The latent factor structure of developmental change in early childhood

Benjamin A. Stenhaug (stenhaug@stanford.edu)
School of Education, 485 Lasuen Mall
Stanford, CA 94305 USA

Michael C. Frank (mcfrank@stanford.edu)
Department of Psychology, 450 Jane Stanford Way
Stanford, CA 94305 USA

Abstract

Piaget proposed that development proceeded in stages; more recently researchers have proposed modular theories in which different abilities develop on their own timetable. Despite the abundance of theory, there is little empirical work on the structure of developmental changes in early childhood. We investigate this question using a large dataset of parent-reported developmental milestones. We compare a variety of factor-analytic item response theory models and find that variation in development from birth to 55 months of age is best described by a model with three distinct dimensions. We also find evidence that dimensionality increases across age, with the youngest children described by a two-factor model. These results provide a model-based method for linking holistic descriptions of early development to basic theoretical questions about the nature of change in childhood.

Keywords: child development; milestones; item response theory; model comparison

Introduction

How do young children grow and change? Is child development a single unified process or a host of different processes, each with their own constraints and timescale? Piaget famously proposed a stage theory in which many seemingly distinct mental processes developed in concert through the operation of the same principles across domains (Flavell, 1963). In contrast, modern theories propose that different facets of children's mental life develop on their own timetable (Gelman & Meck, 1983). And the grandmother of one author of this paper was known to assert that developmental milestones were in compensatory relationships with one another ("children either walk early or else they talk early").

This question is important not only from a theoretical perspective but also for application. The process of assessing children's developmental status critically depends on our assumptions about the nature of that status — in particular, whether there is a single unified process that can be measured via some score derived from subprocesses. In this sense, questions about the nature and structure of development are psychometric questions (Borsboom, 2005). Such psychometric analysis investigating the dimensionality of change has been studied extensively in the case of cognitive aging (e.g., Balinsky, 1941; Li, Nuttall, & Zhao, 1999) but has received less attention in early childhood.

Our goal is to explore the psychometric structure of development. We take as our starting point the idea that psychometric models can instantiate hypotheses about psychological structure in ways that can be assessed via their fit to data. We adopt the framework of item response theory (IRT). IRT models allow us to capture how responses to such questions track both with individual children's abilities as well as with the measurement properties of the questions (and underlying milestones). In particular, our interest is in comparing within a family of multidimensional IRT models in order to gain insight into the underlying dimensionality of early childhood development.

In a standard factor-analytic approach (which multidimensional IRT extends), a solution with N factors partitions observed variance into factors, suggesting dimensions of variation in the sample. One substantial complication to this perspective for analyzing developmental data is the issue that the dimensionality of children's variation could itself change developmentally. Indeed, the dedifferentiation hypothesis of cognitive aging — that distinct factors collapse — is such a hypothesis (Frias, Lövdén, Lindenberger, & Nilsson, 2007). To address this challenge, we use a new set of cross-validation methods to investigate changes in dimensionality.

We use milestone data for our investigation. Global assessment of developmental status via a series of binary questions (e.g., "Can your child walk at least ten steps unassisted?") is both a standard feature of pediatrician visits (Sheldrick et al., 2019) and a gold standard for child development in the research and intervention communities (Bayley, 2009; Bricker et al., 1999; McCoy, Gonzalez, & Jones, 2019). In such assessments, which are typically but not always conducted via parent report, developmental progress is pooled across domains like motor development or language. Thus, these instruments implicitly assume a unifactorial model, although some also provide subscale scores (Bayley, 2009).

Unfortunately, these instruments are commercial products, and hence normative data at the item level are typically not available for analysis. In the current paper, we thus analyze a new set of data from a set of 414 milestone questions administered online to a group of 1946 middle-class Mexican parents of children from 0 to 55 months of age. This very comprehensive milestone set allows us to ask questions about how variation in developmental growth can be partitioned across age and face-valid domains (language, cognition, motor, and socio-emotional development).

We first describe our dataset. We then introduce the family of item response models that we use and the way in which

we compare performance across these models. These models allow us to consider the overall dimensionality of our dataset, which we then follow up on by looking for evidence of change in dimensionality across development. We end by considering the limitations, implications, and next steps for this work.

