1   PREVIC: An adaptive parent report measure of expressive vocabulary in children between

2   3 and 8 years of age

3   Manuel Bohn[1,2], Julia Prein[2], Jonas Engicht[3], Daniel Haun[2], Natalia Gagarina[4], & Tobias

4   Koch[3]

5   [1] Institute for Psychology, Leuphana University Lüneburg, Germany

6   [2] Department of Comparative Cultural Psychology, Max Planck Institute for Evolutionary

7   Anthropology, Leipzig, Germany

8   [3] Institut of Psychology, Friedrich-Schiller-University Jena, Germany

9   [4] Leibniz-Zentrum Allgemeine Sprachwissenschaft, Berlin, Germany

10   Author Note

19   Correspondence concerning this article should be addressed to Manuel Bohn,

20   Leuphana University Lüneburg, Universitätsallee 1, 21335 Lüneburg, Germany. E-mail:

21   manuel.bohn@leuphana.de

Abstract

Parent report measures have proven to be a valuable research tool to study early language development. Caregivers are given a list of words and are asked which of them their child has already used. However, most available measures are not suited for children beyond infancy, come with substantial licensing costs or lack a clear psychometric foundation. Here we present the PREVIC (Parent Report of Expressive Vocabulary in Children) an open access, high quality vocabulary checklist for German-speaking children between three and eight years of age. The PREVIC was constructed leveraging the advantages of Item Response Theory: we designed a large initial item pool of 379 words and collected data from N = 1190 caregivers of children between three and eight years of age. Based on this data, we computed a range of fit indices for each item (word) and used an automated item selection algorithm to compile a final pool that contains items that a) vary in difficulty and b) fit the Rasch (one-parameter logistic) model. The resulting task is highly reliable and shows convergent validity. The IRT-based construction allowed us to design an adaptive version of the task, which substantially reduces the duration of the task while retaining measurement precision. The task – including the adaptive version – was implemented as a website and is freely accessible online at https://ccp-odc.eva.mpg.de/previc-demo/. The PREVIC can easily be adapted to other languages and fills an important gap in the toolkit of researchers interested in language development.

Keywords: language development, vocabulary, individual differences, Item Response Models

Word count: X

PREVIC: An adaptive parent report measure of expressive vocabulary in children between

3 and 8 years of age

Introduction

Learning language is one of the key developmental objectives for children. This learning process is highly variable and leads to persistent individual differences which are related to a wide range of outcome measures later in life (Bleses, Makransky, Dale, Højen, & Ari, 2016; Bornstein, Hahn, Putnick, & Pearson, 2018; Golinkoff, Hoff, Rowe, Tamis-LeMonda, & Hirsh-Pasek, 2019; Marchman & Fernald, 2008; Morgan, Farkas, Hillemeier, Hammer, & Maczuga, 2015; Pace, Alper, Burchinal, Golinkoff, & Hirsh-Pasek, 2019; Pace, Luo, Hirsh-Pasek, & Golinkoff, 2017; Schoon, Parsons, Rush, & Law, 2010; Walker, Greenwood, Hart, & Carta, 1994). For example, in a longitudinal study spanning 29 years, Schoon et al. (2010) found that relatively poorer language skills at age five were associated with lower levels of mental health at age 34. Given the high predictive validity of early language abilities, researchers and practitioners alike need high-quality, easy access measures to assess individual differences. However, such measures are rare and those that exist often come with substantial licensing costs. In this paper, we describe the development of an open, efficient and valid measure of individual differences in expressive vocabulary.

Child language measures can be broadly categorized into two types: direct and parent report measures. Direct measures are generally used with children of three years and older. Direct expressive language assessments involve prompting children to generate words or sentences in response to a stimulus, such as a picture or an object. Direct receptive language assessments reverse the logic and require children to match a verbal prompt with a picture or an object. Various direct measures tailored to different languages and age groups have been developed, including measures for English and German (Bohn et al., 2023; Dunn & Dunn, 1965; Dunn, Dunn, Whetton, & Burley, 1997; Glück & Glück,

70 2011; Golinkoff et al., 2017; Kauschke & Siegmüller, 2002; Kiese-Himmel, 2005; Lenhard,

71 Lenhard, Segerer, & Suggate, 2015). Additionally, standardized cognitive ability tests

72 frequently incorporate direct language measures (e.g., Bayley, 2006; Gershon et al., 2013;

73 Wechsler & Kodama, 1949).

74   Parent report measures in general are widely utilized in psychological research. They

75 are particularly popular as screening methods to identify developmental delays (Diamond

76 & Squires, 1993; Pontoppidan, Niss, Pejtersen, Julian, & Væver, 2017). However, it is

77 important to acknowledge that parent reports come with certain caveats, including the

78 potential for selective reporting and social desirability. As a consequence, providing a

79 comprehensive assessment of the overall quality and usefulness of these measures is

80 challenging (Morsbach & Prinz, 2006). Nonetheless, some parent report measures have

81 been found to be both reliable and valid (Bodnarchuk & Eaton, 2004; Hornman, Kerstjens,

82 Winter, Bos, & Reijneveld, 2013; Ireton & Glascoe, 1995; Macy, 2012; Saudino et al., 1998).

83   In child language research, parent report measures are often utilized with very young

84 children when direct assessment is challenging. One widely used measure is the

85 MacArthur-Bates Communicative Development Inventories (CDI, Fenson et al., 2007). The

86 CDI asks parents to to check those words from a checklist that they believe their child

87 produces and/or understands. This measure has been adapted for a wide range of spoken

88 and signed languages (see Frank, Braginsky, Yurovsky, & Marchman, 2021 for an

89 overview), with various versions available (e.g., Makransky, Dale, Havmose, & Bleses, 2016;

90 Mayor & Mani, 2019), including an online version (DeMayo et al., 2021). Collaborative

91 efforts have facilitated the pooling of CDI data from thousands of children learning

92 different languages into centralized repositories (Frank, Braginsky, Yurovsky, & Marchman,

93 2017; Jørgensen, Dale, Bleses, & Fenson, 2010). Importantly, the CDI exhibits validity as

94 parental reports align with direct observations and assessments of child language

95 (Bornstein & Haynes, 1998; Dale, 1991; Feldman et al., 2005; Fenson et al., 1994).

