

Bayesian Hierarchical Multidimensional Item Response Modeling of Small Sample, Sparse Data for Personalized Developmental Surveillance

Educational and Psychological
Measurement

2021, Vol. 81(5) 936–956

© The Author(s) 2021

Article reuse guidelines:

sagepub.com/journals-permissions

DOI: 10.1177/0013164420987582

journals.sagepub.com/home/epm



Patricia Gilholm^{1,2} , Kerrie Mengersen^{1,2}
and Helen Thompson¹

Abstract

Developmental surveillance tools are used to closely monitor the early development of infants and young children. This study provides a novel implementation of a multidimensional item response model, using Bayesian hierarchical priors, to construct developmental profiles for a small sample of children ($N = 115$) with sparse data collected through an online developmental surveillance tool. The surveillance tool records 348 developmental milestones measured from birth to three years of age, within six functional domains: auditory, hands, movement, speech, tactile, and vision. The profiles were constructed in three steps: (1) the multidimensional item response model, embedded in the Bayesian hierarchical framework, was implemented in order to measure both the latent abilities of the children and attributes of the milestones, while retaining the correlation structure among the latent developmental domains; (2) subsequent hierarchical clustering of the multidimensional ability estimates enabled identification of subgroups of children; and (3) information from the posterior distributions of the item response model parameters and the results of the clustering were used to construct a personalized profile of development for each child.

¹Queensland University of Technology, Brisbane, Queensland, Australia

²Australian Research Council Centre of Excellence for Mathematical and Statistical Frontiers, Brisbane, Queensland, Australia

Corresponding Author:

Patricia Gilholm, School of Mathematical Sciences, Science and Engineering Faculty, Queensland University of Technology, 2 George Street, Brisbane, Queensland 4000, Australia.

Email: p.gilholm@qut.edu.au

These individual profiles support early identification of, and personalized early interventions for, children with developmental delay.

Keywords

multidimensional item response model, developmental surveillance, hierarchical clustering, Bayesian hierarchical modeling

The overarching goal of early intervention in childhood development is to improve and strengthen development for a heterogeneous group of children who have a range of delays, including communication, motor, sensory, and perceptual concerns (Guralnick & Bruder, 2019). The developmental domains targeted for early intervention are unique to each child. Therefore, identifying the specific domains that require more focused attention is important for early interventions to be successful. This article proposes and demonstrates novel methodology for constructing comprehensive personalized developmental profiles that can be used to support routine developmental surveillance and personalized interventions.

The personalized profiles are constructed in three steps: (1) a multidimensional item response model is implemented using a Bayesian hierarchical framework, and applied to data collected through an online developmental surveillance tool; (2) the individual inferences obtained from the item response model are used to identify subgroups of children with similar developmental profiles, through the use of hierarchical clustering; (3) the results from the item response model and the clustering are combined to construct the profiles. The data extracted from the online developmental surveillance tool consist of a small sample of children with sparse measurements. Sparse data and small samples are common in routine data collection (Powell et al., 2003), and this article demonstrates that accurate personalized profiles can be constructed, even for children with very sparse measurements, by embedding the item response model within a Bayesian hierarchical framework. In addition, the uncertainty of the estimates can also be estimated and communicated using this framework.

In this article, a multidimensional item response model is used to jointly estimate children's latent abilities in multiple developmental domains and attributes of the developmental milestones that are used to measure the latent constructs. In item response modeling, the probability of a correct response to an item is modeled as a function of both person characteristics (the latent ability of the person) and item characteristics (the difficulty and discrimination of the items; Fox, 2010). This study focuses on multidimensional item response models (also multidimensional item response theory or MIRT), which are used to model more than one latent trait or ability (Sheng & Wikle, 2007). These latent traits are often not independent; however, MIRT models account for the dependency between traits by directly estimating the correlations between latent traits through the incorporation of all subtests within one model, rather than analyzing each subtest separately (Cheng et al., 2009). MIRT

can either be specified so that each item measures more than one latent trait, also known as between-item multidimensionality (Adams et al., 1997; Zhang, 2012) or as a multi-unidimensional IRT model, or within-item MIRT model, where each item only contributes to one latent ability (Adams et al., 1997; Sheng & Wikle, 2007). This study implements the second type of MIRT.

Implementing IRT models often requires large heterogeneous samples that reflect the range of population characteristics under investigation. When the purpose of the IRT model is for item calibration, then sample sizes of 500 or more are recommended (Reeve & Fayers, 2005). Moreover, most IRT models assume complete responses or only small amounts of nonresponse, as many missing responses can lead to biased estimates of both item and person parameters (Rose et al., 2015). Furthermore, IRT models are often estimated using maximum likelihood methods, which rely on asymptotic approximations that require large sample sizes to obtain accurate parameter estimates. Response data for item response models are typically nonnormally distributed and only have a small amount of sample data to estimate each parameter, particularly at the within-individual level (Fox, 2010). These features of item response models mean that using a maximum likelihood estimation-based approach could lead to inaccurate parameter estimates when used on small samples and sparse data.

Item response models that have been used in developmental research have focused on standardization and assessment of items from preexisting developmental screening tools, such as the Denver Developmental Screening Test (de Lourdes Drachler et al., 2007), the Ages and Stages Questionnaire (Chen et al., 2018), and the Paediatric Evaluation of Disability Inventory (Haley et al., 2010), or in the construction of new developmental screening tools (Lancaster et al., 2018). All of these previous studies used large samples of typically developing children to have sufficient numbers of children across the age range required for standardization.