Data

A child's development can be thought of as the set of developmental milestones that they have reached at a particular point in time. This conceptualization results in data with the same structure as the item response data common to educational measurement. In education, item response data is most typically students responding to test items (i.e., questions) and, in the dichotomous case, getting each question either correct or incorrect. In the context of child development, the child is the "student," and each developmental milestone is the "item."

Data were provided by Kinedu, Inc., a developer of parenting applications. We consider the 1946 children between 2 and 55 months of age whose parents responded to all 414 of the developmental milestones. Kinedu, Inc. mapped each milestone to a face-valid group: physical, cognitive, linguistic, or social & emotional. Table 1 shows the number of developmental milestones in each group along with an example milestone from each group translated to English.

Table 1: Developmental milestone groups and examples

Group	Milestones	Example milestone
Physical	180	Stands on their toes
Cognitive	100	Finds objects on the floor
Linguistic	75	Babbles to imitate conversations
Social & Emotional	59	Complains when play is stopped

Figure 1 shows the age (in months) and number of completed milestones for each child. At 12 months old, most children have reached about 200 developmental milestones. At 24 months old, most children have reached about 300 developmental milestones. At 48 months old, most children have reached about 375 of the 414 developmental milestones.

Methods

We frame the assessment of the dimensionality of child development as a model comparison question.

Models

Item response theory offers a suite of models with which to model item response data. We adopt the notation used in Chalmers (2012). Let $i = 1, \ldots, I$ represent the distinct children and $j = 1, \ldots, J$ the developmental milestones. The item response data is stored in a matrix, y, where element y_{ij} denotes if the ith child has or has not achieved the jth developmental milestone as reported by their parent/guardian. Each model represents the ith child's development using m latent factors $\boldsymbol{\theta}_i = (\theta_1, \ldots, \theta_m)$. The jth milestone's discriminations

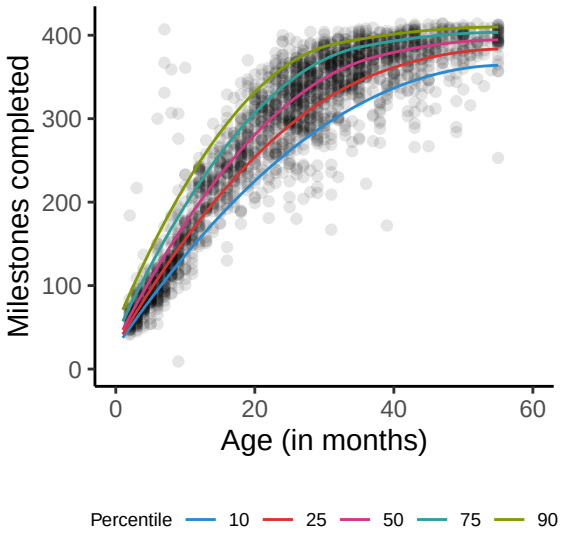

Figure 1: Number of milestones completed by age with percentile curves. Dots represent individual children.

(i.e. slopes) $\boldsymbol{a_j} = (a_1, \ldots, a_m)$ capture the latent factor loadings onto that milestone. We fit five two-parameter logistic (2PL) models where a child's development is represented by $m = 1$, $m = 2$, $m = 3$, $m = 4$ and $m = 5$ latent factors. Hereafter, we, for example, refer to a 2PL model with $m = 4$ latent factors as a 4F 2PL model. As is common, we estimate all models using marginal maximum likelihood estimation (MMLE), which integrates over a generic distribution for $\boldsymbol{\theta}$ and therefore estimates only item parameters (Baker & Kim, 2004). According to the 2PL model, the probability of a child having achieved a developmental milestone is

$$P(y_{ij} = 1 | \boldsymbol{\theta_i}, \boldsymbol{a_j}, b_j) = \sigma(\boldsymbol{a_j}^\top \boldsymbol{\theta_i} + b_j)$$

where b_j is the milestone easiness (i.e. intercept) and $\sigma(x) = \frac{e^x}{e^x+1}$ is the standard logistic function. As an example, Figure 2 shows item characteristic curves from the 1F 2PL model for the items in Table 1. Item characteristic curves show the relationship between θ_i and $P(y_{ij} = 1)$ for a particular item. These curves reveal that babbling is unrelated to development (presumably because parents interpret babbling as including early cooing and hence report that essentially all babies babble). On the other hand, finding objects on the floor is highly related to development with most children with θ_i greater than -1.5 having reached this milestone.