⁹⁶ However, the use of the CDI – in typically developing children – is limited to 37

⁹⁷ months of age. Beyond this point, most children are reported to say all the words on the

⁹⁸ list. Consequently, there is a need for a comparable measure that can be applied to older

⁹⁹ children. Existing instruments focusing on general cognitive development often include

¹⁰⁰ language scales; however, these scales lack detailed information and fail to capture

¹⁰¹ individual differences effectively (Ireton & Glascoe, 1995). For example, the Ages and

¹⁰² Stages Questionnaire (ASQ) at 36 months comprises only six items that encompass general

¹⁰³ communicative behavior, such as whether the child can say their full name when prompted

¹⁰⁴ (Squires, Bricker, Twombly, et al., 2009). One notable example of a dedicated language

¹⁰⁵ measure for older children is the Developmental Vocabulary Assessment for Parents

¹⁰⁶ (DVAP, Libertus, Odic, Feigenson, & Halberda, 2015). The DVAP is derived from the

¹⁰⁷ words used in the Peabody Picture Vocabulary Test (PPVT, Dunn & Dunn, 1965), a

¹⁰⁸ widely used direct measure of receptive vocabulary. As perhaps expected, the DVAP

¹⁰⁹ demonstrates high convergent validity, as evidenced by its strong correlation with the

¹¹⁰ PPVT. However, the proprietary nature of the PPVT limits the utility of the DVAP for

¹¹¹ researchers.[1] As a consequence, it is unlikely that a comparable "success story" – as

¹¹² observed with the CDI – will emerge where researchers have adapted the original English

¹¹³ form to different languages and more efficient forms.

¹¹⁴ A more general issue with existing language measures – including PPVT and DVAP

¹¹⁵ – is a lack of psychometric grounding. Items that make up the scale are selected based on

¹¹⁶ researchers intuitions and there is no clear measurement model that explicates how the

¹¹⁷ different items and test scores are linked to the construct in question (Borsboom, 2006).

¹¹⁸ Item response theory (IRT) provides a theoretical framework and a toolkit to develop tasks

¹¹⁹ with a solid psychometric foundation (Kubinger, 2006; Lord, 2012). Within IRT, all items

---

[1] When the first author approached the license holder of the PPVT in Germany to ask if we could use the German version of the PPVT to build a parental report measure, we were told that we would have to pay for every administration of the new measure and we would not be allowed to openly share the materials.

120 are assumed to measure the same latent construct. Each item is linked to the construct by

121 a probability function (e.g. a logistic curve) which determines how likely a particular

122 response is for individuals with different values on the latent construct. The shape of this

123 curve is defined by the difficulty of an item (i.e. the value of the latent construct when the

124 probability to solve the item is 50%) and its discrimination (i.e. the slope of the curve

125 showing how the probability to solve the item changes with increasing levels on the latent

126 construct). In a *Rasch* model (or one-parameter IRT model), all items are assumed to have

127 equal item discriminations, resulting in parallel item characteristic curves. The great

128 benefit of IRT is that models are testable in that we can quantify the fit of each model and

129 compare competing models. For each item, we can compute fit statistics that indicate how

130 well the model captures the response pattern to the item. Test construction is

131 straightforward in this framework; items are selected that improve the fit to the model.

132 Besides many other advantages (e.g., objective specificity, sum scores can be used as

133 sufficient statistics), the Rasch model also allows for adaptive testing. Here, individuals do

134 not simply respond to all items in the task but only to a subset of optimally informative

135 items given their – constantly updated – value on the latent construct.

136 The downside of IRT-based test construction – and probably the reason it is not used

137 more often – is that it requires a larger initial investment. To be able to remove items with

138 a poor fit during the selection process requires an initial item pool that is substantially

139 larger than the desired size of the final task. Adaptive testing also needs a large item pool

140 so that there is a sufficient number of items that are optimally informative for different

141 regions of the latent construct. Furthermore, to obtain solid estimates for the item

142 parameters it takes large sample sizes. Yet, we believe that these disadvantages are clearly

143 outweighed by the benefits that come with IRT-based test construction in the long run.

The current study

144   Our goal was to develop a high-quality and easy-access vocabulary checklist beyond

145 the CDI for children between three and eight years of age. To ensure the psychometric

146 quality of the task and to allow for adaptive testing, we used IRT to guide item selection

147 and the construction of the item pool. In particular, we constructed the item pool so that

148 all items fit the simplest version of an IRT model, the one-parameter logistic (1PL) or

149 Rasch model (Rasch, 1980). The main reason behind this fairly restrictive approach was

150 that only when the Rasch model holds is the number of solved items (sum score) a

151 sufficient statistic and can be used to represent an individual's value on the latent

152 construct (Birnbaum, 1986). To ensure easy-access, we implemented the checklist as an

153 interactive web-app. Furthermore, the task, the item pool and all associated materials are

154 openly available for other researchers to use. In what follows, we describe the

155 implementation of the task, the construction of the item pool and the item selection

156 procedure. Finally, we report the results on a study assessing the convergent validity of a

157 task that comprises all items in the pool.

Methods

Task design an implementation

161   We decided to use an interactive format instead of presenting parents with a long list

162 of words. The task was implemented as a web-app using `html` and `JavaScript` and ran in

163 every modern web-browser on computers, tablets and smartphones. Words were presented

164 one-by-one and caregivers could indicate whether or not their child says a word either by

165 using the familiar swipe-left/swipe-right functionality, by clicking symbols or using

166 arrow-keys on a keyboard (see Figure 1A). For example, caregivers saw the word "Jacke"

167 (en: jacket) on a card in the center of the screen; to report that their child says the word,

168 they would swipe the card to the right side of the screen which would make the card go

169 away and the next in the deck appear. We included a lightweight back-end that registered
170 the last completed trial so that caregivers could take breaks and even switch devices during
171 the task.

172   For each child we created a personalized link that connected the caregiver's responses
173 to the child's entry in our database. After clicking the link, participants saw a short video
174 in which the first author introduced the rationale of the study. Next, they were introduced
175 to the functionality of the task (Figure 1A) and how to respond. We used the same
176 instructions for how to judge whether a child says a word or not as the German version of
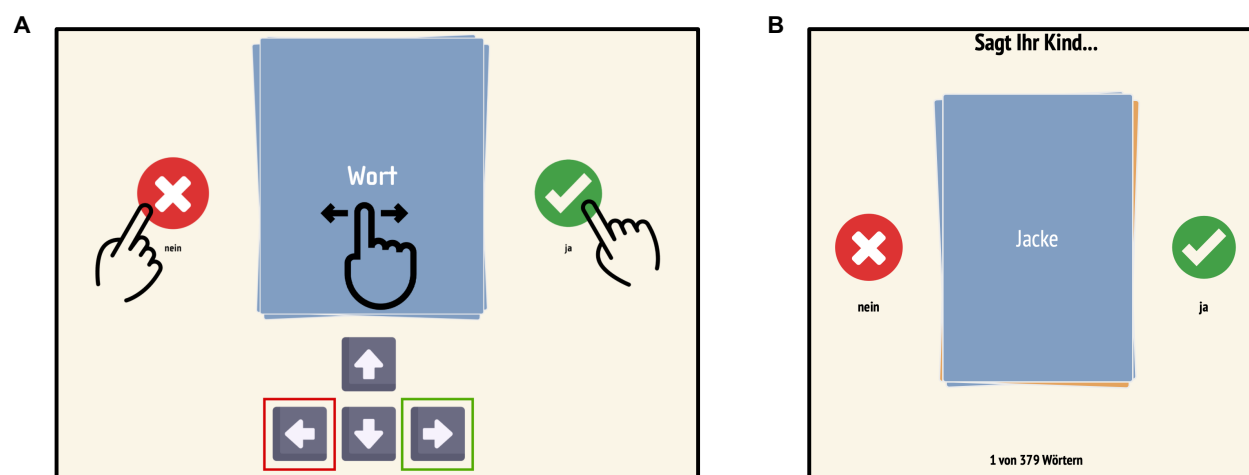177 the CDI (FRAKIS, Szagun, Stumper, & Schramm, 2009).