The sample of children assessed in this research predominantly have a developmental disability, so modeling strategies to account for this small, heterogeneous sample need to be implemented in order to obtain accurate personalized assessments. In addition, the data collected through the online tool is sparse, due to a high level of missing data. This sparsity arises because the children are at different stages through the online program, so some children are not old enough to have responded to the later milestones. In addition, the level of engagement differs between users, resulting in intermittent nonresponse. Sparse data are common in online environments such as recommender systems, where each individual only rates a small number of items from a possibly very large item pool (Demiriz, 2004). Bayesian hierarchical priors are used to overcome the sparse data and small sample problems.

Bayesian hierarchical priors improve parameter estimation for small samples *and* sparse measurements by pooling information from parameters of the same type (König et al., 2020). In the Bayesian hierarchical prior, the parameters at the bottom level of the hierarchy (e.g., the ability parameters for the children or the item parameters of the milestones) are related to each other and are viewed as a sample from a

common population distribution (Gelman et al., 2013). The parameters of this population distribution are also given a prior distribution. By using a hierarchical prior, each individual-level parameter “borrows strength” from the corresponding parameters of other individuals with similar characteristics, and this is accomplished through shrinkage toward the population mean (Fox, 2010; Ntzoufras, 2011). It is this shrinkage property which makes the Bayesian hierarchical prior so useful for small samples and sparse data.

The use of Bayesian MIRT models has been demonstrated on both simulated and real data examples (de la Torre & Patz, 2005; Sheng & Wikle, 2008), but they used large and complete samples. Sheng (2012) and König et al. (2020) implemented Bayesian hierarchical *univariate* item response models for small samples, but both were simulation studies and they did not have sparse measurements. Only one example of a Bayesian hierarchical *multidimensional* item response model applied to small samples could be found in the literature (de la Torre & Hong, 2010); however, the smallest sample size used in the article was 500, which was part of a simulation study. This article presents a novel implementation of a Bayesian hierarchical MIRT model for a real world, small, very sparse sample. The MIRT model is the first step of the methodology developed in this article to construct the personalized developmental profiles. The MIRT models six latent traits, representing different functional domains of development. The functional domains are highly correlated, and the correlation structure is modeled explicitly within the MIRT model.

The second step of our proposed methodology is to use the information from the posterior distributions of the ability parameters to identify subgroups of children that have similar development across the functional domains. Hierarchical clustering is used to obtain these subgroups, as it is relatively easy to implement and provides graphical summaries that are interpretable for a wide audience.

The third step of the methodology combines information derived from the individual posterior distributions with the information collated from clustering the posterior estimates to construct a comprehensive developmental profile for each individual child. As the intention of the developmental profiles is to assist with developmental surveillance, information from the posterior distributions of the children’s abilities, combined with the clustering information, can be used to assist clinicians by (1) identifying at-risk children with lower ability and (2) identifying specific developmental domains to be targeted for personalized early interventions.

Method

Data and Sample Description

The data used in this study came from a parent-reported online surveillance tool, created by The Developing Foundation, a charity that offers support services for families of children or adults who have a brain injury or developmental disability (The Developing Foundation, 2017). The Foundation created the Developing Childhood surveillance tool (The Developing Foundation, 2017) to assist parents and carers to

Table 1. Example 1-, 12-, 18-, and 34-Month Milestones in Each Functional Domain.

Functional domain	1 month	12 months	18 months	34 months
Vision	Instantly blinks at bright light	Television or colorful moving objects capture attention	Visually aware of close and distant world	Recognizes and points out tiny details in pictures
Auditory	Instantly startles to sudden loud noise	Listens to speech without distraction from other sounds	Follows simple two-step commands	Comprehends three key words in a sentence
Tactile	Negative response to pain, positive to comfort	Maintains balance with supported stepping	Begins to identify objects by touch alone	Aware of body size in relation to surroundings
Speech	Nonspecific cry	Sound-making with intent	Social speech used for interacting	Regular use of speech to tell stories and experiences
Movement	Unrestricted range of movement in all limbs	Walks holding on to one hand	Attempts to run but without a lot of control	Can pedal a tricycle with good control
Hands	Hands mostly fist ed or slightly open	Finger feeding with pincer grasp	Stacks 4 to 6 blocks	Can dress and undress completely

monitor and assess their own child’s achievement of early developmental milestones. The online program assesses 348 developmental milestones from birth to 3 years of age. These developmental milestones are categorized into six functional domains—namely, auditory, speech, tactile, movement, vision, and hand function. There are 58 milestones assessed in total in each domain. These milestone measurements are not spread uniformly across the 3-year period. Rather, within each functional domain, there are three milestones assessed each month for the first 12 months, two milestones assessed per month between 13 and 18 months, one milestone assessed per month from 19 to 25 months and the remaining three milestones measured at 28, 31, and 34 months, respectively. Example milestones are found in Table 1, which is reproduced from Gilholm et al. (2020).

The data used in this study were extracted from the Developing Childhood program between February 2015 and February 2017. In total, 115 children used the program during this period. However, not all parents reported on all relevant milestones for their child. The information obtained for each child differs dramatically between children, depending on the age of the child and the level of engagement with the program. Overall, there is approximately 42% missing data across the entire sample, with later milestones having far more missing data compared with earlier ones. This phenomenon is discussed in detail later. Plots presenting the missing data patterns can be found in the Supplemental Material (available online).