We primarily focus here on the latent factor structure of children's ability, but we also examined the structure of individual item models. While we use a 2PL model here (which includes difficulty and discrimination parameters), we also explored 3PL models (which add a guessing parameter for each item). Overall, 3PL models did not fit better than 2PL models and so we omit them in the interest of space. For comparison, we do include a 1F Rasch model where all of the discrimination parameters, a_j, are set to 1 for each item.

Each of these models learn the latent factor structure entirely from the data, making them exploratory. We also fit a

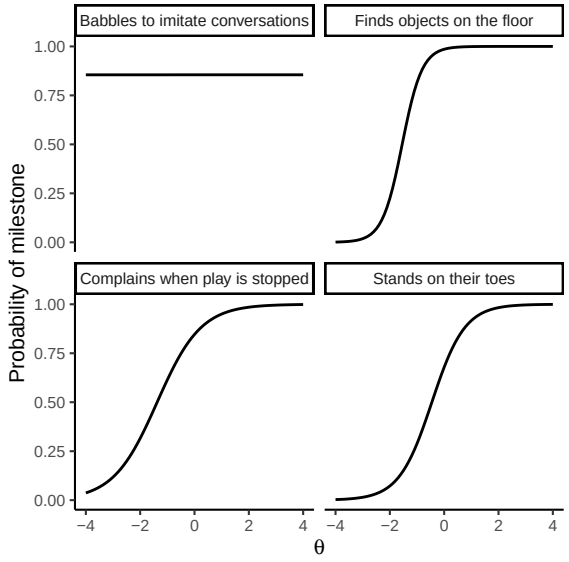

Figure 2: Example item characteristic curves. Babbling is unrelated to a child's development whereas finding objects on the floor is highly related to development.

variety of confirmatory models where milestones are mapped to specific factors according to the four developmental milestone groups shown in Table 1. In the interest of space, we report only the bifactor model, which was the best performing confirmatory model. In the bifactor model, each milestone loads onto a general factor θ_0 and a specific factor θ_s (Cai, Yang, & Hansen, 2011). Accordingly, the probability of a child having achieved a developmental milestone is

$$P(y_{ij} = 1|\theta_0, \theta_s, a_0, a_s) = \sigma(a_0\theta_0 + a_s\theta_s + b_j).$$

Computing is done in R (R Core Team, 2019), model fitting in the R package mirt (Chalmers, 2012), and data wrangling/visualization in the set of R packages known as the tidyverse (Wickham, 2017). Materials to reproduce this paper are available at github.com/stenhaug/kinedu.

Model comparison

Model comparison in IRT typically uses information criterion such as AIC and BIC (Maydeu-Olivares, 2013). However, these methods are not guaranteed to work with modest sample sizes or when the models are misspecified (McDonald & Mok, 1995). Instead, as motivated by Bolt & Lall (2003), we use a marginalized version of cross-validation. In essence, we partition the data into folds based on the children (i.e. the rows of the item response matrix). Then for each fold, we estimate the item parameters using all but that fold, and calculate the likelihood of that fold by integrating over $g(\boldsymbol{\theta})$.