Figure 1. Task implementation. (A) Instructions provided to parents demonstrating the functionality of the task. The word was presented on a card in the middle of the screen. Parents could indicate whether or not their child says a word by swiping the card left (no) or right (yes), touching or clicking the "yes" or "no" symbol or pressing the left or right arrow key on the keyboard. (B) Screenshot from the task presenting the German word Jacke (en: jacket).

Item pool generation

Our goal was to create an item pool with items of varying difficulty. We used age-of-acquisition (AoA) ratings as a rough indicator of anticipated item difficulty. Previous work has shown strong associations between AoA ratings and how likely children are to know a word (Bohn et al., 2023; Bohn, Tessler, Merrick, & Frank, 2021). We started the process by compiling a list with AoA ratings for 3,921 German words from various sources (Birchenough, Davies, & Connelly, 2017; Łuniewska et al., 2019; Schröder, Gemballa, Ruppin, & Wartenburger, 2012). We excluded words with AoA ratings above ten. The remaining words were ordered by rated AoA and then split into ten lists with 344 words each. A research assistant with a background in linguistics went through the lists and selected words that a) were indicative of language abilities more broadly (avoiding very specialized terms) and b) that were different from one another in that they were semantically unrelated (to avoid words that are learned in the same context). For each list, we aimed for roughly 35 words; 17 nouns, nine verbs and nine adjectives. The so-generated item pool had 379 words, of which 197 were nouns, 92 were verbs and 90 were adjectives. Figure 2A shows how the items were distributed across AoA ratings and word types.

Data collection

Next, we aimed to collect data for all 379 items from a large sample of parents with children between three and eight years of age. Our goal was to have at least 100 complete responses per year (e.g. 100 parents with children between 3.0 and 4.0). This data would then be used to estimate item parameters to be used during the selection process.

Participants.   Participants were recruited via a database of children whose caregivers indicated an interest in participating in studies on child development and who additionally signed up for online studies. All children lived in Leipzig, Germany, an urban Central-European city with approximately 600,000 inhabitants. The city-wide median

Table 1

Participants per age group

and sex

Age group	N	female
3 - 4 years	176	82
4 - 5 years	191	84
5 - 6 years	221	113
6 - 7 years	291	142
7 - 8 years	308	148
> 8 years	3	1

Note. Children in the > 8 years group were very close to 8, see Figure 2C

individual monthly net income in 2021 was ~ 1,600€. Children growing up in Leipzig mostly live in nuclear two-generational families. Socioeconomic status was not formally recorded, although the majority of families in the database come from mid to high socioeconomic backgrounds with high levels of parental education. In addition, it is very likely that the online format caused selective responding and skewed the sample towards highly motivated and interested families. Caregivers received an email with a personalized link to the study. Approximately one week after the first email, they received a reminder if they had not yet finished the study. We contacted caregivers of 4094 children; caregivers of 1826 children started the study of which 1190 (29.00 %) completed all 379 items. All subsequent analyses are based only on the complete data. Table 1 shows the age and sex distribution of participants

²¹⁴   **Descriptive results.**   Figure 2 visualizes the results. On an item level, we saw

²¹⁵ strong negative correlations between caregiver's responses and rated ages-of-acquisition.

²¹⁶ The less likely caregiver's were to say their child says a word, the higher was the rated

²¹⁷ AoA. This relation was the same for nouns, verbs and adjectives. On a child-level, we saw

²¹⁸ that the older children were, the more words they used according to their caregiver's.

²¹⁹ These results reflect highly expected patterns and served as a sanity check for the design
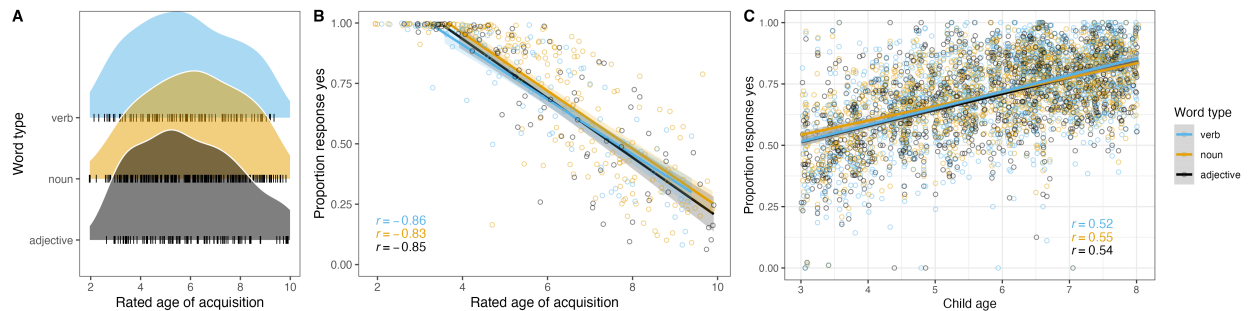
²²⁰ and implementation of the task.

Figure 2. Initial item pool. (A) Distribution of items across word types and rated age-of-acquisition. (B) Item-based correlation between rated age-of-acquisition and caregiver responses by word type averaged across participants. (C) Child-based correlation between age and caregiver responses by word type averaged across items.

Item selection

²²² The goal of the item selection procedure was to generate an item pool with items

²²³ that fit the Rasch model. We selected items in three steps. In the first step, we excluded

²²⁴ items using conventional cut-offs for indices that quantify the fir of each item to the Rasch

²²⁵ model. The goal was to remove a large number of items with a poor fit to reduce the

²²⁶ computational burden in subsequent steps. In step 2, we used an automated item selection

²²⁷ procedure to select an optimal subset of items from the remaining pool. We focused on the

²²⁸ fit of the Rasch model as well as variation in item difficulty (to measure in different regions

²²⁹ of the latent dimension). Finally, in step 3, we submitted the items selected in step 2 to an

²³⁰ analysis of differential item functioning (DIF).