Information regarding the disability status (i.e., whether the child has a disability and the type of disability) was not recorded for all participants. The children in the sample that did record this information have a diverse range of disabilities, including autism spectrum disorder, cerebral palsy, Down syndrome, and speech and hearing impairments. Although not all disability statuses were known, it is assumed that this sample had more children with a developmental delay or disability than the general population as they were using the services provided by the Developing Foundation. The results reported in this article and the supplemental material, as well as the information provided on Github, are provided at an aggregate level and preserve the privacy of the individual subjects in the study. For this reason, and since the data are nonidentifiable, the Queensland University of Technology's Human Research Ethics committee waived the need for consent from the parents or guardians for the data used in this research.

Model

In this study, a two-parameter logistic MIRT model was used. The difficulty parameter refers to a milestone's location on the continuous latent functional domain and the discrimination parameter refers to the slope of the milestone's item characteristic curve (i.e., milestones with a steeper slope are better at discriminating between children who have high ability and those who have low ability). In addition to the two item parameters, the model also measures the ability of the children, which refers to the child's location on the continuous latent functional domain (de Ayala, 2013).

The two-parameter MIRT model measures the probability of milestone achievement as follows (de la Torre & Patz, 2005):

$$Y_{ij(d)} \sim \text{Bernoulli}(p_{ij(d)}),$$

$$\log\left(\frac{p_{ij(d)}}{1 - p_{ij(d)}}\right) = \alpha_{j(d)}(\theta_{i(d)} - \beta_{j(d)}), \quad (1)$$

where

d denotes the functional domain, $d = 1, \dots, D$;

$Y_{ij(d)}$ is child i 's response to milestone j for functional domain d ;

$p_{ij(d)}$ is the probability that child i achieves milestone j for functional domain d ;

$\theta_{i(d)}$ is the ability parameter for child i for functional domain d , where

$\boldsymbol{\theta}_i = (\theta_{i(1)}, \dots, \theta_{i(D)})$ is the vector of all abilities for child i ;

$\alpha_{j(d)}$ is the discrimination parameter for milestone j within functional domain d ;

and

$\beta_{j(d)}$ is the difficulty parameter for milestone j within functional domain d .

The MIRT model was embedded in a Bayesian hierarchical framework (Fox, 2010), where prior distributions for the model parameters are imposed at two levels

of the model hierarchy. At the first level, the individual parameters of the same type are aggregated and a prior distribution for the population of parameters is specified. At the second level, the parameters for this prior distribution, the hyperparameters, are also given prior distributions, or hyperpriors.

The priors adopted for this study are as follows. The vector of ability parameters for each child, θ_i , are assumed to be independently distributed from a multivariate normal distribution (Bürkner, 2019; Fox, 2010):

$$\theta_i \sim \text{MVN}(\mu_\theta, \Sigma_\theta). \quad (2)$$

The hyperprior distribution for the multivariate normal hyperparameters μ_θ and Σ_θ is a normal-inverse-Wishart, which is specified as follows (Gelman et al., 2013):

$$\begin{aligned} \mu_\theta | \Sigma_\theta &\sim \text{MVN}\left(\mu_0, \frac{\Sigma_\theta}{N_0}\right), \\ \Sigma_\theta &\sim \text{InvWishart}(\nu_0, \Lambda_0). \end{aligned} \quad (3)$$

where μ_0 is the vector of prior means, N_0 is the number of prior measurements and ν_0 and Λ_0 are the degrees of freedom and scale matrix for the inverse-Wishart distribution (Gelman et al., 2013). In this implementation, the following values were chosen for the hyperpriors, $\nu_0 = p + 1 = 7$, $\Lambda_0 = I_p$, $N_0 = 1$, where p is the number of latent dimensions. A noncentered reparameterization was performed for the multivariate normal hyperparameters, which is outlined in Equation 4.

For the item discrimination parameters, $\alpha_{j(d)}$, an independent truncated normal prior, $\alpha_{j(d)} \sim N_+(\mu_\alpha, \sigma_\alpha^2)$, is used, which ensures that the discrimination parameters are positive (Curtis, 2010). A normal distribution, $N(1, 1)$, is used as a hyperprior distribution for the hyperparameter, μ_α and an inverse-gamma hyperprior, $\text{InvGamma}(1, 1)$, is used for σ_α^2 (Fox, 2010). For the item difficulty parameters, $\beta_{j(d)}$, a standard normal prior, $\beta_{j(d)} \sim N(0, 1)$, is used. To ensure model identifiability, no hyperpriors are specified for the difficulty hyperparameters (Fox, 2010). The complete hierarchical structure of the model is represented in the directed acyclic graph in Figure 1.

The posterior distributions for all parameters were approximated through Markov chain Monte Carlo (MCMC) simulation. Specifically, the Hamiltonian Monte Carlo No-U-turn sampler (HMC NUTS) was implemented (Carpenter et al., 2017). For a thorough introduction to this method, please refer to Monnahan et al. (2017). The HMC NUTS sampler uses gradients to explore the target distribution more efficiently (Betancourt & Girolami, 2015). Because of this efficient exploration of the target space, the HMC NUTS sampler is able to overcome some of the computational problems associated with hierarchical models, such as sampling from funnel distributions resulting from high correlations between local and global parameters (Betancourt & Girolami, 2015).