Mathematically — following notation similar to Vehtari, Gelman, & Gabry (2017) — we partition the data into 6 subsets $y^{(k)}$ for $k = 1, \ldots, 6$. Each model is fit separately to each training set $y^{(-k)}$ yielding item parameter estimates which we compactly denote $\Psi_j^{(-k)}$. The predictive (i.e. out-of-sample or cross-validated) likelihood of $y^{(k)}$ is

$$p(y^{(k)}|y^{(-k)}) = \prod_{i \in i^{(k)}}^{I} \int_{\theta} \prod_{j=1}^{J} \hat{\Pr}(y_{ij}^{(k)}|\Psi_j^{(-k)}, \theta) g(\theta) d\theta.$$

The ultimate quantity of interest for each model is the log predictive likelihood for the entire item response matrix, which is defined as

$$\text{lpl } y = \sum_{k=1}^{K} \log p(y^{(k)}|y^{(-k)}).$$

Results

Table 2 shows the number of parameters, the in-sample log likelihood (which necessarily increases with more parameters), and the lpl y defined in the model comparison section. The 3F 2PL model performs best, which is evidence that child development between the ages of 2 and 55 months follows a multidimensional path.

Table 2: Model performance: The 3F 2PL performs best as measured by lpl y

Model	Parameters	log-likelihood (in-sample)	lpl y (out-of-sample)
1F Rasch	415	-254984	-255442
1F 2PL	828	-222073	-223106
2F 2PL	1241	-212896	-214491
Bifactor	1242	-210030	-211682
3F 2PL	1653	-208806	-210961
4F 2PL	2064	-208114	-211023
5F 2PL	2474	-207316	-211036

Understanding the latent factor structure

To understand each of the three factors in the best performing model, we fit the model to the the full dataset. We then estimated the factor loadings (i.e. discriminations or slopes) using a varimax rotation. The varimax rotation results in orthogonal and, therefore, more interpretable factors (Kaiser, 1959). Under the varimax rotation, the first factor explains 41% of the variance, the second factor explains 16% of the variance, and the third factor explains 3% of the variance.

Figure 3 shows the distribution of factor loadings for each group on each of the three factors. The first factor loads mainly on cognitive and linguistic milestones. The second factor is a combination of each of the groups with the strongest loadings on the physical and social & emotional milestones. The third factor mainly loads positively on linguistic milestones and, interestingly, negatively on physical milestones.

We estimate the factor scores for each child using expected a posteriori (EAP) with a three dimensional standard normal distribution as calculated by Gauss-Hermite quadrature with 61 points (Embretson & Reise, 2013). Figure 4 shows the relationship between age and factor score for each factor. The first factor, perhaps unsurprisingly, has a high correlation (r = 0.9) with age. The second factor has a strong association with

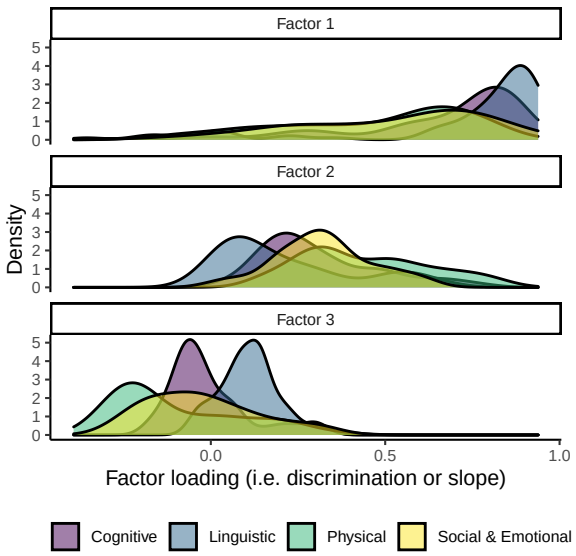

Figure 3: Factor loadings by group

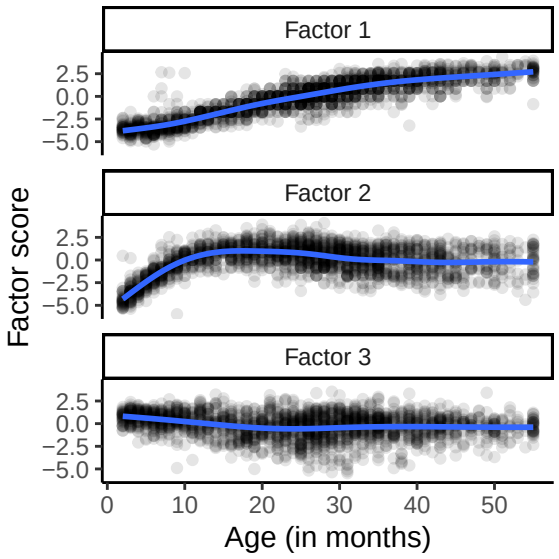

Figure 4: The first factor is highly associated with age

age from 2 to 16 months but thereafter is unrelated to age. By and large, the third factor does not have any association with age.