²³¹   IRT-models were implemented in a Bayesian framework in `R` using the `brms` package
²³² (Bürkner, 2017, 2019) unless otherwise stated. We predicted the probability of a correct
²³³ answer based on a participant's latent ability and item characteristics. We fit two classes of
²³⁴ models: Rasch (1PL) Birnbaum (2PL) models. The main difference between these two
²³⁵ models lies in their assumption about how the probability of solving an item changes with
²³⁶ ability levels (item discrimination). Here, the Rasch model assumes that the rate of change
²³⁷ (i.e. the slope of the logistic curve) is the same for all items while the 2PL model allows
²³⁸ item discrimination parameters to vary between items.

²³⁹   **Step 1: Fit-based item selection.** We fit a Rasch Model to the data and
²⁴⁰ computed In- and Outfit values based on draws from the model's posterior predictive
²⁴¹ distribution (using the function `add_epred_draws`) for each item and person combination.
²⁴² In- and Outfit quantify the deviation of a person's response to an item from what the
²⁴³ model predicts based on the difficulty of an item and the person's ability parameter. The
²⁴⁴ result was a distribution of values for each item and index. For each item, we then
²⁴⁵ computed the mode for each index. The closer to 1 the index is, the better the fit. Figure 3
²⁴⁶ visualizes the results. We used the cut-off values suggested in the literature (Bond & Fox,
²⁴⁷ 2013; Debelak, Strobl, & Zeigenfuse, 2022) and excluded items with In- or Outfit values
²⁴⁸ below 0.7 and above 1.3. Like all heuristics, these cut-offs are to some extent arbitrary.
²⁴⁹ Yet, in the present context they served the purpose of removing a large number of
²⁵⁰ potentially unsuitable items. This procedure led us to exclude 167 of the 379 items in the
²⁵¹ pool, leaving 212 for the automated item selection.

²⁵²   **Step 2: Automated item selection.** The goal of this step was to select items
²⁵³ with different levels of difficulty that fit the Rasch model. Selecting items based on these
²⁵⁴ criteria ensured that a) the final item pool allowed for precise measurement in different
²⁵⁵ regions of the latent ability and b) the number of solved items is a sufficient statistic for an
²⁵⁶ individual's ability. Such an item pool is then optimally suited for adaptive testing.

First we defined an objective function that reflected the selection criteria which would later be used in the automated selection process. Items should vary in their difficulty; we quantified this requirement as the standard deviation of the distance (in difficulty estimates) between adjacent items. Lower values indicate smaller distances and thus an overall more equal spacing. Items should also fit the Rasch model; we quantified this requirement in three ways. First and second, we used the In- and Outfit values for each item computed in the previous step. Third, we computed modification indices for each item. For this, we re-fitted the Rasch model using the package `lavaan` (Rosseel, 2012) and used the function `modindices` to obtain modification indices. Broadly speaking, modification indices quantify the improvement in model fit (in terms of the chi-square test statistic) when an item would be dropped.

The objective function was the sum of these four components. Before summation, we multiplied the different components by constants to bring them on a comparable scale and to emphasize certain components over others: the standard deviation for item difficulties was multiplied by -1/3, Infit values by -4, Outfit values by -2 and modification indices by -1/100. The resulting score was always negative so that larger individual values led to more negative values. Because the process described below aims to maximize the score, this meant minimizing the individual values.

Following Bohn et al. (2023), we employed simulated annealing (Kirkpatrick, Gelatt Jr, & Vecchi, 1983) as a method to identify the most optimal items for any given subset size. The process involves systematically exploring the vast space of possible subsets, commencing from a randomly selected initial subset. Subsequently, small random changes are proposed by exchanging some items within the subset under consideration with others located outside it. If a proposed change leads to an improvement in the objective function's value, the proposal is accepted, and the enhanced subset becomes the starting point for subsequent proposals.

To prevent the process from becoming trapped in local optima, it probabilistically accepts proposals that decrease the value of the objective function. The probability of accepting a proposal that reduces the objective function is influenced by a parameter known as "temperature," which gradually decreases from an initially high value to a lower value during the simulation. In the early "hot" phase, the process explores the search space more freely, accepting decreasing proposals often enough to enable movement between local optima separated by less effective subsets, facilitating the discovery of global optima. As the simulation progresses into the later "cool" phases, the process converges towards a more focused "hill climbing" search, where only increasing proposals are accepted. This fine-tunes the best subset discovered during the hot phase, resulting in a more refined and optimized solution.

The simulated annealing algorithm finds the optimal items for a given size of the subset but does not answer the question of what the optimal size is. To answer this question, we applied the algorithm to subsets of different sizes. Our goal was to find the largest subset for which the Rasch model provided a good fit. For each size, we therefore compared the fit of a Rasch model to a 2PL model using Bayesian approximate leave-one-out cross-validation (Vehtari, Gelman, & Gabry, 2017) based on differences in expected log posterior density (ELPD) estimates and the associated standard error (SE). Based on suggestions in the literature (Sivula, Magnusson, & Vehtari, 2020), we considered models to be equivalent up to a point when the ELPD in favor of a model exceeded two times the standard error of the difference.

Figure 3B visualizes the model comparison and shows that the Rasch model provided a good fit for subsets up to 90 items. We therefore decided on 90 items as the size of the final item pool. We ran the simulated annealing algorithm 20 times and selected the 90 items that were returned most often (86 items were returned on every run).

Step 3: Differential item functioning.   The final step of item selection consisted of assessing differential item functioning (DIF, see Bürkner, 2019). DIF describes a

situation when items show differential characteristics for subgroups that otherwise have the

same overall score (Holland & Wainer, 2012). We assessed DIF based on sex (male and

female). We estimated separate item parameters for the two groups and assessed whether

their 95% CrI overlapped. Figure 3C shows that the item parameters were very similar in

the two subgroups. However, one item ("verloben", en: to get engaged) had to be

excluded. Thus, the size of the final item pool was 89 items, 43 (48%) of which were nouns,

