged with it entails a set of skills in young children that are, themselves, very closely related to language and executive functioning, such as sitting still, listening to instructions, following up on and remembering instructions, and inhibiting unwanted behaviour (such as walking away). If tasks or task batteries are long, they also require persistence and sustained attention from the child. Hence, it is conceivable that task completion in itself is predictive of children's later academic skills, in line with earlier findings showing that non-response can be predictive of later task behaviour (Bathurst & Gottfried, 1987; van der Velde & Junge, 2020).

Because the study sample that we investigated was older than the infants investigated in many previous studies, we differentiated between non-completion and non-engagement. Non-completion refers to situations where children did not respond to one or more items of a task, for example, by remaining silent or indicating they do not want to answer. Non-completion thus leads to missing data. Non-engagement refers to situations where children are not engaged with the items of a task: they may respond to them, but express verbally or nonverbally that they are not interested. Multiple reasons for low levels of engagement may be at stake, including feelings of anxiety, shyness, but also boredom or concentration problems. Whereas it is difficult to assess why infants do not respond to task stimuli, studying older children allows to make a distinction between cases where children do not comply with a task in the sense that they are not engaged with it, but still respond to its items (which might make their responses unreliable), and cases where children do comply, in the sense that they are engaged with the task, but do not respond (for example, if a task is too difficult). It is currently unclear whether non-completion and non-engagement are strongly related or even overlapping constructs, or whether they can be treated as separate theoretical constructs.

All in all, we addressed a number of related research questions in our study. First, we focused on the issue of non-response itself by assessing whether task completion and engagement represented two separate constructs in our data or whether they should best be seen as a single construct. Second, we asked whether non-response at ages 2 and 3 years was predicted by factors like age, gender, parental education, home language background or test location. Do these factors predict whether children completed or engaged with behavioural tasks? Finally, we asked whether non-response was stable over time - from 2 to 3 years, and whether it predicted early academic skills in

grade 1. To address these questions, we investigated the relations between children's background characteristics, task completion and engagement at ages two and three, and early academic skills at age six through structural equation modelling (SEM), in a large sample of 3667 children in the Netherlands, which will be described in more detail below.

By answering these questions, we hoped to increase our understanding of the impact that non-response might have on a study sample and the ways in which non-response would compromise the generalizability of obtained results. Ideally, information about non-response is considered when designing new experiments, in order to minimize bias a priori. In addition, it is important to establish which factors we could include in our statistical models to control for the portion of non-response that is inevitable in work with young children. Recent research on causal inference suggests we should be careful when including control variables in statistical analyses, because they might amplify bias in particular cases or even lead to incorrect results (Hernán & Robins, 2020; Imbens & Rubin, 2015; Pearl, 2009; Pearl et al., 2016). More insight into the factors relating to non-response might help us minimize non-response in future studies both before setting them up, as well as after carrying them out.

4 | METHOD

4.1 | Participants

Data were analysed from children who participated in the pre-COOL and COOL studies which took place in the Netherlands. In these longitudinal studies, a large cohort of preschool and primary school children was followed to gain insight into their cognitive, linguistic, and (pre-)academic development (Mulder et al., 2014, 2017; Verhagen et al., 2017, 2019). Data from these studies has been used in policy reports concerning equality in young children (Zumbuehl & Dillingh, 2020) and to develop a metric that is used by the Dutch government to allocate school funds that should help decrease inequality in primary education (Posthumus et al., 2019). The full cohort study comprised different waves, with various tasks being administered at these waves during children's preschool and primary school trajectories. The first of these waves took place when children were 2 years old and the final one at the end of primary school, when children were 11 years old. Across these different waves, data were collected from 3667 children, but not all children were tested at each wave during their (pre)school trajectory.

In this paper, we report on data collected at three waves of the pre-COOL and COOL studies: when children were 2 years old, when they were 3 years old, and at the end of grade 1, when they were 6 years old. The first two waves were included because we were interested in non-response at toddler age and because many children were tested at these waves (i.e., around 2500 children). Data from ages four and five were not analysed because much fewer children were tested at these ages (i.e, around 750 children), and also because most of the tasks at these ages were very different from those at ages two and three.

Information about parental education level and linguistic status at home was obtained through parent questionnaires. Parental education was assessed on a four-point scale with (1) 'primary school', (2) 'lower vocational training', (3) 'secondary school and/or vocational training', and (4) 'higher education (i.e., college or university degree)' as its scale points, and averaged for both parents. When questionnaires were not filled out, school registry information on parental education was used where available. Children's home language background was assessed through a question asking whether children were only exposed to Dutch, or to (an)other language(s) in addition to or instead of Dutch at home. In case information on home language background was missing, research assistants obtained information about children's home language situation from parents or teachers.

The sample ($N = 3667$) included 1765 girls and 1848 boys (for 54 children, information about gender was missing). At test wave 1, mean age was 28 months ($SD = 3$, min-max = 20–37). At test wave 2, mean age was 42 months ($SD = 3$, min-max = 34–52). In grade 1, mean age was 85 months ($SD = 5$, min-max = 68–108). Information about

parental education was available for 2502 children (68%), with a mean score of 3.14 ($SD = 0.81$, min-max = 1–4). Most children were from monolingual Dutch families ($N = 2012/55\%$). The remaining children were from families where other languages were spoken next to, or instead of Dutch ($N = 866/23.5\%$), or data about their linguistic background was missing ($N = 789/21.5\%$). Most children were tested at their daycare ($N = 2398/65.5\%$), but many others were tested at home ($N = 993/27\%$). For the remaining children we either did not have information about location, or they were not tested at ages two or three. At each wave, about 30 assessors were involved. These assessors were from various geographical areas in the Netherlands, often from areas that were close to where the children lived. About a third of the assessors ($N = 11$) were involved in both waves. Thus, the majority of assessors were only involved in wave 1 or 2. Assessors in grade 1 were primary school teachers, so there was no overlap in assessors between waves 1 and 2 versus grade 1.

4.2 | Materials

Children were administered a series of tasks measuring executive functioning and language skills (see also: Mulder et al., 2014, 2017; Verhagen et al., 2017, 2019). In the analyses presented here, we only included tasks that were administered at ages two and three. We therefore excluded a memory for location task that was only administered at age two, and an inhibition task and two early numeracy tasks that were only administered at age three. Two delay of gratification tasks administered at ages two and three were excluded, since not responding was the targeted behaviour on this task (i.e., suppressing the dominant response to reach for a reward). A working memory task, administered at both ages, was not included in our analyses either, because children almost always gave responses to all items at age three and completion rates for this task were thus at ceiling (see our supplementary materials for descriptives). In grade 1, a widely-used test for language, literacy, and mathematical skills from the Dutch National Institute for Educational Measurement (Cito) was administered (Engzell et al., 2021; Vlug, 1997). We describe each of these tasks in more detail below.