In addition to using the HMC NUTS sampler, a noncentered parameterization was used for modeling the multidimensional ability distributions, which removes the

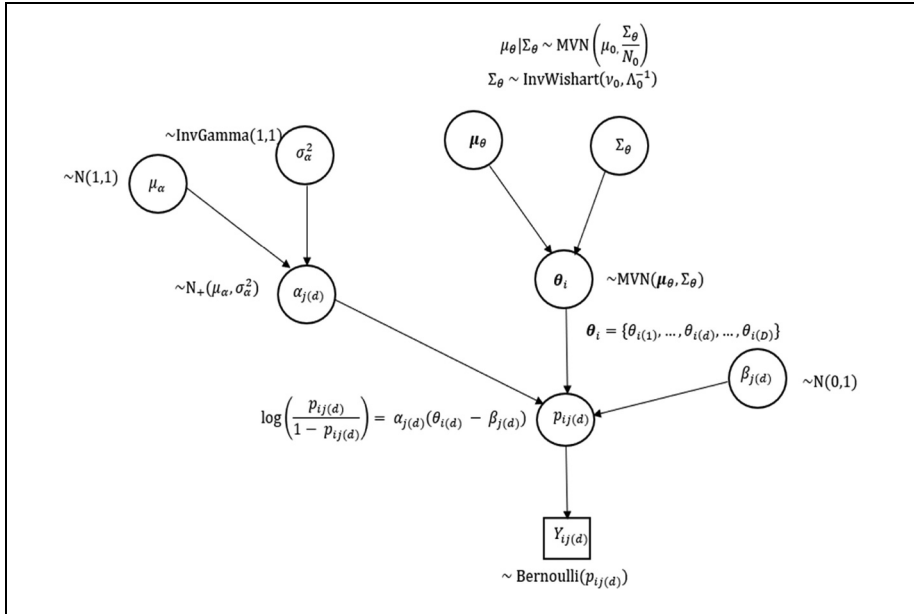


Figure 1. A directed acyclic graph of the hierarchical structure of the model.

dependency between the ability parameters θ_i and the global mean parameter μ_θ (König et al., 2020). This can also help alleviate the computational problems associated with correlations between the parameters at different levels.

The noncentered parameterization for θ was implemented as follows:

$$\begin{aligned} \theta &= \mu_\theta + L\delta, \\ LL^T &= \Sigma_\theta, \end{aligned} \quad (4)$$

where the elements of δ are independently and identically distributed standard normal and L is the Cholesky decomposition of the covariance matrix Σ_θ .

The full model with the noncentered reparameterization was written in the Stan programming language (Carpenter et al., 2017). The sampler ran four chains for 50,000 iterations each. Convergence of the chains for all model parameters was checked by assessing the trace plots, and by inspecting the \hat{R} statistic and the effective sample size for each parameter (Bürkner, 2019).

Hierarchical Clustering

Hierarchical clustering of the posterior distributions of the ability parameters for each child was performed to identify children with similar developmental profiles. First, a standardized posterior mean, $z_{\theta_{i(d)}} = \frac{\mu_{\theta_{i(d)}}}{\sigma_{\theta_{i(d)}}}$, was computed for each functional domain

for every child, where $\mu_{\theta_{i(d)}}$ and $\sigma_{\theta_{i(d)}}$ are, respectively, the posterior mean and standard deviation of the ability parameter $\theta_{i(d)}$, $i = 1, \dots, N$, $d = 1, \dots, D$. This standardization allows fairer comparisons between children, regardless of how many milestones they had completed.

Second, the $z_{\theta_{i(d)}}$ were used as input for hierarchical agglomerative clustering (James et al., 2013). In hierarchical agglomerative clustering, each data point starts in its own cluster. Then data points are iteratively merged based on a measure of similarity between points, until all data points are merged into one cluster. The hierarchical structure of the merges can be displayed graphically in the form of a dendrogram (James et al., 2013). The measure of similarity implemented in this article was the Euclidean distance (James et al., 2013), where the distance between pairs of observations is defined as

$$D(x_i, x_{i'}) = \sum_{j=1}^p d_j(x_{ij} - x_{i'j})^2, \quad (5)$$

where $d_j(x_{ij} - x_{i'j})^2$ is the Euclidean distance between child i and child i' for the j th latent dimension.

To implement the hierarchical clustering, the Euclidean distance (James et al., 2013) and Ward's linkage (Murtagh & Legendre, 2014) were used; both of these are common methods that are readily available in popular statistical programs. To identify the optimal number of clusters obtained from the hierarchical clustering, the NbClust package was used (Malika et al., 2014), which uses 30 different indices to assess the number of clusters, and the majority rule across the 30 indices was taken as the optimal number.

For this study, the clustering was performed in two stages. In the first stage, multivariate clustering was performed through hierarchical clustering of all six functional domain standardized posterior means. In the second stage, univariate clustering was performed by clustering each functional domain independently; this allowed subgroups to be identified both within and between functional domains.

All models were implemented in R (R Core Team, 2019), and the Bayesian hierarchical MIRT was implemented in Stan using the Rstan package (Stan Development Team, 2020). The code used to implement the models is available on the first author's Github (Gillholm, 2020).

Results

Bayesian Hierarchical Multidimensional Item Response Model

Posterior estimates of the population parameters of the MIRT model and the associated convergence statistics, effective sample size and \hat{R} , are provided in Table 2. The parameter estimates and convergence statistics for the individual ability and item parameters are available in the supplemental material. The posterior correlations between the functional domains, obtained as a by-product of the MCMC sampler,

Table 2. Posterior Distribution and Convergence Statistics for the Population Parameters.

