The latent factor structure of child development

Anonymous Cogsci Submission

Abstract

to do

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 distinct 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 describe 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. 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 our model family specification and evaluation measures. We consider the overall dimensionality of our dataset and then turn to the evidence for change in dimensionality across development. We end by considering the limitations and implications of 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

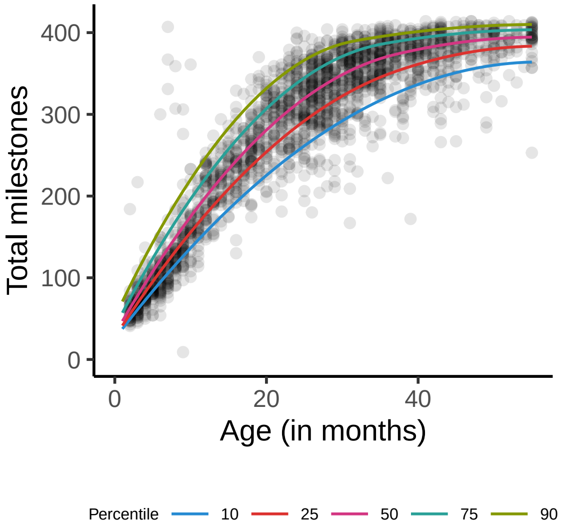

Figure 1: Number of milestones by age

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."

We use Kinedu, a Mexico-based child development app, as a source for this type of data. When parents first start using the Kinedu app, they are asked a series of questions about which developmental milestones their child has reached. We consider the 1946 children between 2 and 55 months of age whose parents responded to all 414 of the developmental milestones. Each developmental miletone on Kinedu is mapped to a milestone group: physical, cognitive, linguistic, or social & emotional. Table 1 shows the number of developmental milestones in each group along with an example milestone translated to English.

Table 1: Developmental milestone groups and examples

Group	Count	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 interrupted

Figure 1 shows the age (in months) and number of developmental 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. Finally, at 48 months old, most children have reached about 375 of the 414 developmental milestones.

probably should describe where the percentile curves come from

Empirical assessment of the dimensionality of child development

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 & others (2012). Let $i = 1,\ldots,I$ represent the distinct children and $j = 1,\ldots,J$ the developmental milestones. The Kinedu 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 (i.e. slopes) $\boldsymbol{a_j} = (a_1,\ldots,a_m)$ capture the latent factor loadings onto that milestone.

We fit four two-parametric logistic (2PL) models where a child's development is represented by $m = 1$, $m = 2$, $m = 3$, and $m = 4$ latent factors. Hereafter, we, for example, refer to a 2PL model with $m = 4$ latent factors as a 4F 2PL model. 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. We also fit a 1F Rasch model where each of the discriminations is fixed to 1. As an example, Figure 2 shows item characteristic curves from 1F 2PL model for the items in Table 1. Most children babbling and babbling is unrelated with development. 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.

The 2PL models learn the latent factor structure entirely from the data, making them exploratory. The bifactor model offers an alternative specification where each milestone loads onto a general factor θ_0 and a specific factor θ_s (Cai, Yang, & Hansen, 2011). The assignment of each developmental milestone to its specific factor is an opportunity to specify the latent factor structure, making the model confirmatory as opposed to exploratory. We map each milestone to its specific factor according to the four developmental milestone groups shown in Table 1. For the bifactor model, the probability of a child having achieved a developmental milestone is

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

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 misspecification (McDonald & Mok, 1995). Instead, we prefer 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 $g(\theta)$.