All tasks measuring executive functioning and language skills were administered at both ages two and three. To avoid floor or ceiling effects, the tasks had been constructed such that a subset of the items was held constant from ages two to three and more difficult items replaced easier items at age three. This allowed to assess the difficulty of all items and scale them on an underlying latent ability scale through Item Response Theory modelling (Grimm et al., 2013; McArdle et al., 2009) in studies using this dataset, which looked at children's development of language and executive functioning (e.g., Verhagen et al., 2019). Crucially, we do not report on children's performance (i.e., accuracy) on the tasks in the current paper, but rather, on the degree to which children completed the items of a task as well as on how engaged they were when performing the task. Task completion was operationalized as the number of items per task completed by the child. Note that because of partly different items for the tasks at ages two and three, completion rates cannot be directly compared across ages. Task engagement was based on a rating scale filled out by the assessors. Specifically, assessors rated each child's level of engagement immediately after each task had been completed or aborted, using a seven-point scale with 1 ('not engaged at all'), 4 ('moderately engaged'), and 7 ('very much engaged') as its scale points. Engagement scores were weakly to moderately correlated with accuracy scores for the tasks we report on at both age two ($0.209 < r < 0.568$) and age three ($0.037 < r < 0.349$). These weak to moderate correlations signal that assessors' ratings of engagement were at least in part independent of children's level of performance on the tasks. We did not calculate such correlations between accuracy and completion scores, because these depend on one another: Participants who complete fewer items are more likely to have lower accuracy scores than participants who complete more items, or more extreme ones, in the case of proportions (a participant who only completed a single item either has an accuracy of 0 or 100, if we only look at completed items; and only a possible accuracy of 0 or 1/total number of items, if we look at all possible items).

4.2.1 | Selective attention

A visual search task was used to assess selective attention (Mulder et al., 2014, 2017). In this task, which was administered on a laptop, children were encouraged to search for targets (elephants) that were surrounded by distractors (bears and donkeys) that were similar in colour and size. Children were allowed to search for the elephants for 40 s. There were three task items, which were preceded by three training items in which children were instructed to point at the elephants only and do so as quickly as they could. At age two, the stimuli were presented in a 6 × 8 grid which showed 8 targets and 40 distractors on all three test items. At age three, the first two items were the same as at age two. The last item at this age was slightly more difficult, showing a 9 × 8 grid with 8 targets and 64 distractors. A completion score was calculated as the number of items for which children obtained a score. Children who did not find any targets in an item, did not obtain a score and were classified as non-responders. Likewise, when a child did not look at the screen at all during the 40 s search time, that item was set to missing for that child.

4.2.2 | Verbal short-term memory

A nonword repetition task was used to assess verbal short-term memory (Verhagen et al., 2017, 2019). In this task, children were instructed to repeat novel words. Specifically, they watched short video clips in which a novel object appeared, while they heard pre-recorded sentences labelling the novel object with its non-word name, as follows: 'Look a hiemup! Say hiemup'. The task started with two training items, followed by 12 task items. Repetition attempts were classified by the assessors as 'correct', 'incorrect', or 'uncodable' immediately after a child's response. A completion score on this task represented the number of repetition attempts children made.

4.2.3 | Phoneme identification

Children's knowledge of Dutch phonemes was assessed using a phoneme identification task (Gerrits, 2003; Kuijpers, 1996). In this task, children were shown two pictures on a laptop screen, of which the labels formed a minimal pair in Dutch (i.e., pear and bear), and then heard the label of one of these two pictures (e.g., bear). The child was then asked to point to the picture that corresponded to the auditorily presented word. The rationale behind the task is that children can only distinguish between the two words if they can distinguish between the phonemes that differ between the two words of the minimal pair (i.e., /p/ and /b/). Children were presented with two training items and 12 task items. Responses were classified as 'correct', 'incorrect', or 'no response'. The completion score on this task was calculated as the number of responses a participant gave, that is, the number of items for which children selected a picture.

4.2.4 | Vocabulary

Vocabulary was assessed using a shortened version of the Dutch Peabody Picture Vocabulary Test (PPVT-III-NL, Dunn & Dunn, 2005). In this task, children were instructed to select one out of four pictures after an orally presented word. Two adaptations to the original task were made (see Verhagen et al., 2019). First, for administrative reasons, the task was carried out on a laptop, rather than as a pencil and paper test. Second, items that were found to be either too easy (accuracy >70%) or too difficult (accuracy <30%) in pilot studies with children of the relevant age ranges were removed. This was done in order to avoid administering items that did not differentiate well across children. A fixed number of 24 items was administered to all children. A completion score was computed by summing the number of responses a participant gave, that is, the number of items for which children selected a picture.

4.2.5 | Sentence comprehension

Children's sentence comprehension and early grammatical skills were measured using a picture selection task that assessed children's knowledge of grammatical structures (Johnson et al., 2005; Sekerina et al., 2004). During this task, children saw two pictures on a laptop screen, describing two events that differ in only one grammatical feature (i.e., gender, number, tense, word order). Children then heard a sentence describing one of these two pictures and were asked to point to the corresponding picture. For example, they would see a picture of a boy washing a pear ('He is washing a pear') and a picture of a girl washing a pear ('She is washing a pear'), and hear the sentence 'She is washing a pear'. The task contained three training items, followed by 12 task items. Responses were classified as 'correct', 'incorrect', or 'no response'. A completion score on this task represented the number of responses a participant gave, that is, the number of items for which children selected a picture.

4.2.6 | Working memory

Children's working memory was assessed using the Six Boxes task, adapted from Diamond et al. (1997). In this task, six identical boxes were placed in front of the child. The assessor hid a toy in each box while the child was watching. The child was then asked to search for the toys by lifting one box at a time, in six search trials. In between trials, children were distracted for 6 s by the assessor (i.e., asked to look away at the assessor's hand, while they counted from 1 to 6 out loud). The task for the child thus was to remember which boxes they had already emptied and which still contained a toy at each search trial, and hence, to update the information in their working memory at each trial. A completion score on this task represented the number of responses a participant gave, that is, the number of items for which children selected a box.

4.2.7 | Early academic skills

Children's early academic skills were assessed through the Cito test, which is a standardized national test that is conducted twice a year during children's school trajectories, as part of a pupil monitoring system (Engzell et al., 2021; Vlug, 1997). Here, we report on the test that is administered across multiple days at the end of grade 1 and that covers four domains: general mathematical skills, vocabulary, decoding skills, and reading comprehension. All subtests are carried out as paper and pencil tests. For mathematical skills, the teacher reads aloud exercises that test number knowledge, mental arithmetic, and basic geometry. Children write down their answers. For vocabulary, the teacher reads aloud a word, and children are asked to select which picture out of three pictures corresponds to this word. For decoding, children are presented with three lists of words of increasing complexity, of which they have to read as many as they can within 3 min. For reading comprehension, children read a text and answer questions alongside the text assessing their understanding of the text. The National Institute for Educational Measurement (Cito) in the Netherlands employs Item Response Theory to scale children's raw scores on different versions of the test on a latent ability scale, considering the difficulty level of the items. For each of the early academic skills measures, the score on this latent ability scale for that academic skill was used as the outcome measure.