Parameter	Mean	SD	2.5%	97.5%	ESS	\hat{R}
Mean difficulty	0.00	—	—	—	—	—
Variance difficulty	1.00	—	—	—	—	—
Mean discrimination	2.29	0.12	2.07	2.54	10416.56	1
Variance discrimination	0.61	0.08	0.46	0.77	10225.92	1
Mean auditory ability	1.09	0.19	0.73	1.46	17973.76	1
Mean hands ability	1.58	0.21	1.19	1.99	25027.99	1
Mean movement ability	1.52	0.23	1.08	1.98	30963.69	1
Mean speech ability	0.92	0.20	0.54	1.31	21673.03	1
Mean tactile ability	1.27	0.17	0.93	1.61	20881.46	1
Mean vision ability	1.31	0.18	0.96	1.67	21041.68	1
Variance auditory ability	1.35	0.30	0.86	2.04	25287.92	1
Variance hands ability	1.53	0.35	0.97	2.33	32639.60	1
Variance movement ability	2.50	0.60	1.54	3.89	35590.82	1
Variance speech ability	1.83	0.39	1.20	2.70	28341.05	1
Variance tactile ability	0.62	0.14	0.39	0.94	33220.98	1
Variance vision ability	0.92	0.22	0.57	1.42	30075.11	1
<i>Correlations between ability dimensions</i>						
Auditory and hands	0.71	0.07	0.56	0.82	42589.57	1
Auditory and movement	0.41	0.10	0.19	0.59	36811.07	1
Auditory and speech	0.64	0.07	0.49	0.77	38016.51	1
Auditory and tactile	0.72	0.06	0.58	0.83	46304.51	1
Auditory and vision	0.86	0.04	0.76	0.92	47944.27	1
Hands and movement	0.64	0.08	0.47	0.78	34462.74	1
Hands and speech	0.61	0.08	0.43	0.75	27286.67	1
Hands and tactile	0.80	0.06	0.67	0.89	37015.26	1
Hands and vision	0.75	0.07	0.60	0.86	48103.71	1
Movement and speech	0.51	0.09	0.32	0.67	34199.17	1
Movement and tactile	0.73	0.07	0.57	0.84	53111.04	1
Movement and vision	0.44	0.10	0.23	0.63	51453.35	1
Speech and tactile	0.50	0.09	0.31	0.67	46535.76	1
Speech and vision	0.62	0.08	0.45	0.75	50962.31	1
Tactile and vision	0.76	0.06	0.62	0.86	54596.81	1

Note. ESS = effective sample size.

are also provided in Table 2. The magnitude of these correlations provide support for the choice of a multidimensional model. The high correlations among the functional domains is to be expected from a developmental context, as children do not develop in each functional domain independently but rather as a whole system of interconnected parts, where delays in one domain can effect development in other domains (Ayoub & Fischer, 2006).

This model produced posterior distributions for the children's abilities for each functional domain and the difficulty parameters for every milestone. The ability estimates are adjusted for the difficulty of the milestone, where the difficulty is estimated

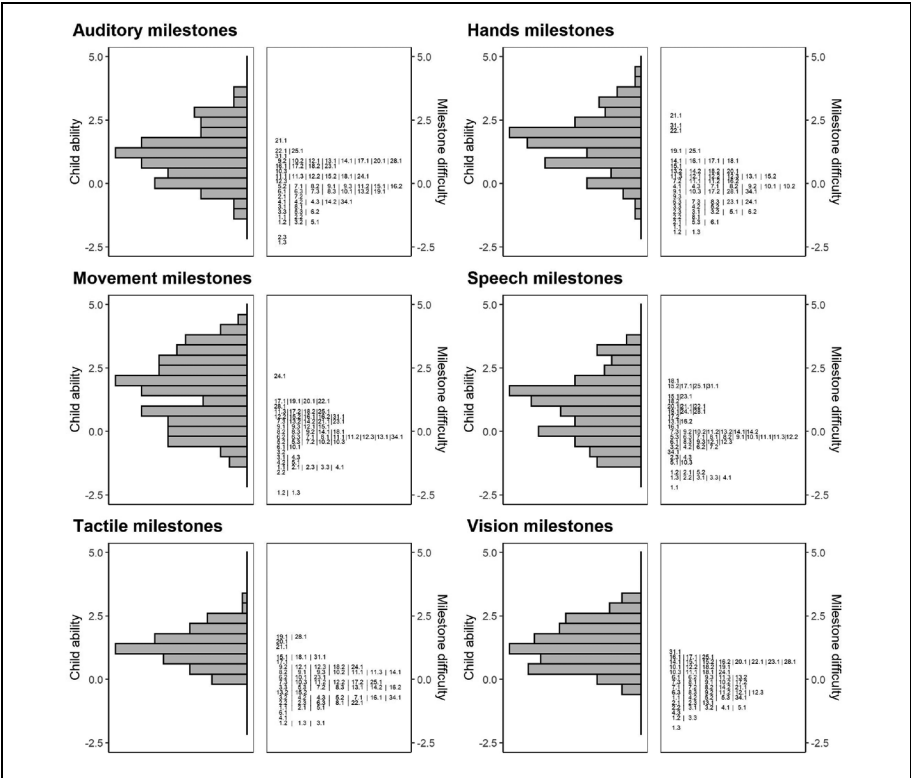


Figure 2. Wright maps of the posterior means for the milestone difficulty and child ability parameters for the six functional domains. The number for each milestone references the month and the sequence within the month that the milestone is expected to be achieved—for example, 1.1 is the first expected milestone to be achieved when a child is 1 month old, 1.2 is the second milestone expected to be achieved when a child is 1 month old, and so on.

from the data. The posterior means of the children’s abilities and the milestone difficulties for each dimension are displayed in the Wright maps in Figure 2.

Overall, the distribution of the posterior means for the children’s abilities is higher on the latent dimensions compared with the item difficulty distributions, showing that, on average, the children were able to achieve most of the milestones. There is also variation in the distributions of both the child abilities and the milestone difficulties between the functional domains. For example, there is more variation in the abilities and difficulties within the movement and speech domains compared with the tactile and vision domains.