Dimensionality across the age-span

For the entire dataset, we've shown evidence that the 2PL model with 3 factors performs best. But is this latent factor dimensionality consistent across age? For example, perhaps for very young children 1-factor is sufficient and then later on 2 and then 3 factors become valuable. We take two approaches to assessing the dimensionality of child development across the age-span. First, we examine the performance of each of the models by age. Second, we partition the data by age and use the same cross-validation procedure to find the best fitting model in each partition.

Full model Figure 5 displays the mean cross-validated log likelihood for each model by age, which comes from the k-fold cross-validation described in the model comparison section. For each student, we calculate the marginalized out-of-sample likelihood based on the item parameters $\Psi_j^{(-k)}$ from fitting the model to $y^{(-k)}$, the folds of data that do not include the student. As a reminder, students are assigned to folds randomly and not by age.

Figure 5 shows how both the 3F 2PL and bifactor models compare to the 2F 2PL model in terms of cross-validated log likelihood for each age. The 2F 2PL outperforms both models for children younger than 7 months old. For children older than 11 months old, both the 3F 2PL and bifactor models outperform the 2F 2PL model with the 3F 2PL model tending to perform best.

Age-partitioned models As another method of examining the dimensionality of child development across the age span, we create four partitions of the data based on the ages of the children. We then cross-validate the 2PL models independently in each partition. This analysis allows us to examine the dimensionality for each age group separately. For each age partition, we drop milestones where less than 2.5% or greater than 97.5% of children have reached the milestone because they contain little information and make the models less stable. This process results in, for example, 432 children and 359 milestones in the 13-24 month old partition.

Figure 6 shows the results of this analysis. Consistent with our findings in the previous section, the best fitting model contains a lower dimensional factor structure for younger children. The best fitting model is the 2F 2PL for the partition of data containing children two to 12 months old, whereas the best fitting model is the 3F 2PL for the partitions containing older children.

Discussion

Is child development a single unified process or a host of different processes? Stage theories assume synchronization in developmental changes across distinct domains like language, social/emotional development, and cognition. In contrast, more modern modular theories tend to assume that particular aspects of development proceed "on their own schedule" (Spelke, Breinlinger, Macomber, & Jacobson, 1992). Here, inspired by psychometric studies of age-related changes in cognition, we explored this issue in a large dataset of children's developmental milestones. Our premise was that understanding the nature of variation in milestones could help shed light on whether children's developmental change covaries across domains within a single factor or whether it is split into multiple factors.

Using multi-factor item response theory models and a new cross-validation method for model comparison, we found that a three-factor model best described developmental variation across the first 55 months. While the first factor described a