4.3 | Procedure

At ages two and three, tasks were administered by trained research assistants in a quiet room at children's homes or daycare centres. Children assessed at their daycare centre were tested in a quiet room, to minimize distraction by other children and daycare staff. As for children assessed at home, we also attempted to minimize distraction as

much as possible. That is, when assistants phoned children's parents to make an appointment, they asked parents to refrain from intervening during the test session, and make sure that other children and pets would not be in the same room as the child during the test session. At visit, these requests were repeated, if necessary. The five tasks here reported on were intermixed with the tasks excluded from the current study that were described above ('see under Materials') and administered in a fixed order: phoneme identification, vocabulary, selective attention, verbal short-term memory, sentence comprehension. Teachers administered the Cito tasks at the end of grade 1 according to a standardized protocol.

4.4 | Analyses

We used SEM, through the R package *lavaan* (Rosseel, 2012). Our analysis consisted of two steps. First, we used confirmatory factor analysis (CFA) to test whether engagement and completion represented two separate factors at age two and at age three in our data. To determine how early academic skills should be incorporated in our model, we also assessed whether early academic skills could be represented as different manifest variables or as a single latent construct. These CFAs showed that completion and engagement were best represented as separate latent variables at both age two and age three, while early academic skills were best represented as four different manifest variables. In a second step, we ran a SEM with the latent variables completion, engagement, and the Cito manifest variables, which also included the background variables home language background, parental education, and age, to see how these background variables related to completion and engagement. In each of these steps, model fit was assessed using the following cut-off criteria: RMSEA <0.08, with a preferred upper limit of the 90% CI < 0.10, CFI > 0.90, SRMR < 0.08 (Kline, 2005). We did not use the chi-square index, because it is sensitive to sample size, and typically significant in large samples (Brown, 2015; Little, 2013). In all analyses, we used full information maximum likelihood estimation with robust (Huber-White) standard errors to handle missing data. Scripts and data that were used for our analyses can be found on our OSF page: https://osf.io/7cwej

5 | RESULTS

5.1 | Descriptives

Descriptive statistics for completion and engagement rates per task are presented in Table 1; descriptives for performance on the tests assessing early academic skills are given in Table 2. Recall that completion for each task refers to the number of items a child completed, and engagement refers to assessor ratings on a seven-point scale assessing children's engagement. As can be seen from Table 1, we also report the drop-out for each of the tasks. This is the number of children that did not complete at least half of the items of a given particular task (i.e., 2/3; 6/12; 12/24), which was used as an exclusion criterion in earlier studies using this test battery (Mulder et al., 2014, 2017; Verhagen et al., 2017). Here, these children were included, because we were interested in seeing which background factors would predict completion and engagement rates in the entire sample.

5.2 | Completion and engagement as two separate constructs?

In our CFA investigating whether completion and engagement at age two should be considered two separate latent variables or a single latent variable, we compared a model in which completion and engagement scores of all five tasks loaded on a single variable to a model in which engagement scores from all tasks loaded on one variable and completion scores for all tasks on another variable. This comparison showed that the two-factor model fitted our data better than

TABLE 1 Descriptives for completion and engagement per task at ages two and three years

	Engagement				Completion				
	M	SD	Min-max	N	M	SD	Min-max	N	Drop-out
Age 2									
Selective attention	5.29	1.96	1…7	2386	2.71	0.83	0…3	2386	216 (9.1%)
Verbal short- term memory	4.92	2.02	1…7	2336	8.74	4.42	0…12	2339	553 (23.6%)
Phoneme identification	5.75	1.64	1…7	2446	10.42	3.17	0…12	2449	228 (9.3%)
Vocabulary	5.54	1.70	1…7	2409	21.88	5.27	0…24	2409	147 (6.1%)
Sentence comprehension	4.87	1.99	1…7	2311	9.41	3.91	0…12	2311	407 (17.6%)
Working memory	5.97	1.50	1…7	2284	5.80	0.90	0…6	2256	59 (2.6%)
Age 3									
Selective attention	6.54	0.94	1…7	2697	2.98	0.26	0…3	2697	19 (0.7%)
Verbal short- term memory	6.14	1.27	1…7	2691	11.29	2.41	0…12	2691	122 (4.5%)
Phoneme identification	6.71	0.85	1…7	2709	11.81	1.30	0…12	2709	32 (1.2%)
Vocabulary	6.51	0.95	1.0.7	2702	23.56	2.28	0…24	2702	25 (0.9%)
Sentence comprehension	6.43	0.99	1…7	2688	11.84	1.12	0…12	2688	24 (0.9%)
Working memory	6.68	0.70	1…7	2596	5.98	0.21	0…6	2631	3 (0.1%)

Note: The selective attention task contained three items. The tasks measuring verbal short term memory, phoneme identification and sentences comprehension contained 12 items. The vocabulary task contained 24 items.

TABLE 2 Descriptives for performance on early academic skills (Cito) in grade 1

	M	SD	Min-max	N
Mathematical skills	139.53	31.25	26.63…245.65	2205
Decoding	37.65	17.64	1…102	2160
Reading comprehension	114.81	28.22	8…198	1632
Vocabulary	51.76	21.21	−3…113	1554

the one-factor model ($\Delta\chi^2$ (1) = 30.58, $p < 0.001$). Even though the two-factor model fitted the data better than the one-factor model, this model still fitted the data poorly (RMSEA = 0.198, 90% CI [0.192 … 0.204], CFI = 0.756, SRMR = 0.071, $N = 2466$). We therefore included covariances between the latent variables completion and engagement, and between completion and engagement scores for each individual task. The model including these covariances fitted the data well (RMSEA = 0.071, 90% CI [0.065 … 0.077], CFI = 0.973, SRMR = 0.032, $N = 2466$).

For completion and engagement at age three, we also compared two models: a model in which completion and engagement scores of all five tasks loaded on a single variable and a model in which engagement scores from all tasks loaded on one variable and completion scores for all tasks on another variable. This comparison showed that the two-factor model fitted our data better than the one-factor model ($\Delta\chi^2$ (1) = 45.17, $p < 0.001$). As with the latent variables at age two, however, the two-factor model did not have a good enough fit (RMSEA = 0.171, 90% CI [0.166 … 0.177], CFI = 0.802, SRMR = 0.073, $N = 2709$). After adding covariances between the latent variables

completion and engagement, and between completion and engagement scores for each individual task, the model fitted the data well (RMSEA = 0.071, 90% CI [0.065 … 0.077], CFI = 0.971, SRMR = 0.031, $N = 2709$).