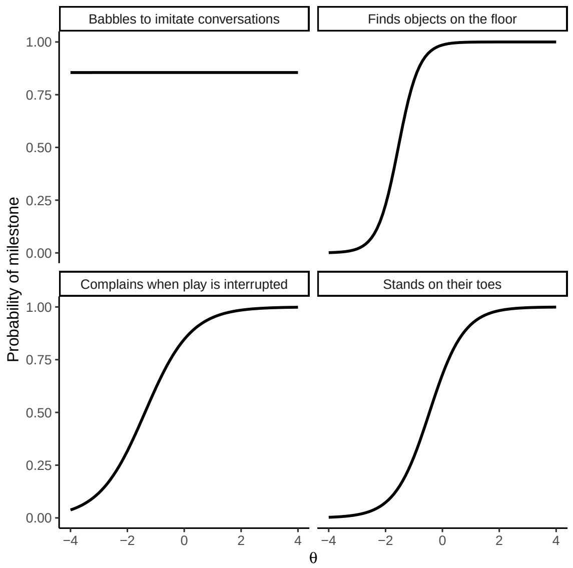

Figure 2: Example item characteristic curves

Mathematically and following notation similar to Vehtari, Gelman, & Gabry (2017), we partition the data into K subsets $y^{(k)}$ for $k = 1, \ldots, K$. 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

fascinating that dropping 1 month olds changed the winner from 3F to 4F

Table 2 shows the number of parameters, the in-sample log likelihood (which neccessarily increases with more parameters), and the lpl y defined in the model comparison section. The 4F 2PL model performs best which is evidence that child development between the ages of 2 and 55 months follows a multidimensional path.

need to tie results to literature

Computing is done in R (R Core Team, 2019), model fitting in the R package mirt (Chalmers & others, 2012), and data wrangling/visualization in the set of R packages known as the tidyverse (Wickham, 2017).

think about presentation of this table. is it worth displaying in sample numbers? would it be better as a graph?

Table 2: Model performance: The 4F 2PL performs best

model	npars	in_log_lik	lpl y (out-of-sample)
1F Rasch	415	-254983.9	-255443.5
1F 2PL	828	-222073.0	-223072.0
2F 2PL	1241	-212957.9	-214498.2
Bifactor	1242	-210030.2	-211682.2
3F 2PL	1653	-208887.6	-210943.4
4F 2PL	2064	-208124.1	-210793.5

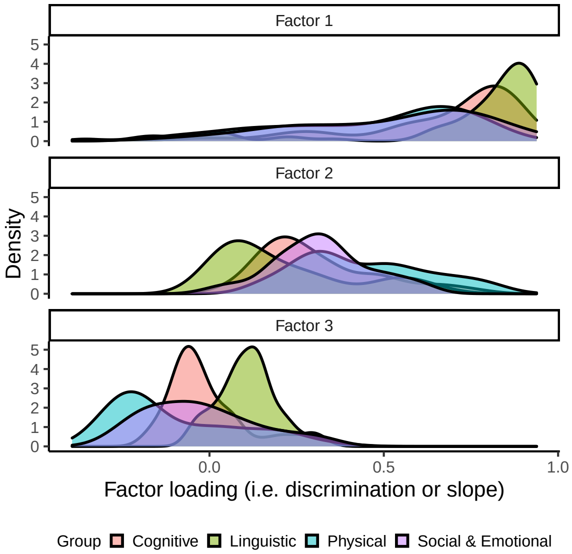

Figure 3: Factor loadings by group

Understanding the latent factor structure

section is written as if 3F is winner may need to edit to 4F

To understand each of the three factors in the best performing model, we fit the model to the the full dataset. We then estimate 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). Figure 3 shows the distribution of factor loadings for each group on each of the three factors. The first factor load 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 mainly load positively on linguistic milestones and, interestingly, negatively on physical milestones.

We also estimate the factor scores for each child using expected a posteriori (EAP) with a three dimensional standard normal distribution (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 = 0.90) with age. The second factor has a strong association with 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.