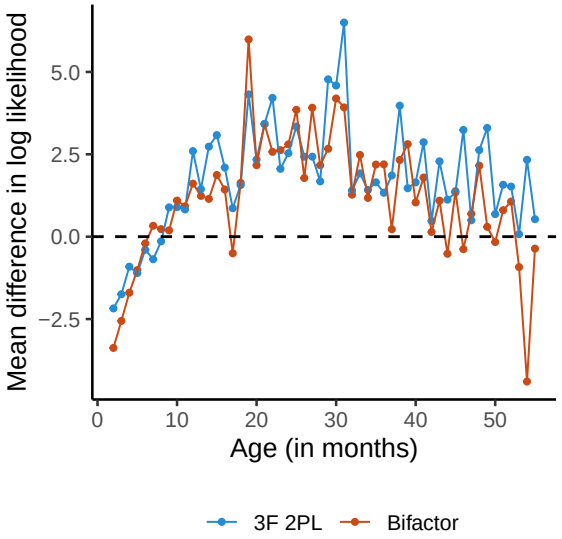

Figure 5: Comparing the 3F 2PL and Bifactor models to the 2F 2PL

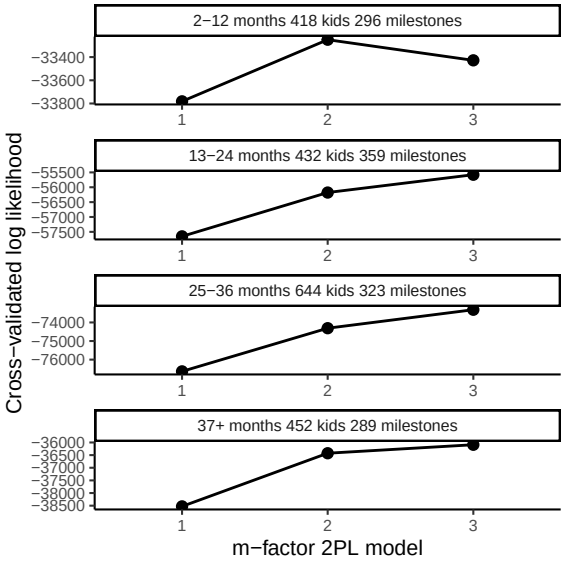

Figure 6: 2F 2PL best for young kids; 3F 2PL best for older kids

large amount of shared variation in development, the structure of these factors did suggest some differentiation between cognitive/linguistic development, physical development, and socio/emotional development. Further, we found that the dimensionality of variation increased developmentally: 2–12 month-olds were best described by a two-dimensional model, while older groups were best described by a three-dimensional model. This analysis provides tentative support for a developmental differentiation hypothesis, where different domains of development vary across individuals in a way that is increasingly more independent over age.

Our study has a number of limitations that should inform future work. Our dataset is cross-sectional, meaning that we are only describing variation across individuals rather than the coherence of factors within individuals. Second, we relied on parent report, which can have significant biases and limitations, especially in its precision regarding capacities that are difficult to observe (e.g., cognitive abilities; Feldman et al., 2000; Frank, Braginsky, Marchman, & Yurovsky, 2020). Third, our data come from a very specific population group (middle- and upper-class Mexican parents whose children were in group care) and hence caution is warranted in generalizing to other populations. Fourth, our cross-validation procedure evaluates item parameters by integrating over a generic ability distribution $g(\boldsymbol{\theta})$, which is consistent with how IRT models are typically estimated (MMLE), but does not map directly onto a practical prediction task. Future work should explore other cross-validation procedures. Fifth, it's important to note that the best-fitting model describes only the optimal dimensionality with regard to child development as defined by the Kinedu milestones; it is at best a distant reflection of the structure of any latent variables internal to the child (Bork, Epskamp, Rhemtulla, Borsboom, & Maas, 2017; Maraun, 2003; Van Der Maas et al., 2006).

This work has significant practical implications: Measures of developmental change should not assume that a single score captures all of the variance in developmental change. Thus, understanding the generality of our conclusions is an important practical goal that could affect the structure of a variety of standardized developmental inventories in broad clinical and research use.

The nature of developmental variation is of core importance to our theorizing about the mechanisms of child development. Yet this variation has often been assumed to be unifactorial or multifactorial without formal evaluation. Our work here takes a first step towards using psychometric models to evaluate this dimensionality empirically.

Acknowledgements

Thank you to Kinedu, Inc. for support and data sharing. Thank you to George Kachergis, Alex Carstensen, Ben Domingue, and Ann Weber for comments on early versions of this paper.

References

Baker, F. B., & Kim, S.-H. (2004). *Item response theory: Parameter estimation techniques*. CRC Press.

Balinsky, B. (1941). An analysis of the mental factors of various age groups from nine to sixty. *Genetic Psychology Monographs*.

Bayley, N. (2009). *Bayley-iii: Bayley scales of infant and toddler development*. Giunti OS.

Bolt, D. M., & Lall, V. F. (2003). Estimation of compensatory and noncompensatory multidimensional item response models using markov chain monte carlo. *Applied Psychological Measurement*, *27*(6), 395–414.