To test whether early academic skills could be represented as a single latent variable rather than as four separate manifest variables, we ran a final CFA, in which all four Cito measures loaded onto a single latent variable. The one-factor model showed good factor loadings for decoding (0.505), vocabulary (0.454), mathematics (0.563) and reading comprehension (0.903), but did not fit the data well enough (RMSEA = 0.119, 90% CI [0.095 … 0.144], CFI = 0.947, SRMR = 0.043, $N = 2270$). We therefore used the four variables as manifest variables in our subsequent analysis.

5.3 | Relations between engagement, completion, and early academic skills

To address our second and third research questions of (1) whether completion and engagement at ages two and three were predicted by children's background variables age, gender, home language background, parental education, and test location, and (2) how completion and engagement at ages two and three related to children's early academic skills at age six, a SEM was run with the latent and manifest variables that we obtained in the CFA analyses presented above. This model thus included latent variables for completion at age two, engagement at age two, completion at age three and engagement at age three. It also included four different manifest variables representing children's Cito scores. The model included all covariances between the tasks assessing early academic skills. We used home language background, parental education, test location and gender as control variables. From these variables, test location was only included as a control variable for the completion and engagement scores, but not for the Cito scores, since all Cito scores were conducted at school. We also controlled for age (in months) by including a centred value at ages two and three, which was calculated by subtracting the mean age of all participants at the particular test moment from the age of each individual participant. This centred score was then used as a control variable for the latent variable at that respective test wave.

The model furthermore included covariances between age at wave 1, and home language and parental education. This was because children from higher socio-economic backgrounds tended to go to daycare at a younger age, and were thus younger when tested at wave 1 than children from lower socio-economic backgrounds. The model also included covariances between parental education, and home language and test location, as children with higher educated parents were more often tested at home, and less often from families in which another language was spoken alongside or instead of Dutch. Finally, the model included the earlier mentioned covariances between completion and engagement scores for each individual task at each age, but not across tasks and ages.

The model showed good data fit (RMSEA = 0.044, 90% CI [0.042 … 0.045], CFI = 0.926, SRMR = 0.040, $N = 3667$). The results of the model can be found in Figures 1 and 2. Although one model was fitted, its results are presented in separate figures for ease of visual presentation. Figure 1 presents all relations with the variables age, gender, parental education, home language background and test location, shedding light on the characteristics of non-responders. Figure 2 presents the relation between completion and engagement scores at ages two and three, and early academic skills at age six, and thus informs us about the predictive component of non-response. For an overview of how the different tasks loaded onto the latent factors completion and engagement, as well as the correlations across the different early academic skills, see the Appendix (Figure A1).

As can be seen from Figure 1, age correlated weakly but significantly with completion and engagement at age two. Thus, the older children were, the more items they completed ($\beta = 0.228$, $p < 0.001$) and the more engaged they were ($\beta = 0.253$, $p < 0.001$). In addition, effects of home language background were found: children who did not have Dutch as their (sole) home language generally completed fewer items ($\beta = -0.231$, $p < 0.001$) and obtained lower engagement scores ($\beta = -0.212$, $p < 0.001$) than children from homes in which only Dutch was spoken. We also found significant correlations between parental education and non-response: children with higher educated parents completed more items ($\beta = 0.162$, $p < 0.001$) and were also more engaged ($\beta = 0.222$, $p < 0.001$) than children with less highly educated parents. Furthermore, we saw significant relations between gender and non-response: girls

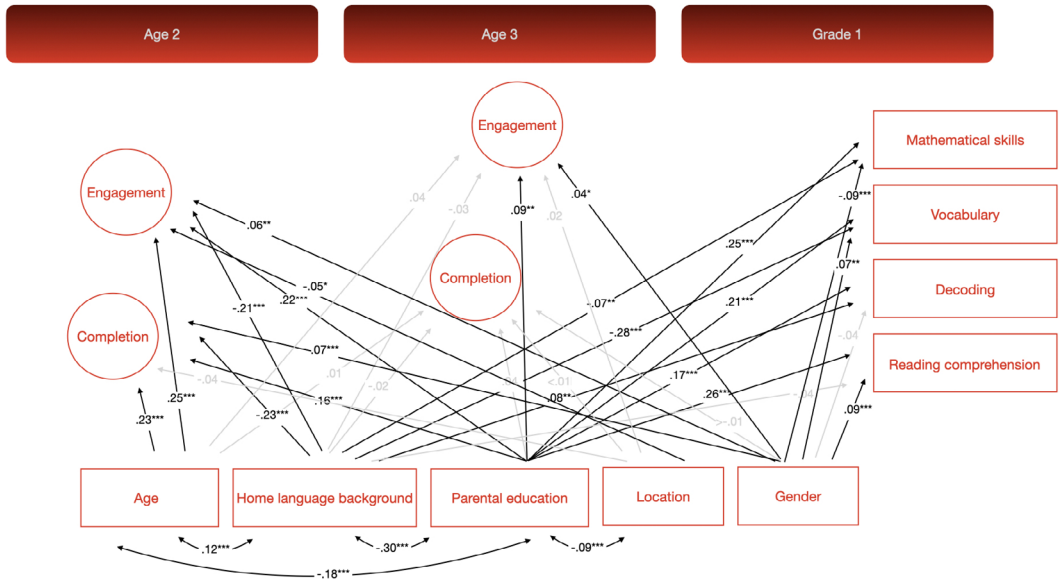

FIGURE 1 Structural equation model showing all relations with the background variables. Standardized estimates are presented. Gray lines represent non-significant coefficients. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Covariates between age and the other background variables were only included for age 2 (see under 'Analyses' for an explanation).