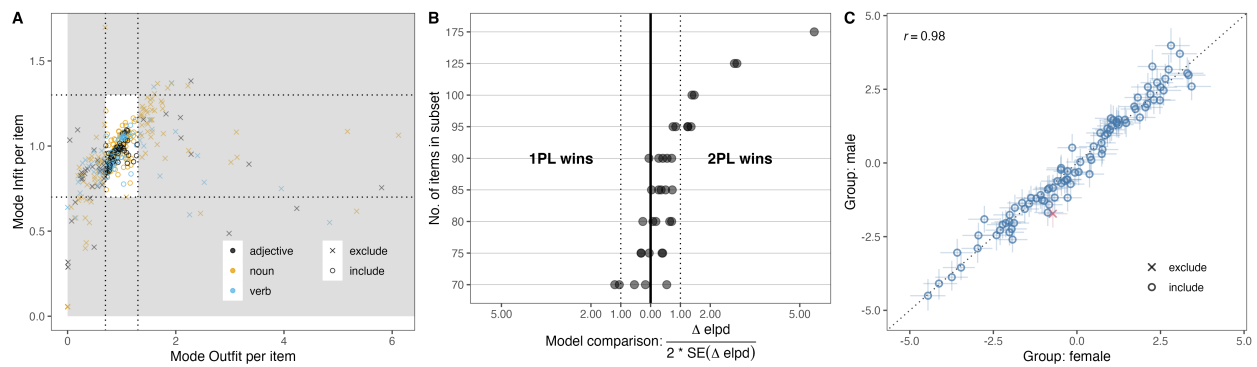20 (22%) were verbs and 26 (29%) were adjectives.

Figure 3. Three steps of item selection. (A) In- and Outfit values for all 379 items in the
initial item pool. Items that fell into the grey region (crosses) were excluded. Color shows the
different word types. Dashed lines show cut-off values of 0.7 and 1.3. (B) Model comparison
ratio comparing the fit of a Rasch model to the fit of a 2PL model for different numbers
of items (y-axis). Each point shows an independent run of the item selection procedure
and subsequent model comparison (five per subset). The x-axis title shows how the ratio is
computed. Values left of 0 indicate a better fit of the Rasch model, values to the right a
better fit of the 2PL model. The dashed line marks a ratio of 1, which we assumed to be the
point when one of the models clearly provided a better fit. (C) Correlation between item
parameters estimated separately by sex. Points show the mode of the posterior distribution
for each item with 95% CrIs. Point color and shape denote items that were excluded.

Psychometric properties of the final item pool

The final item pool consisted of 89 items of varying difficulty that fit the Rasch model (see Figure 4A). Next, we investigated the reliability and convergent validity of a task including the full item pool (PREVIC).

Reliability. We computed KR-20 (Kuder & Richardson, 1937) and Andrich Reliability (Andrich, 1982). Both indices indicated excellent reliability (KR-20 = 0.97; Andrich = 0.97).

Convergent validity. We assessed convergent validity by comparing PREVIC scores to a direct assessment of children's receptive vocabulary using the oREV (Bohn et al., 2023). The oREV asks children to select a picture (out of four) upon hearing a word. It has 22 items which fit the Rasch model. Because the oREV is also available as a web application, we sent out emails to all caregivers who provided complete data in the data collection that led to the construction of the PREVIC (N = 1190) and asked them to have their child complete the oREV. We obtained oREV data from 692 children (337 female, m_{age} = 5.78, range = 3.02 - 8.00) which corresponds to a response rate of ~ 58%.

We found a substantial correlation between caregiver's answers to questions about their children's expressive vocabulary in the PREVIC and a direct assessment of children's receptive vocabulary in the oREV (r = 0.54; 95% CI = 0.48 – 0.59; Figure 4B)). Furthermore, when predicting PREVIC scores by oREV scores and age in a binomial model, we found that oREV scores had a large and positive effect (Figure 4C). These results speak to the convergent validity of the PREVIC.

Adaptive testing

The large and diverse item pool allowed us to to create an adaptive version of the PREVIC in addition to the complete checklist. The general idea of an adaptive test is to only show the caregiver the most informative items given the (continuously updated)
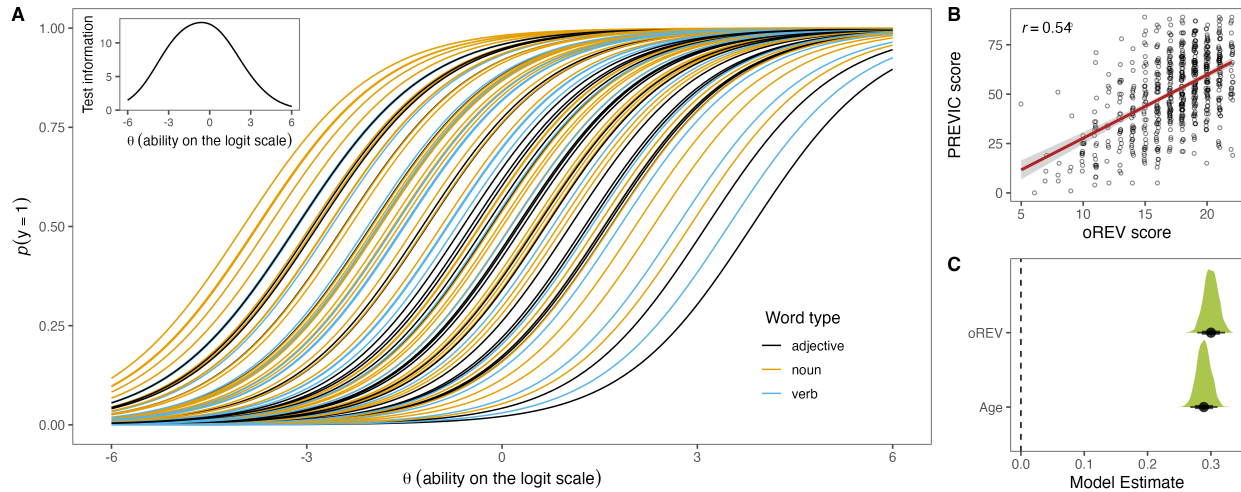
Figure 4. Item characteristics and validity. (A) Item characteristic curves for the 89 items colored by word type. Inset on the upper left shows the test information curve. (B) Correlation between PREVIC and oREV scores.Points show aggregated scores of individuals in the two tasks. Points have minimal horizontal noise added to avoid overplotting. The red line shows a regression line (with 95% CI) based on a linear model. (C) Posterior model estimates for oREV scores and age (scaled) in a model predicting PREVIC scores. Points show posterior means with 95% CrI.

₃₄₂ ability. As a consequence, items that are too easy or too difficult are omitted and the test

₃₄₃ becomes substantially shorter while retaining the same level of measurement precision.