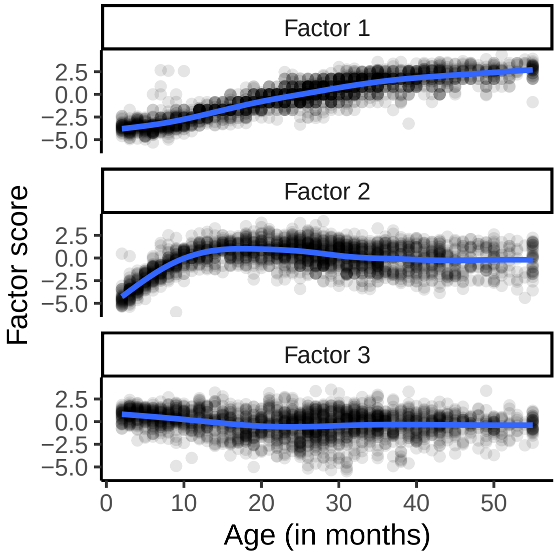

Figure 4: The first factor is highly associated with age

Dimensionality across the age-span

For all of the data, we've shown evidence that 4-factors performs best. But is this latent factor dimensionality consistent across age-span? 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 currently shows up at end of paper. if we go with this graph, probably want more smoothing

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 does not include the student. As a reminder, students are assigned to folds randomly and not by age.

Figure 5 shows how both the 4F 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 4F 2PL and bifactor models outperform the 2F 2PL model with the 4F 2PL model tending to perform best. Interestingly, the bifactor model performs more similarly across the age-span to the 4F 2PL than the 2F 2PL despite having a similar number of parameters to the 2F 2PL.

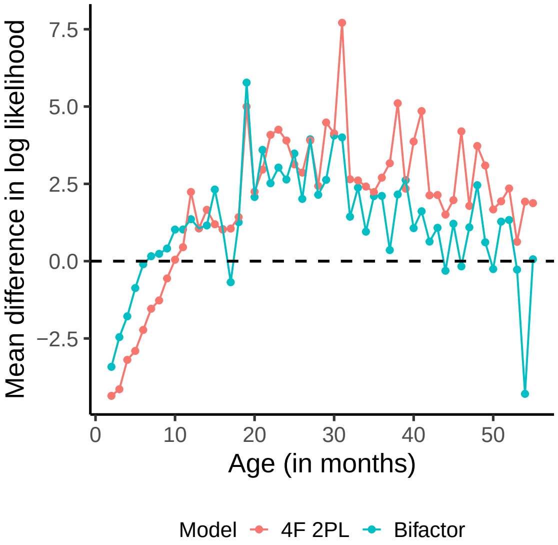

Figure 5: Comparing the 4F 2PL and Bifactor models to the 2F 2PL

Age-partitioned models

figure currently shows up at end of paper. need to add in 4F 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 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. Dropping milestones is done because these milestones do not contain much information and make IRT models less stable. This process results in, for example, 432 children and 359 milestones in the 13-24 month old partition.

should we note (or sensitivity check) the tendency smaller datasets to prefer less flexible models?

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

to do

Acknowledgements

We'd like to thank Kinedu for providing the data that made this research possible.

References

Balinsky, B. (1941). An analysis of the mental factors of

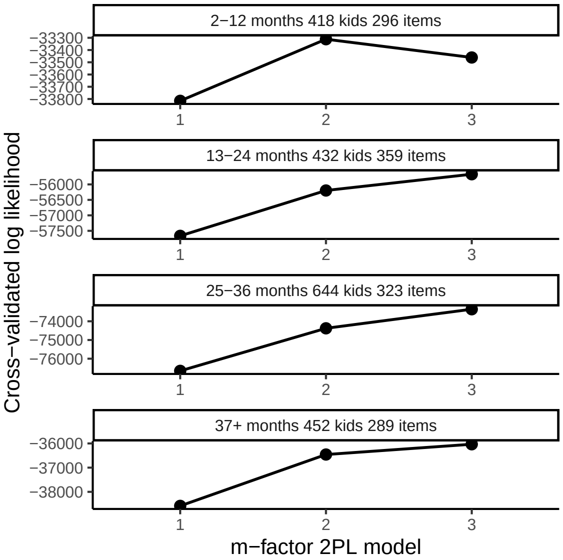

Figure 6: 2F 2PL best for young kids; 3F 2PL best for older kids

various age groups from nine to sixty. *Genetic Psychology Monographs*.

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

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., & others. (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.

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

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.

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).

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`