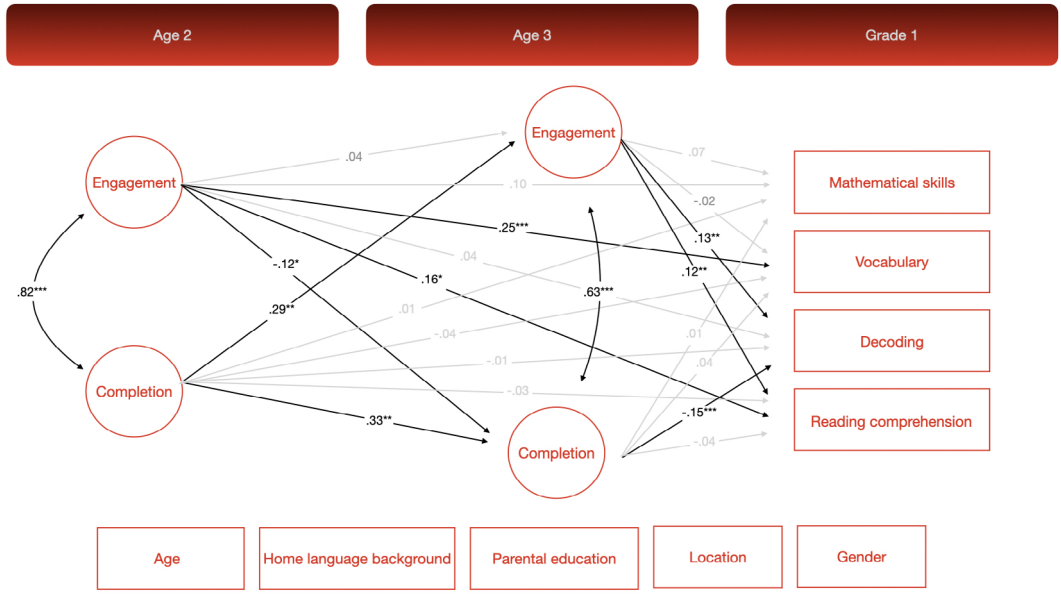

FIGURE 2 Structural equation model showing all relations between completion and engagement and early academic skills. Standardized estimates are presented. Gray lines represent non-significant coefficients. *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$.

completed slightly more items ($\beta = 0.074$, $p < 0.001$) and were somewhat more engaged ($\beta = 0.063$, $p = 0.002$) than boys. Finally, we saw a significant relation between test location and engagement, with children being less engaged when they were tested at daycare ($\beta = -0.047$, $p = 0.027$). Children also completed fewer items when tested at their daycare, but this relation was not significant ($\beta = -0.036$, $p = 0.103$).

We observed fewer significant, and overall weaker, relations between these background variables and non-response at age three: gender ($\beta = 0.044$, $p = 0.028$) and parental education ($\beta = 0.090$, $p = 0.004$) were related to engagement in a similar way as at age two, but age ($\beta = 0.041$, $p = 0.071$), home language background and test location ($\beta = 0.020$, $p = 0.392$) were not ($\beta = -0.030$, $p = 0.202$). There were no significant correlations between completion at age 3 and any of the background variables.

The results also indicated weak, but significant relations between parental education and each of the early academic skills: mathematical skills ($\beta = 0.253$, $p < 0.001$), vocabulary ($\beta = 0.207$, $p < 0.001$), decoding ($\beta = 0.166$, $p < 0.001$) and reading comprehension ($\beta = 0.263$, $p < 0.001$). Thus, children from families in which parents were highly educated generally obtained higher scores than children from families in which parents were less highly educated on all early academic skills. Moreover, significant relations with home language background were found for all tests except for the test measuring reading comprehension: children from families in which Dutch was not the (sole) home language obtained lower scores on the tests measuring vocabulary ($\beta = -0.271$, $p < 0.001$) and mathematics ($\beta = -0.068$, $p = 0.005$) than children from Dutch-only homes. However, children who heard another language than only Dutch at home scored slightly better on decoding than children who heard only Dutch ($\beta = 0.077$, $p = 0.002$). Finally, we observed a couple of significant relations between children's gender and performance on early academic skills: girls performed slightly better than boys on vocabulary ($\beta = 0.072$, $p = 0.001$) and reading comprehension ($\beta = 0.094$, $p < 0.001$), whereas boys performed slightly better on mathematics ($\beta = -0.089$, $p < 0.001$), and on decoding ($\beta = -0.037$, $p = 0.080$), although this latter relation was not significant.

The results in Figure 2 show a number of significant relations between completion and engagement scores on the one hand and early academic skills on the other hand, but these are generally weak. Specifically, we found weak but significant relations between reading comprehension and engagement at both age two ($\beta = 0.162$, $p = 0.009$) and age three ($\beta = 0.119$, $p = 0.004$). Engagement at age three was found to be positively related to decoding ($\beta = 0.134$, $p = 0.001$), whereas completion at this age was negatively related to decoding ($\beta = -0.151$, $p < 0.001$). Furthermore, there was a weak but significant relation between engagement at age two and vocabulary in grade 1 ($\beta = 0.251$, $p < 0.001$). Regarding the relations between completion and engagement between ages two and three, the results in Figure 2 show that engagement at age two was negatively related to completion at age three ($\beta = -0.123$, $p = 0.018$), but not to engagement at that same age ($\beta = 0.040$, $p = 0.449$). However, there was a moderate significant relation between completion at age two and engagement at age three ($\beta = 0.289$, $p < 0.001$). There was also a relation of comparable strength between completion at age two and completion at age three ($\beta = 0.334$, $p = 0.001$). Furthermore, although our CFA indicated that completion and engagement should best not be considered a single factor, they were highly correlated at both age two ($\beta = 0.819$, $p < 0.001$) and age three ($\beta = 0.632$, $p = 0.026$).

6 | DISCUSSION

In the present study we examined non-response in a sample of over 3000 two-and three-year olds who were administered a series of developmental tasks at ages two and three and tests of early academic skills in grade 1. The aim of our study was to examine whether non-response at toddler age could be predicted by children's background characteristics and whether it was predictive of early academic skills 3 to 4 years later. More specifically, we assessed (1) whether two aspects associated with non-response - here labelled task completion and engagement - were separate constructs, (2) which background characteristics (i.e., age, gender, parental education, home language and test location) were predictive of non-response, (3) whether non-response was stable from age two to age three and (4) whether non-response at ages two and three predicted early academic skills in grade 1.

Using SEM analyses, we found no evidence that task completion and engagement were represented by a single latent construct. Rather, our analyses showed that, both at ages two and three, these two factors were better

represented as separate variables, than as a single latent variable. However, in our final model the two were highly correlated, especially at age two. These outcomes suggest that the various behaviours associated with data loss are separable yet interrelated, such that children who do not respond are likely to be (perceived as) not engaged, and vice versa. It is important to note, however, that non-response may reflect childrens' competences as measured by a particular task, such that children do not respond to items they do not know the answer for. Yet, if this were the case, it is still important to understand which children are likely not to give a response, even if this is merely an indication of their abilities to respond accurately in a task. Taking these children out of a data set could lead to the overestimation of a particular effect in the remaining sample, because it would be based on good performers. Likewise, excluding the items to which children do not respond might lead to a slight overestimation of performance in children with relatively many of such items. Hence, even if non-response reflects to some degree children's abilities to perform a task, it is still crucial to see whether and when non-response is selective, so as to be able to minimize it in studies with young children.

Earlier research (Slaughter & Suddendorf, 2007; van der Velde & Junge, 2020) has shown that several background characteristics are associated with non-response. Our results corroborate these findings, as we found that age, gender, parental education, and home language background (i.e., being from a Dutch-only home or not) predicted both completion and engagement at age two, and test location predicted engagement at age 2, although this latter effect was rather small. At age three, gender and parental education predicted engagement. It is important to point out that non-completion at age three was relatively low: seemingly, the tasks were less demanding for most children at this age, with the exception of the nonword repetition task, which was the sole task requiring a verbal response from the child. Our results suggest that when tasks are demanding for a particular age group, children who do not speak the target language (well) or have lower educated parents are less likely to respond to the items of a task than children who do speak the target language well and have higher educated parents. These findings highlight the need to include children from various language and socio-economic backgrounds in study samples to obtain representative samples, and, in fact, oversample children from these groups to compensate for higher degrees of data loss.