The posterior distributions for the item difficulty and discrimination parameters for each milestone are summarized in Figure 3. Here it can be seen that earlier milestones are less difficult to achieve than later ones across all functional domains, as

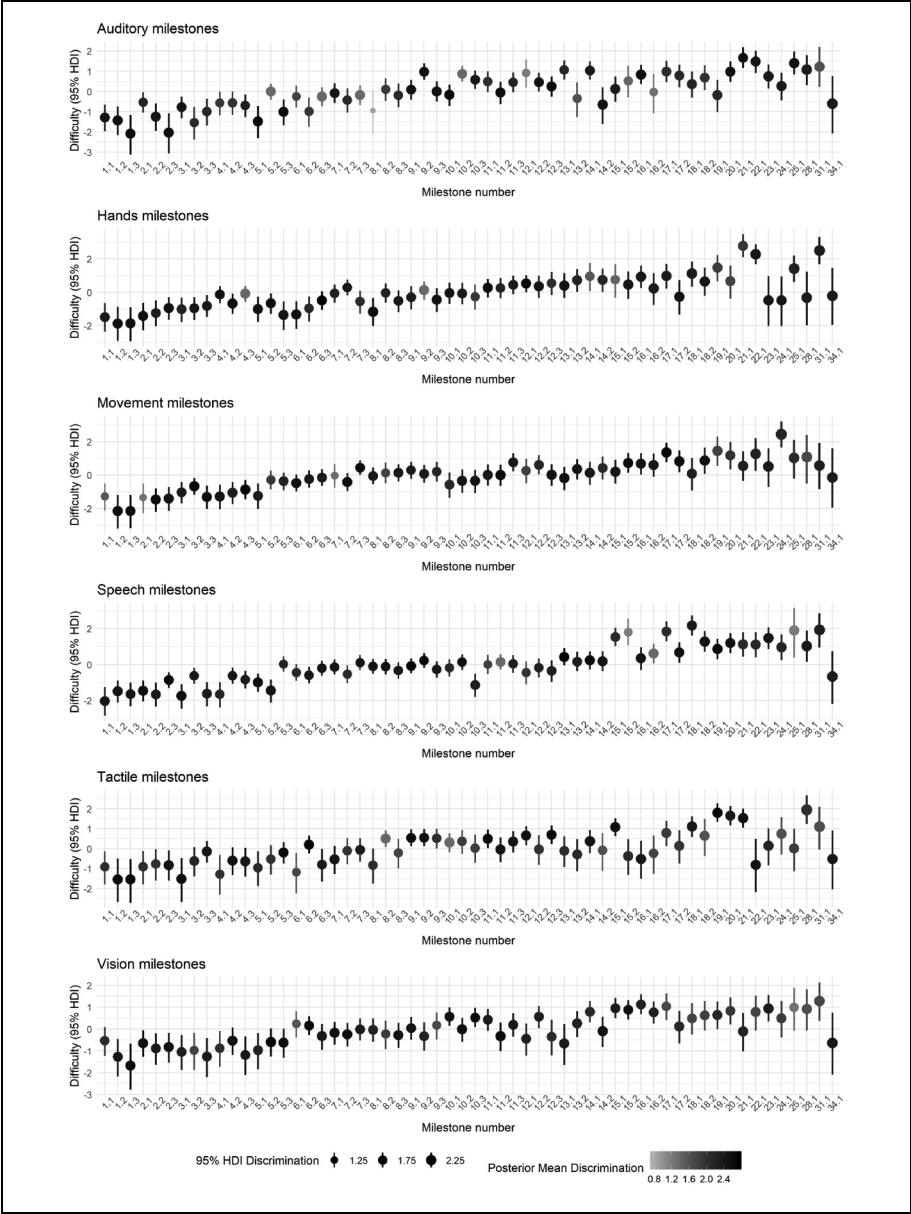


Figure 3. Bubble plots of the difficulty and discrimination parameters for each milestone within each functional domain. The location of the bubble and the bars refer to the posterior mean and the 95% HDI (highest density interval) of the difficulty parameter for that milestone. The color and size of the bubble references the posterior mean and 95% HDI of the discrimination parameter, respectively.

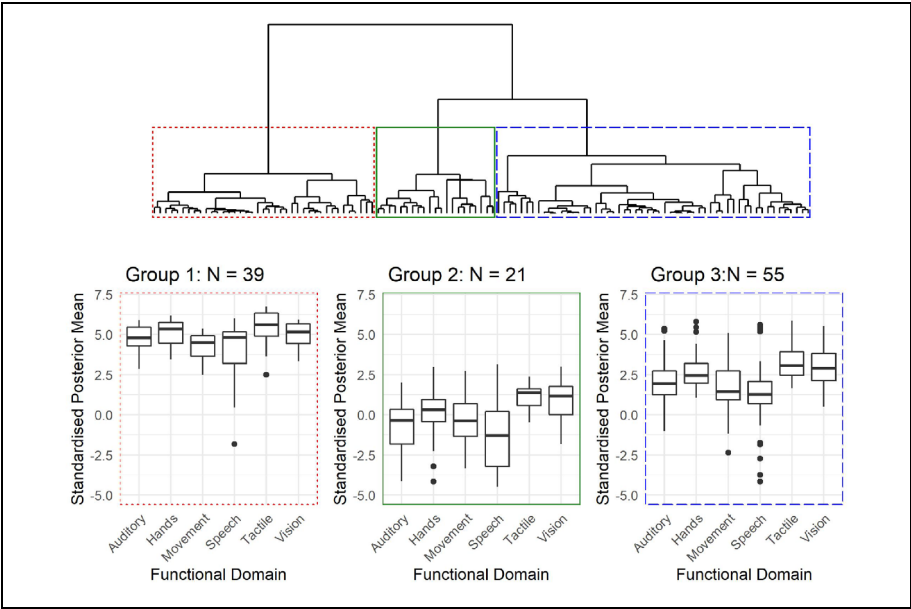


Figure 4. Dendrogram and profile plots of the three clusters obtained from the multivariate hierarchical clustering.