Bork, R. van, Epskamp, S., Rhemtulla, M., Borsboom, D., & Maas, H. L. van der. (2017). What is the p-factor of

psychopathology? Some risks of general factor modeling. *Theory & Psychology*, *27*(6), 759–773.

Borsboom, D. (2005). *Measuring the mind: Conceptual issues in contemporary psychometrics*. Cambridge University Press.

Bricker, D., Squires, J., Mounts, L., Potter, L., Nickel, R., Twombly, E., & Farrell, J. (1999). Ages and stages questionnaire. *Paul H. Brookes: Baltimore*.

Cai, L., Yang, J. S., & Hansen, M. (2011). Generalized full-information item bifactor analysis. *Psychological Methods*, *16*(3), 221.

Chalmers, R. P. (2012). Mirt: A multidimensional item response theory package for the r environment. *Journal of Statistical Software*, *48*(6), 1–29.

Embretson, S. E., & Reise, S. P. (2013). *Item response theory*. Psychology Press.

Feldman, H. M., Dollaghan, C. A., Campbell, T. F., Kurs-Lasky, M., Janosky, J. E., & Paradise, J. L. (2000). Measurement properties of the macarthur communicative development inventories at ages one and two years. *Child Development*, *71*(2), 310–322.

Flavell, J. H. (1963). The developmental psychology of jean piaget.

Frank, M., Braginsky, M., Marchman, V., & Yurovsky, D. (2020). Variability and consistency in early language learning. *Child Development*.

Frias, C. M. de, Lövdén, M., Lindenberger, U., & Nilsson, L.-G. (2007). Revisiting the dedifferentiation hypothesis with longitudinal multi-cohort data. *Intelligence*, *35*(4), 381–392.

Gelman, R., & Meck, E. (1983). Preschoolers' counting: Principles before skill. *Cognition*, *13*(3), 343–359.

Kaiser, H. F. (1959). Computer program for varimax rotation in factor analysis. *Educational and Psychological Measurement*, *19*(3), 413–420.

Li, C., Nuttall, R. L., & Zhao, S. (1999). A test of the piagetian water-level task with chinese students. *The Journal of Genetic Psychology*, *160*(3), 369–380.

Maraun, M. (2003). Myths and confusions: Psychometrics and the latent variable model. *Unpublished Manuscript. Retrieved from Http://Www. Sfu. Ca/˜ Maraun/Myths-and-Confusions. Html*.

Maydeu-Olivares, A. (2013). Goodness-of-fit assessment of item response theory models. *Measurement: Interdisciplinary Research and Perspectives*, *11*(3), 71–101.

McCoy, D. C., Gonzalez, K., & Jones, S. (2019). Preschool self-regulation and preacademic skills as mediators of the long-term impacts of an early intervention. *Child Development*, *90*(5), 1544–1558.

McDonald, R. P., & Mok, M. M.-C. (1995). Goodness of fit in item response models. *Multivariate Behavioral Research*, *30*(1), 23–40.

R Core Team. (2019). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from `https://www.R-project.org/`

Sheldrick, R. C., Schlichting, L. E., Berger, B., Clyne, A., Ni, P., Perrin, E. C., & Vivier, P. M. (2019). Establishing new norms for developmental milestones. *Pediatrics*, *144*(6).

Spelke, E. S., Breinlinger, K., Macomber, J., & Jacobson, K. (1992). Origins of knowledge. *Psychological Review*, *99*(4), 605.

Van Der Maas, H. L., Dolan, C. V., Grasman, R. P., Wicherts, J. M., Huizenga, H. M., & Raijmakers, M. E. (2006). A dynamical model of general intelligence: The positive manifold of intelligence by mutualism. *Psychological Review*, *113*(4), 842.

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical bayesian model evaluation using leave-one-out cross-validation and waic. *Statistics and Computing*, *27*(5), 1413–1432.

Wickham, H. (2017). *Tidyverse: Easily install and load the 'tidyverse'*. Retrieved from `https://CRAN.R-project.org/package=tidyverse`