As for longitudinal relations between non-response and early academic skills, we observed that completion and engagement predicted early academic skills at least to a certain extent. Specifically, engagement at age two predicted both reading comprehension and vocabulary in grade 1. Engagement at age three predicted reading comprehension and decoding in grade 1. A final, somewhat unexpected finding was the negative relation between completion at age three and decoding in grade 1. It is important to note that completion scores at age three were high overall, and non-completion was largely due to the nonword repetition task. Perhaps, children at this age rather did not repeat a non-word at all than repeat it incorrectly, in which case non-completion at age three would be indicative of children's inhibition of wrong answers or self-awareness, which might, in turn, be related to decoding in grade 1. All other relations between non-response and early academic skill were in line with the idea that children who are engaged with tasks when they were toddlers constitute a special group that does well at early academic tests at school when they are 6 years old, at least on tests measuring language (−related) skills.

Non-response turned out to be somewhat stable across ages two and three: completion at age two was significantly related to completion and engagement at age three. Yet, we also observed that engagement at age two was negatively related to completion at age three. Again, a speculative interpretation of this result is that non-completion on the non-word repetition task reflected children's inhibition of wrong answers or metalinguistic awareness, along the lines of the ideas sketched above. Taken together, our findings suggest that children who do not complete tasks at age two tend to be the same children that will not respond at age three. The overall picture that emerged from our results then is that non-response rates in developmental tasks at 2 years of age are subject to effects of child characteristics (home language background, parental education) and have some predictive value for non-response at age three and for early academic skills at age six.

6.1 | General discussion

All in all, our results support earlier work that non-response in children can be predicted by several background characteristics (Bathurst & Gottfried, 1987; van der Velde & Junge, 2020) and that this is not only the case for infants, but also for slightly older children aged two and 3 years. Moreover, our CFAs showed that task completion and engagement are two separable, but related, constructs, which are related to child characteristics such as age and home language. Yet, our results indicate that especially the linguistic competences that are required to participate might form a hindrance: children from non Dutch-monolingual homes were more likely to be non-responders, and non-response was related to linguistic skills in grade 1. The idea that linguistics competence plays a role in task completion and engagement is also supported by the fact that, in general, non-response was low at age three, whereas we did observe quite some non-response at age two. Taken together, there seems to be a fine line between the complexity of the task procedure, and linguistic competence necessary for the task, on the one hand, and non-response on the other hand. The tasks in this test battery were quite difficult for a subset of the two-year-olds, but did not seem too difficult anymore for three-year-olds. Older children likely will not show non-response on the tasks as they were administered here, but they might show non-response if the tasks become more complex, for example, in terms of their linguistic demands or response requirements. In this respect, it is also important to note that different non-response patterns might exist: children could respond less towards the end of a task, or rather respond less in the beginning, and there are likely many other possible non-response patterns. Because of the wide variety of patterns that could occur, which might differ across tasks as well, and the fact that we were not interested in the mechanics of a particular task, we did not investigate such non-response patterns further in this study. However, future studies that delve deeper into the mechanics of a particular task, could take non-response patterns into account when looking at task complexity.

Another factor that might have accounted, at least in part, for the non-response in our study is that it contained five computerized tasks that were presented successively. The highest rates of non-response were found for the fifth task (at age two) and the fourth task (at age three) in the sequence, which suggests that fatigue may have played a role. Future work could examine effects of task duration (and task order) on non-response. In addition, it would be worth exploring which properties of tasks induce minimal non-response in toddlers, and see how these properties may be incorporated in other tasks. In our study, for example, we noticed that the working memory task (i.e., Six Boxes task, Diamond et al., 1997) was performed by virtually all children at age 3. While this might be partly due to the fact that guessing was not penalized (unlike, for example, in the visual search task where negative feedback was given), this could also be due to the fact that children performed a fun task that fitted their developmental stage (i.e., look for toys). Perhaps, looking for tangible toys in boxes seduced even the least responsive children in the sample to perform the trials, suggesting that it may be worthwhile to see which characteristics of tasks eliciting high response rates can be incorporated in other tasks as well. Our study has a number of limitations. First, we cannot exclude that non-response effects are partially caused by the high number of language tasks (four out of five) in our study. Nevertheless, we used a wide variety of tasks that required various types of stimuli (visual, auditory, laptop, tangible objects) and various types of responses (pointing, grasping objects, producing language). We also tried to keep verbal demands to participate low by using scripted language involving a small number of very short sentences containing only high-frequency words, supported with gestures. Future research could address to what extent our outcomes generalize to tasks in other domains. A second limitation is that home language background information was missing for a considerable number of children (21%). In addition, children from families with highly-educated parents were overrepresented in our sample. Since effects were still found, this might indicate that we are actually underestimating the observed effects. Finally, children's engagement was based on an assessor rating scale that was not very specific. In future work, clear protocols, coding systems, and interrater reliability checks should be included to make measurements more reliable, and as such, allow for a more thorough investigation of how engagement relates to completion as well as other factors associated with non-response.

Despite its limitations, our study constitutes a first step towards understanding to what extent non-response in developmental research can be predicted by children's background and whether it is predictive of later scholastic achievement. In developmental studies, low completion rates often lead to exclusion of children, but low

engagement typically does not, except in younger children where it is subsumed under the general notion of 'fussiness'. Since these two constructs are not necessarily the same, studies should be clear about which criteria they use for excluding children. In fact, as non-response appears not to be random, researchers should be careful to exclude children and ideally use analytical techniques in which non-complete data can be included, such as mixed effects modelling or Item Response Theory, in which scores at the item level can be analysed. It is also recommended to include attrition analyses and fine-tune conclusions towards the sample being looked at, to avoid overgeneralizing or overstating results. At the very least, consistent use of such analyses and sample descriptions will help give us some idea of sampling bias in our field, and of the consequences of such biases for theory construction.

AUTHOR CONTRIBUTIONS

Sybren Spit: Conceptualization; formal analysis visualization; writing - original draft. **Hanna Mulder:** Data curation; methodology; writing – review and editing. **Carolien van Houdt:** Data curation; writing – review and editing. **Josje Verhagen:** Conceptualization; methodology; writing – original draft.

CONFLICT OF INTEREST

The authors declare that they have no conflict of interest.

PEER REVIEW

The peer review history for this article is available at https://publons.com/publon/10.1002/icd.2376.

DATA AVAILABILITY STATEMENT

All data, stimuli and analysis scripts are available on our OSF page https://osf.io/7cwej.