₃₄₄   In order to determine the most informative items, the ability of the child has to be

₃₄₅ estimated during the test. To achieve this, we implemented a maximum likelihood

₃₄₆ estimator in `html` and `TypeScript` (which is compiled to native `JavaScript`). As a

₃₄₇ consequence, the adaptive version is still fully portable and can be run in any modern

₃₄₈ web-browser. The estimated ability "is the ability value that maximizes the likelihood

₃₄₉ function $L(\theta)$" (Magis & Raîche, 2012) given the item response y_i (either 0 or 1) and the

₃₅₀ item difficulty α_i (Eid, 2011; Eid & Schmidt, 2014):

$$L(\theta) = \prod_{i=1}^{p} \frac{\exp^{y_i*(\theta-\alpha_i)}}{1 + \exp^{\theta-\alpha_i}}$$

351 The maximum likelihood estimation is implemented using a line search algorithm that

352 converges when the maximum of the likelihood distribution has been reached. Based on

353 the estimated ability, the program will then select the next item from the pool so that the

354 difficulty is nearest to the current ability level (Urry, 1970). This procedure is equivalent to

355 selecting items with the maximum information criterion when using a Rasch model (Magis

356 & Raîche, 2012).

357   At the beginning of the test, the ability level is set to 0. A person specific ability

358 estimation is also not possible using Maximum Likelihood after the first item and we

359 therefore followed the convention to set the ability level to -4 if the answer was "no" and 4

360 if the answer was "yes" (e.g. implemented in the R package catR, Magis & Barrada, 2017).

361 The test then continues until a pre-specified level of measurement precision (standard error

362 of the ability estimate) is reached or until all items have been used. Users can set the

363 desired level of measurement precision at the beginning of the test (e.g. SE of 0.3, 0.4 or

364 0.5), which again influences its length (larger SE means shorter test). In the end, the user

365 downloads a file containing the final ability level, the SE at which the test stopped, the

366 answered items (including the word itself and its difficulty) and the participant's response

367 pattern.

368   We validated the implementation of our estimator by comparing its ability estimates

369 and selected items to those of the catR package in a number of simulations. The results

370 were identical and only differed beyond the fifth decimal because JavaScript and R differ

371 in how many decimal places they use for computations. The code to run the simulations

372 can be found in the associated online repository.

Discussion

This paper describes the construction and validation of the PREVIC, an adaptive parent report measure of productive vocabulary in German-speaking children between three and eight years of age. Following the logic of widely-used vocabulary checklists for younger children (Fenson et al., 2007), the PREVIC presents caregivers with individual words and asks if the child speaks this word. The items (words) that make up the PREVIC were selected using Item-Response theory (IRT): we started with a large initial item pool of 379 words from which we selected 89 items that fit the Rasch model. The resulting task is highly reliable and shows convergent validity when contrasted with a direct receptive vocabulary measure. Leveraging IRT allowed us to devise an adaptive version of the task in which only the most informative items are presented. The task is implemented as a web-app and can be used with any device that runs a modern web browser. The task itself (adaptive and complete checklist) as well as the source code are freely available online.

The PREVIC fills an important gap in the tool kit of researchers studying language development beyond infancy. It complements direct assessments of children's vocabulary by providing an additional perspective on children's vocabulary skills. Parents observe children for extended periods of time and their assessment therefore provides a more aggregated measure. Parent reports are also immune to momentary fluctuations in children's motivation and attention that might influence the results of direct assessments. Nevertheless, parental reports remain indirect measures and are ideally combined with direct assessments whenever feasible. Given that the PREVIC is short – in particular the adaptive version rarely takes more than five minutes to complete – it can easily be filled out by parents during a lab visit. Its implementation as a web-app even allows for sending it to families before or after a visit.

At present, the PREVIC is only available in German. However, with inclusivity and broad applicability in mind, we have made the entire source code available. This not only

facilitates its adaptation to other languages and allows researchers to use the same user

interface. Encouragingly, preliminary feedback indicates that parents find the interface

intuitive and user-friendly. The CDI has seen expansive adaptation across various

languages (see Frank et al., 2021 for a summary). Such adaptations are usually not

complete translations in that some words are removed and others added to capture the

linguistic nuances and specificities of each language. For the adaptation of the PREVIC to

new languages, a comprehensive reassessment of item properties would be highly desirable.

For adaptive testing, it would even be mandatory. Taken together, we hope that our

commitment to openness will put the PREVIC on a similar trajectory as the CDI.

Limitations

The sample we tested was not a representative sample: It only contained families

living in Leipzig, Germany who volunteered to participate in research on child development

and who additionally indicated that they were interested in participating in online studies.

These multiple steps of (self-)selection most likely skewed the sample to more affluent and

educated parents, though we have no demographic data to assess this claim. We think the

most likely consequence is that the variation in our sample was reduced compared to the

general population and that the probability of knowing a particular word would be

somewhat lower in a representative sample. The data we collected during the construction

of the PREVIC should therefore not be seen as a normative data set. Instead, the

PREVIC is first and foremost a research tool that can be used to measure variation in

receptive vocabulary in a given sample.

Conclusion

We designed the PREVIC with a commitment to psychometric rigor; its grounding in

Item-Response Theory provides a clear measurement model and specifies how individual

items relate to each other and the underlying psychological ability. This approach not only

⁴²⁴ strengthens the PREVIC's validity in assessing receptive vocabulary but also serves as a

⁴²⁵ methodological reference for developing tests in other areas. By making the PREVIC

⁴²⁶ openly accessible, we actively contribute to the collective resource pool for researchers in

⁴²⁷ language development, ensuring that they have another reliable tool at their disposal.

⁴²⁸   **Open Practices Statement**

⁴²⁹   The task can be accessed via the following website:

⁴³⁰ https://ccp-odc.eva.mpg.de/previc/. The corresponding source code can be found in the

⁴³¹ following repository: https://github.com/ccp-eva/previc-demo. The data sets generated

⁴³² during and/or analysed during the current study are available in the following repository:

⁴³³ https://github.com/ccp-eva/previc-analysis/. Data collection was preregistered at:

⁴³⁴ https://osf.io/utzfh.

References

Andrich, D. (1982). An index of person separation in latent trait theory, the traditional KR. 20 index, and the guttman scale response pattern. *Education Research and Perspectives*, *9*(1), 95–104.

Bayley, N. (2006). *Bayley scales of infant and toddler development–third edition.* San Antonio, TX: Harcourt Assessment.

Birchenough, J. M., Davies, R., & Connelly, V. (2017). Rated age-of-acquisition norms for over 3,200 german words. *Behavior Research Methods*, *49*(2), 484–501.

Birnbaum, A. (1986). Test scores, sufficient statistics, and the information structures of tests. In F. M. L. & M. R. Novick (Ed.), *Statistical theories of mental test scores.* Addison & Wesley.

Bleses, D., Makransky, G., Dale, P. S., Højen, A., & Ari, B. A. (2016). Early productive vocabulary predicts academic achievement 10 years later. *Applied Psycholinguistics*, *37*(6), 1461–1476.