the difficulty steadily increases over time. However, Figure 3 also shows that there is much more variation in the posterior distributions for the difficulty parameters for the later milestones. This may be because only a small proportion of children have responded to these later milestones, which results in more uncertainty in the posterior distributions. The posterior distributions for the item parameters were also ranked to identify the most and least difficult or discriminatory milestones, respectively. The plots displaying the results of the ranking for each domain can be found in the supplemental material.

Hierarchical Clustering

The standardized posterior means of the ability parameters were clustered using hierarchical agglomerative clustering to identify subgroups of children showing similar developmental patterns. The clustering was executed using the standardized posterior means to adjust for the number of milestones completed for each child. Plots demonstrating the effect of the adjustment for each functional domain can be found in the supplemental material.

The clustering was performed in two stages, multivariate clustering and univariate clustering. The clusters identified from the multivariate clustering are displayed in Figure 4. Three clusters were identified, as determined by the cluster validation

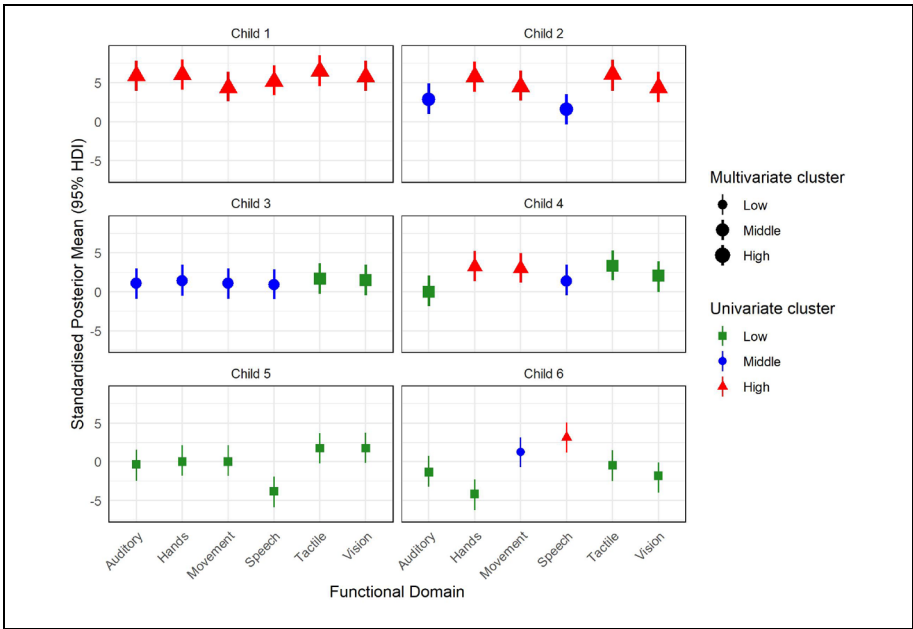


Figure 5. Standardized posterior mean and 95% HDI (highest density interval) for each functional domain. The bubble color references the univariate cluster membership and the size of the bubbles references the multivariate cluster membership.

indices, and can be summarized as a high-ability cluster (left), a low-ability cluster (middle), and a moderate-ability cluster (right). Although these clusters are useful for identifying the average ability of each child across the six functional domains, they are not very informative in terms of specific ability within each functional domain. To identify subgroups within each of the functional domains, univariate hierarchical clustering was performed by clustering the standardized posterior means for each functional domain separately. The cluster validation indices revealed three clusters of high, moderate, and low ability for the auditory, hands, movement and speech functional domains, and two clusters of high and low ability for the tactile and vision functional domains. Descriptive statistics for the univariate clusters are displayed in the Supplemental Material.

Personalized Profiles

To create a personalized profile of development for every child, the information from the individual posterior ability distributions from the MIRT and the results of the multivariate and univariate clustering were combined. Figure 5 displays six of these profiles. The top, middle, and bottom rows of Figure 5 each contain two children who were classified into the high-, moderate-, and low-ability multivariate clusters,