ORCID

Sybren Spit https://orcid.org/0000-0002-9724-8912

REFERENCES

Andringa, S., & Godfroid, A. (2020). Sampling bias and the problem of generalizability in applied linguistics. *Annual Review of Applied Linguistics*, *40*, 134–142. https://doi.org/10.1017/S0267190520000033

Arnett, J. J. (2008). The neglected 95%: Why American psychology needs to become less American. *American Psychologist*, *63*(7), 602–614. https://doi.org/10.1037/0003-066X.63.7.602

Bathurst, K., & Gottfried, A. W. (1987). Untestable subjects in child development research: Developmental implications. *Child Development*, *58*(4), 1135–1144. https://doi.org/10.2307/1130552

Bell, J. C., & Slater, A. (2002). The short-term and longer-term stability of non-completion in an infant habituation task. *Infant Behavior & Development*, *25*(2), 147–160. https://doi.org/10.1016/S0163-6383(02)00118-2

Brown, T. A. (2015). *Confirmatory factor analysis for applied research*. Guilford Press.

Diamond, A., Prevor, M. B., Callender, G., & Druin, D. P. (1997). Prefrontal cortex cognitive deficits in children treated early and continuously for PKU. *Monographs of the Society for Research in Child Development*, *62*(4), i–208.

Dunn, L., & Dunn, L. M. (2005). *Peabody picture vocabulary test-III-NL. Nederlandse versie door Liesbeth Schlichting [Dutch version by Liesbeth Schlichting]*. Harcourt Assessment B.V.

Engzell, P., Frey, A., & Verhagen, M. (2021). Learning loss due to school closures during the COVID-19 pandemic, supplementary information. *PNAS*, *118*(17), e2022376118. https://doi.org/10.1073/pnas.2022376118

Espy, K. A., Sheffield, T. D., Wiebe, S. A., Clark, C. A., & Moehr, M. J. (2011). Executive control and dimensions of problem behaviors in preschool children. *Journal of Child Psychology and Psychiatry, and Allied Disciplines*, *52*(1), 33–46. https://doi.org/10.1111/j.1469-7610.2010.02265.x

Frank, M. C., Bergelson, E., Bergmann, C., Cristia, A., Floccia, C., Gervain, J., Hamlin, J. K., Hannon, E. E., Kline, M., Levelt, C., & Lew-Williams, C. (2017). A collaborative approach to infant research: Promoting reproducibility, best practices, and theory-building. *Infancy*, *22*(4), 421–435. https://doi.org/10.1111/infa.12182

Gerrits, E. (2003). Speech perception of young children at risk for dyslexia and children with specific language impairment. Proceedings of the 15th International Conference of Phonetic Sciences (ICPhS), .

Grimm, K. J., Kuhl, A. P., & Zhang, Z. (2013). Measurement models, estimation, and the study of change. *Structural Equation Modeling: A Multidisciplinary Journal*, *20*(3), 504–517. https://doi.org/10.1080/10705511.2013.797837

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). Most people are not WEIRD. *Nature*, *466*(7302), 29. https://doi.org/10.1038/466029a

Hernán, M. A., & Robins, J. M. (2020). *Causal inference: What if?* Chapman & Hall/CRC.

Hughes, C., & Ensor, R. (2005). Executive function and theory of mind in 2 year olds: A family affair? *Developmental Neuropsychology*, *28*(2), 645–668. https://doi.org/10.1207/s15326942dn2802_5

Imbens, G. W., & Rubin, D. B. (2015). *Causal inference for statistics, social, and biomedical sciences. An Introduction*. Cambridge University Press.

Johnson, V. A., de Villiers, J. G., & Seymour, H. N. (2005). Agreement without understanding? The case of third person singular/s. *First Language*, *25*(3), 317–330. https://doi.org/10.1177/0142723705053120

Klein-Radukic, S., & Zmyj, N. (2015). Dropout in looking time studies: The role of infants' temperament and cognitive developmental status. *Infant Behavior and Development*, *41*, 142–153. https://doi.org/10.1016/j.infbeh.2015.10.001

Kline, R. B. (2005). *Principles and practice of structural equation modeling* (2nd ed.). Guilford Press.

Kruskal, W., & Mosteller, F. (1979). Representative sampling, III: The current statistical literature. *International Statistical Review*, *47*(3), 245–265. https://doi.org/10.2307/1402647

Kuhn, L. J., Willoughby, M. T., Parramore Wilbourn, M., Vernon-Feagans, L., Blair, C. B., & Family Life Project Investigators. (2014). Early communicative gestures prospectively predict language development and executive function in early childhood. *Child Development*, *85*(5), 1898–1914. https://doi.org/10.1111/cdev.12249

Kuijpers, C. (1996). Perception of the voicing contrast by Dutch children and adults. *Journal of Phonetics*, *24*(3), 367–382. https://doi.org/10.1006/jpho.1996.0020

Kukull, W., & Ganguli, M. (2012). Generalizability: The trees, the forest, and the low-hanging fruit. *Neurology*, *78*(23), 1886–1891. https://doi.org/10.1212/WNL.0b013e318258f812

Little, T. D. (2013). *Longitudinal structural equation modeling*. Guilford Press.

ManyBabies Consortium. (2020). Quantifying sources of variability in infancy research using the infant-directed-speech preference. *Advances in Methods and Practices in Psychological Science*, *3*(1), 24–52. https://doi.org/10.1177/2515245919900809

McArdle, J. J., Grimm, K. J., Hamagami, F., Bowles, R. P., & Meredith, W. (2009). Modeling life-span growth curves of cognition using longitudinal data with multiple samples and changing scales of measurement. *Psychological Methods*, *14*(2), 126–149. https://doi.org/10.1037/a0015857

Mulder, H., Hoofs, H., Verhagen, J., Van der Veen, I., & Leseman, P. P. M. (2014). Psychometric properties and convergent validity of an executive function test battery for 2-year-olds using item response theory. *Frontiers in Psychology*, *5*, 733. https://doi.org/10.3389/fpsyg.2014.00733

Mulder, H., Verhagen, J., Van der Ven, S. H. G., Slot, P. L., & Leseman, P. P. M. (2017). Early executive function at age two predicts emergent mathematics and literacy at age five. *Frontiers in Psychology*, *8*, 1706. https://doi.org/10.3389/fpsyg.2017.01706

Nielsen, M., Haun, D., Kärtner, J., & Legare, C. H. (2017). The persistent sampling bias in developmental psychology: A call to action. *Journal of Experimental Child Psychology*, *162*, 31–38. https://doi.org/10.1016/j.jecp.2017.04.017

Onland-Moret, N. C., Buizer-Voskamp, J. E., Albers, M. E. W. A., Brouwer, R. M., Buimer, E. E. L., Hessels, R. S., de Heus, R., Huijding, J., Junge, C. M. M., Mandl, R. C. W., Pas, P., Vink, M., van der Wal, J. J. M., Hulshoff Pol, H. E., & Kemner, C. (2020). The YOUth study: Rationale, design, and study procedures. *Developmental Cognitive Neuroscience*, *46*, 100868. https://doi.org/10.1016/j.dcn.2020.100868

Pearl, J. (2009). *Causality: Models, reasoning and inference*. Cambridge University Press.