Bodnarchuk, J. L., & Eaton, W. O. (2004). Can parent reports be trusted?: Validity of daily checklists of gross motor milestone attainment. *Journal of Applied Developmental Psychology*, *25*(4), 481–490.

Bohn, M., Prein, J., Koch, T., Bee, R. M., Delikaya, B., Haun, D., & Gagarina, N. (2023). oREV: An item response theory-based open receptive vocabulary task for 3-to 8-year-old children. *Behavior Research Methods*, 1–11.

Bohn, M., Tessler, M. H., Merrick, M., & Frank, M. C. (2021). How young children integrate information sources to infer the meaning of words. *Nature Human Behaviour*, *5*(8), 1046–1054.

Bond, T. G., & Fox, C. M. (2013). *Applying the rasch model: Fundamental measurement in the human sciences.* Psychology Press.

Bornstein, M. H., Hahn, C.-S., Putnick, D. L., & Pearson, R. M. (2018). Stability of core language skill from infancy to adolescence in typical and atypical development. *Science*

462   *Advances*, *4*(11), eaat7422.

463   Bornstein, M. H., & Haynes, O. M. (1998). Vocabulary competence in early childhood:

464   Measurement, latent construct, and predictive validity. *Child Development*, *69*(3),

465   654–671.

466   Borsboom, D. (2006). The attack of the psychometricians. *Psychometrika*, *71*(3), 425–440.

467   Bürkner, P.-C. (2017). Brms: An r package for bayesian multilevel models using stan.

468   *Journal of Statistical Software*, *80*(1), 1–28.

469   Bürkner, P.-C. (2019). Bayesian item response modeling in r with brms and stan. *arXiv*

470   *Preprint arXiv:1905.09501*.

471   Dale, P. S. (1991). The validity of a parent report measure of vocabulary and syntax at 24

472   months. *Journal of Speech, Language, and Hearing Research*, *34*(3), 565–571.

473   Debelak, R., Strobl, C., & Zeigenfuse, M. D. (2022). *An introduction to the rasch model*

474   *with examples in r.* Crc Press.

475   DeMayo, B., Kellier, D., Braginsky, M., Bergmann, C., Hendriks, C., Rowland, C. F., . . .

476   Marchman, V. (2021). Web-CDI: A system for online administration of the

477   MacArthur-bates communicative development inventories. *Language Development*

478   *Research.*

479   Diamond, K. E., & Squires, J. (1993). The role of parental report in the screening and

480   assessment of young children. *Journal of Early Intervention*, *17*(2), 107–115.

481   Dunn, L. M., & Dunn, L. M. (1965). *Peabody picture vocabulary test.*

482   Dunn, L. M., Dunn, L. M., Whetton, C., & Burley, J. (1997). British picture vocabulary

483   scale 2nd edition (BPVS-II). *Windsor, Berks: NFER-Nelson.*

484   Eid, M. (2011). *Testtheorie* (Erscheint: ET unbestimmt). Göttingen: Hogrefe.

485   Eid, M., & Schmidt, K. (2014). *Testtheorie und testkonstruktion.* Hogrefe Verlag GmbH &

486   Company KG.

487   Feldman, H. M., Dale, P. S., Campbell, T. F., Colborn, D. K., Kurs-Lasky, M., Rockette,

488   H. E., & Paradise, J. L. (2005). Concurrent and predictive validity of parent reports of

489   child language at ages 2 and 3 years. *Child Development*, *76*(4), 856–868.

490   Fenson, L. et al. (2007). *MacArthur-bates communicative development inventories*. Paul H.

491   Brookes Publishing Company Baltimore, MD.

492   Fenson, L., Dale, P. S., Reznick, J. S., Bates, E., Thal, D. J., Pethick, S. J., . . . Stiles, J.

493   (1994). Variability in early communicative development. *Monographs of the Society for*

494   *Research in Child Development*, i–185.

495   Frank, M. C., Braginsky, M., Yurovsky, D., & Marchman, V. A. (2017). Wordbank: An

496   open repository for developmental vocabulary data. *Journal of Child Language*, *44*(3),

497   677–694.

498   Frank, M. C., Braginsky, M., Yurovsky, D., & Marchman, V. A. (2021). *Variability and*

499   *consistency in early language learning: The wordbank project*. MIT Press.

500   Gershon, R. C., Slotkin, J., Manly, J. J., Blitz, D. L., Beaumont, J. L., Schnipke, D., et

501   al.others. (2013). IV. NIH toolbox cognition battery (CB): Measuring language

502   (vocabulary comprehension and reading decoding). *Monographs of the Society for*

503   *Research in Child Development*, *78*(4), 49–69.

504   Glück, C. W., & Glück, C. W. (2011). *Wortschatz-und wortfindungstest für 6-bis 10-jährige*

505   *(WWT 6-10)*. Urban & Fischer.

506   Golinkoff, R. M., De Villiers, J. G., Hirsh-Pasek, K., Iglesias, A., Wilson, M. S., Morini, G.,

507   & Brezack, N. (2017). *User's manual for the quick interactive language screener*

508   *(QUILS): A measure of vocabulary, syntax, and language acquisition skills in young*

509   *children*. Paul H. Brookes Publishing Company.

510   Golinkoff, R. M., Hoff, E., Rowe, M. L., Tamis-LeMonda, C. S., & Hirsh-Pasek, K. (2019).

511   Language matters: Denying the existence of the 30-million-word gap has serious

512   consequences. *Child Development*, *90*(3), 985–992.

513   Holland, P. W., & Wainer, H. (2012). *Differential item functioning*. Routledge.

514   Hornman, J., Kerstjens, J. M., Winter, A. F. de, Bos, A. F., & Reijneveld, S. A. (2013).

515   Validity and internal consistency of the ages and stages questionnaire 60-month version

516   and the effect of three scoring methods. *Early Human Development*, *89*(12), 1011–1015.

517   Ireton, H., & Glascoe, F. P. (1995). Assessin children's development using parents' reports:

518   The child development inventory. *Clinical Pediatrics*, *34*(5), 248–255.

519   Jørgensen, R. N., Dale, P. S., Bleses, D., & Fenson, L. (2010). CLEX: A cross-linguistic

520   lexical norms database. *Journal of Child Language*, *37*(2), 419–428.

521   Kauschke, C., & Siegmüller, J. (2002). *Patholinguistische diagnostik bei*

522   *sprachentwicklungsstörungen: Diagnostikband phonologie.* Urban & Fischer.

523   Kiese-Himmel, C. (2005). AWST-r-aktiver wortschatztest für 3-bis 5-jährige kinder

524   (AWST-r–active vocabulary test for 3-to 5-year-old children). *Göttingen: Hogrefe.*

525   Kirkpatrick, S., Gelatt Jr, C. D., & Vecchi, M. P. (1983). Optimization by simulated

526   annealing. *Science*, *220*(4598), 671–680.

527   Kubinger, K. D. (2006). *Psychologische diagnostik: Theorie und praxis psychologischen*

528   *diagnostizierens.* Hogrefe Verlag.

529   Kuder, G. F., & Richardson, M. W. (1937). The theory of the estimation of test reliability.

530   *Psychometrika*, *2*(3), 151–160.

531   Lenhard, A., Lenhard, W., Segerer, R., & Suggate, S. (2015). *Peabody picture vocabulary*

532   *test-4. Ausgabe: Deutsche fassung.* Frankfurt am Main: Pearson Assessment.

533   Libertus, M. E., Odic, D., Feigenson, L., & Halberda, J. (2015). A developmental

534   vocabulary assessment for parents (DVAP): Validating parental report of vocabulary

535   size in 2-to 7-year-old children. *Journal of Cognition and Development*, *16*(3), 442–454.

536   Lord, F. M. (2012). *Applications of item response theory to practical testing problems.*

537   Routledge.

538   Łuniewska, M., Wodniecka, Z., Miller, C. A., Smolik, F., Butcher, M., Chondrogianni, V.,

539   et al.others. (2019). Age of acquisition of 299 words in seven languages: American

540   english, czech, gaelic, lebanese arabic, malay, persian and western armenian. *PloS One*,

541   *14*(8), e0220611.

542   Macy, M. (2012). The evidence behind developmental screening instruments. *Infants &*

Young Children, 25(1), 19–61.

Magis, D., & Barrada, J. R. (2017). Computerized adaptive testing with R: Recent updates of the package catR. *Journal of Statistical Software, Code Snippets, 76*(1), 1–19. https://doi.org/10.18637/jss.v076.c01

Magis, D., & Raîche, G. (2012). Random Generation of Response Patterns under Computerized Adaptive Testing with the R Package catR. *Journal of Statistical Software, 48*(8). https://doi.org/10.18637/jss.v048.i08

Makransky, G., Dale, P. S., Havmose, P., & Bleses, D. (2016). An item response theory–based, computerized adaptive testing version of the MacArthur–bates communicative development inventory: Words & sentences (CDI: WS). *Journal of Speech, Language, and Hearing Research, 59*(2), 281–289.

Marchman, V. A., & Fernald, A. (2008). Speed of word recognition and vocabulary knowledge in infancy predict cognitive and language outcomes in later childhood. *Developmental Science, 11*(3), F9–F16.

Mayor, J., & Mani, N. (2019). A short version of the MacArthur–bates communicative development inventories with high validity. *Behavior Research Methods, 51*(5), 2248–2255.

Morgan, P. L., Farkas, G., Hillemeier, M. M., Hammer, C. S., & Maczuga, S. (2015). 24-month-old children with larger oral vocabularies display greater academic and behavioral functioning at kindergarten entry. *Child Development, 86*(5), 1351–1370.

Morsbach, S. K., & Prinz, R. J. (2006). Understanding and improving the validity of self-report of parenting. *Clinical Child and Family Psychology Review, 9*, 1–21.

Pace, A., Alper, R., Burchinal, M. R., Golinkoff, R. M., & Hirsh-Pasek, K. (2019). Measuring success: Within and cross-domain predictors of academic and social trajectories in elementary school. *Early Childhood Research Quarterly, 46*, 112–125.

Pace, A., Luo, R., Hirsh-Pasek, K., & Golinkoff, R. M. (2017). Identifying pathways between socioeconomic status and language development. *Annual Review of*

⁵⁷⁰ *Linguistics*, *3*, 285–308.

⁵⁷¹ Pontoppidan, M., Niss, N. K., Pejtersen, J. H., Julian, M. M., & Væver, M. S. (2017).

⁵⁷² Parent report measures of infant and toddler social-emotional development: A

⁵⁷³ systematic review. *Family Practice*, *34*(2), 127–137.

⁵⁷⁴ Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests.*

⁵⁷⁵ University of Chicago Press.

⁵⁷⁶ Rosseel, Y. (2012). Lavaan: An r package for structural equation modeling. *Journal of*

⁵⁷⁷ *Statistical Software*, *48*, 1–36.

⁵⁷⁸ Saudino, K. J., Dale, P. S., Oliver, B., Petrill, S. A., Richardson, V., Rutter, M., . . .

⁵⁷⁹ Plomin, R. (1998). The validity of parent-based assessment of the cognitive abilities of

⁵⁸⁰ 2-year-olds. *British Journal of Developmental Psychology*, *16*(3), 349–362.

⁵⁸¹ Schoon, I., Parsons, S., Rush, R., & Law, J. (2010). Children's language ability and

⁵⁸² psychosocial development: A 29-year follow-up study. *Pediatrics*, *126*(1), e73–e80.

⁵⁸³ Schröder, A., Gemballa, T., Ruppin, S., & Wartenburger, I. (2012). German norms for

⁵⁸⁴ semantic typicality, age of acquisition, and concept familiarity. *Behavior Research*

⁵⁸⁵ *Methods*, *44*(2), 380–394.

⁵⁸⁶ Sivula, T., Magnusson, M., & Vehtari, A. (2020). Uncertainty in bayesian leave-one-out

⁵⁸⁷ cross-validation based model comparison. *arXiv Preprint arXiv:2008.10296*.

⁵⁸⁸ Squires, J., Bricker, D. D., Twombly, E., et al. (2009). *Ages & stages questionnaires*. Paul

⁵⁸⁹ H. Brookes Baltimore, MD.

⁵⁹⁰ Szagun, G., Stumper, B., & Schramm, S. A. (2009). *Fragebogen zur frühkindlichen*

⁵⁹¹ *sprachentwicklung (FRAKIS) und FRAKIS-k (kurzform)*. Universitätsverlag Potsdam.

⁵⁹² Urry, V. W. (1970). *A monte carlo investigation of logistic mental test models* (PhD

⁵⁹³ thesis). United States – Indiana. Retrieved from

⁵⁹⁴ https://www.proquest.com/docview/302519686/citation/F24B5FB611144881PQ/1

⁵⁹⁵ Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical bayesian model evaluation using

⁵⁹⁶ leave-one-out cross-validation and WAIC. *Statistics and Computing*, *27*(5), 1413–1432.

597 Walker, D., Greenwood, C., Hart, B., & Carta, J. (1994). Prediction of school outcomes

598   based on early language production and socioeconomic factors. *Child Development*,

599   *65*(2), 606–621.

600 Wechsler, D., & Kodama, H. (1949). *Wechsler intelligence scale for children* (Vol. 1).

601   Psychological corporation New York.