Pearl, J., Glymour, M., & Jewell, N. P. (2016). *Causal inference in statistics. A primer*. John Wiley & Sons Ltd.

Posthumus, H., Scholtus, S., & Walhout, J. (2019). De nieuwe onderwijsindicator primair onderwijs. Samenvattend rapport. https://www.cbs.nl/-/media/_pdf/2019/45/de-nieuwe-onderwijsachterstandenindicator-primair-onderwijs.pdf

Poulin-Dubois, D., Blaye, A., Coutya, J., & Bialystok, E. (2011). The effects of bilingualism on toddlers' executive functioning. *Journal of Experimental Child Psychology*, *108*(3), 567–579. https://doi.org/10.1016/j.jecp.2010.10.009

Rosseel, Y. (2012). Lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1–36. https://doi.org/10.18637/jss.v048.i02

Scott, R. M., He, Z., Baillargeon, R., & Cummins, D. (2012). False-belief understanding in 2.5-year-olds: Evidence from two novel verbal spontaneous-response tasks. *Developmental Science*, *15*(2), 181–193. https://doi.org/10.1111/j.1467-7687.2011.01103.x

Sekerina, I., Stromswold, K., & Hestvik, A. (2004). How do children and adults process referentially ambiguous pronouns? *Journal of Child Language*, *31*(1), 123–152. https://doi.org/10.1017/S0305000903005890

Slaughter, V., & Suddendorf, T. (2007). Participant loss due to 'fussiness' in infant visual paradigms: A review of the last 20 years. *Infant Behavior & Development*, *30*(3), 505–514. https://doi.org/10.1016/j.infbeh.2006.12.006

Stets, M., Stahl, D., & Reid, V. M. (2012). A meta-analysis investigating factors underlying attrition rates in infant ERP studies. *Developmental Neuropsychology*, *37*(3), 226–252. https://doi.org/10.1080/87565641.2012.654867

van der Velde, B., & Junge, C. (2020). Limiting data loss in infant EEG: Putting hunches to the test. *Developmental Cognitive Neuroscience*, *45*, 100809. https://doi.org/10.1016/j.dcn.2020.100809

Verhagen, J., de Bree, E. H., Mulder, H., & Leseman, P. P. M. (2017). Effects of vocabulary and Phonotactic probability on two-year-Olds' nonword repetition. *Journal of Psycholinguistic Research*, *46*(3), 507–524. https://doi.org/10.1007/s10936-016-9448-9

Verhagen, J., Boom, J., Mulder, H., de Bree, E. H., & Leseman, P. P. M. (2019). Reciprocal relationships between nonword repetition and vocabulary during the preschool years. *Developmental Psychology*, *55*(6), 1125–1137. https://doi.org/10.1037/dev0000702

Vlug, K. V. (1997). Because every pupil counts: The success of the pupil monitoring system in The Netherlands. *Education and Information Technology*, *2*(4), 287–306. https://doi.org/10.1023/A:1018629701040

Willoughby, M. T., Blair, C. B., Wirth, R. J., Greenberg, M., & Family Life Project Investigators. (2010). The measurement of executive function at age 3 years: Psychometric properties and criterion validity of a new battery of tasks. *Psychological Assessment*, *22*(2), 306–317. https://doi.org/10.1037/a0018708

Willoughby, M. T., Wirth, R. J., Blair, C. B., & Family Life Project Investigators. (2012). Executive function in early childhood: Longitudinal measurement invariance and developmental change. *Psychological Assessment*, *24*(2), 418–431. https://doi.org/10.1037/a0025779

Zumbuehl, M., & Dillingh, R. (2020). Ongelijkheid van het jonge kind. Centraal Planbureau. https://www.cpb.nl/sites/default/files/omnidownload/CPB-Notitie-Ongelijkheid-van-het-jonge-kind.pdf

How to cite this article: Spit, S., Mulder, H., van Houdt, C., & Verhagen, J. (2023). Can we predict non-response in developmental tasks? Assessing the longitudinal relation between toddlers' non-response and early academic skills. *Infant and Child Development*, *32*(1), e2376. https://doi.org/10.1002/icd.2376

APPENDIX A

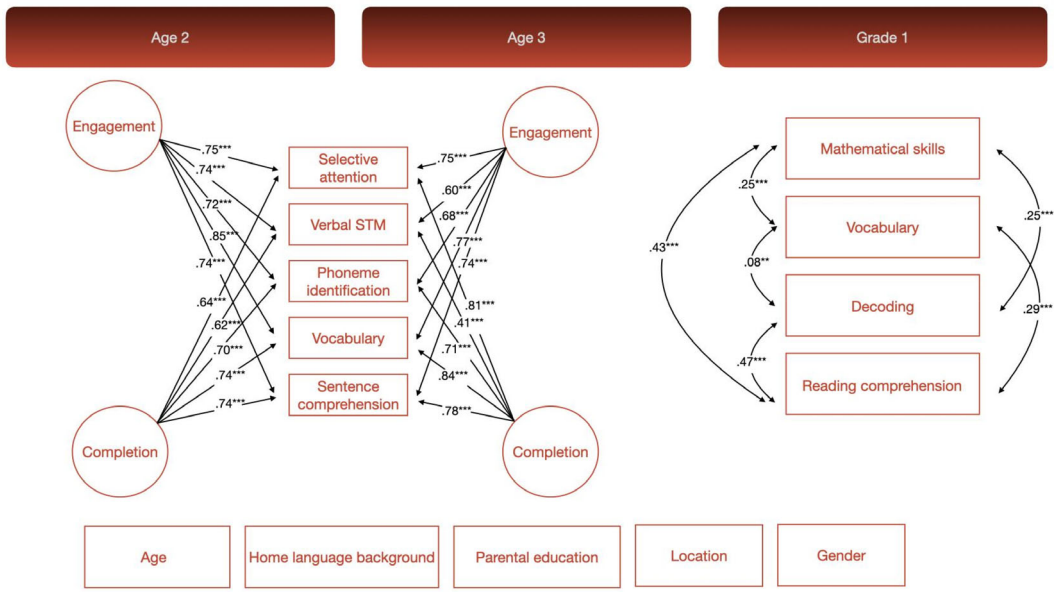

FIGURE A1 Factor loadings and covariates of the Structural Equation Model. Structural equation model showing loadings of tasks onto the latent variables and the relevant covariates. Standardized estimates are presented. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$