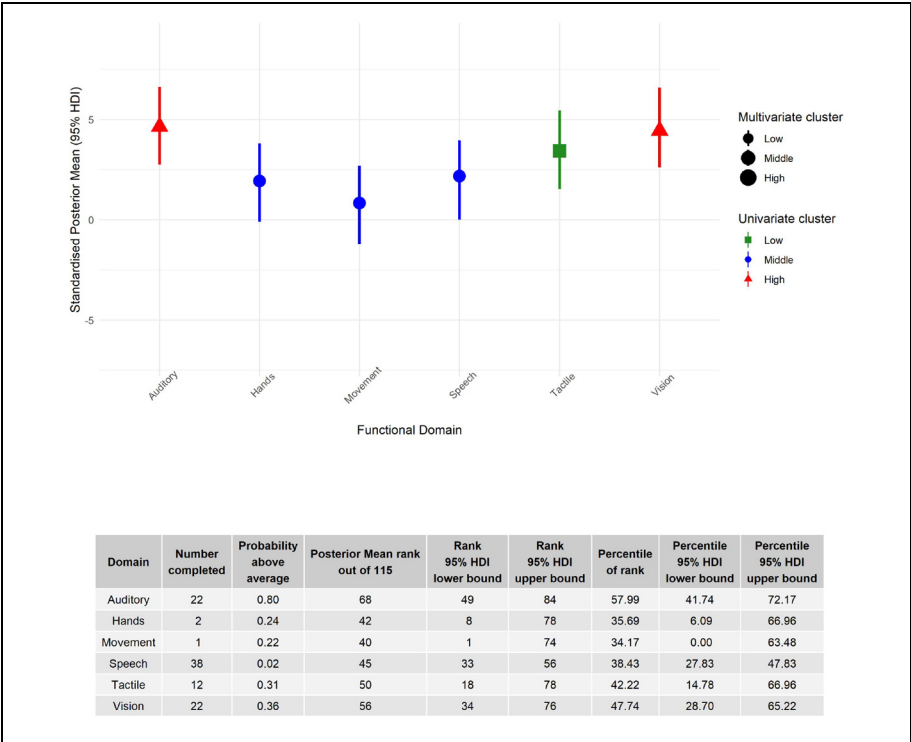


Figure 6. Example of a child's profile with additional information from the posterior distribution.

respectively. When comparing the two children within each multivariate cluster, it can be seen that the children in the left column have similar ability across the six functional domains (indicated by the univariate cluster color and shape); however, the children in the right column have variable ability across the six functional domains. This information is useful for identifying which specific functional areas need to be targeted for early intervention.

In addition, by virtue of the MCMC, probability statements are able to be made from the posterior distributions of the parameter estimates. These probability statements can be made for each child and this personalized information can be included in the profile. Figure 6 contains an example of such information. The table in Figure 6 contains the number of milestones the child has completed for each functional domain, the probability that the child is above the population average for each domain, the mean rank and the percentile of the mean rank, as well as the corresponding 95% highest density intervals (HDI). By reporting the 95% HDI, the uncertainty in the estimates can be conveyed. For example, in Figure 6, although this child has a similar rank score for the movement and speech domains, the 95% HDI is

much wider for the movement domain given that this child has only completed one milestone for the movement domain.

Discussion

This research has demonstrated how personalized developmental profiles can be created through the use of multidimensional item response modeling, embedded in a Bayesian hierarchical framework, applied to a small sample of children with sparse information. Although the sample size was small and the measurements were sparse, by using a Bayesian hierarchical modeling framework, posterior estimates of ability were able to be constructed by borrowing strength from similar children. In addition, by combining the information obtained from clustering the posterior means of the ability estimates, this model provided estimates at the individual, subgroup, and population levels, providing a comprehensive picture of development for this population of children.

In this research, the individual-level posterior distributions of the ability parameters for each child not only show how each child is developing within each functional domain but also provide an estimate of uncertainty in this prediction. Measuring and communicating uncertainty in the estimates is especially important when using small sparse samples, so that parents and clinicians do not place undue confidence in the results for those with sparse data (Fischhoff & Davis, 2014). Communicating statistical uncertainty to nonscientific audiences is difficult and is often omitted from visualizations (Roberts & Gough, 2016). In this research, effort was made to both account for and communicate the uncertainty of the estimates. However, a 95% HDI may not be intuitive to understand for parents. Therefore, future research directions include appropriate approaches for visualizing and communicating uncertainty, for example, through the use of color coding or an uncertainty score.

This model also provided information on each milestone, through the posterior distributions for the difficulty and discrimination parameters. Probabilistically ranking the item parameters identified the developmental milestones that were both difficult and easy to achieve. In addition, milestones that were and were not discriminating between those with high and low ability were also identified. In particular, locating nondiscriminatory items can be used to improve the surveillance tool by removing or revising these redundant items.

The individual results of the Bayesian hierarchical MIRT model were able to be further explored by clustering the standardized posterior means of the individual ability parameters. By clustering the latent trait estimates, subgroups were able to be identified using the whole sample, as posterior ability estimates were obtained for all children, regardless of the number of milestones they had responded to. In addition, using the latent trait estimates greatly reduced the complexity of the clustering problem by reducing the dimensionality from 348 binary milestone measurements to six latent trait estimates. Both multivariate and univariate clustering were explored to identify clusters of average ability across all domains and ability within each domain,

respectively. Through this process, children with low ability can be identified through the multivariate clustering, and then personalized interventions could be tailored to specific functional domains flagged as low ability identified through the univariate clustering.

This methodology has a number of limitations. First, no covariates were available for this sample of children. Incorporating covariates into a multilevel IRT model can help explain the differences in ability between individuals (Fox, 2004). A number of covariates have been identified that are associated with developmental delay, such as poverty, caregiver cognitive impairment, and low parental education (Scarborough et al., 2009). Incorporating covariates such as these into the model may help explain the differences in ability between children. Notwithstanding this, there is interest in fitting a model with no such covariates to evaluate the intrinsic magnitude and variation between ability scores of the children and difficulty of the milestones. Second, this study performed the clustering as a secondary modeling step. Mixture IRT models do exist, but they have limitations in terms of computation time and convergence problems and are not effective for small sample sizes (Finch & French, 2012). As the model was already complex, with more than 1,400 individual, item, and population parameters being estimated using a small, very sparse sample, adding in the mixture components would have exponentially increased the number of parameters to be estimated and convergence of such a model would have been very unlikely.

The current methodology is cross-sectional in nature, which means the model would need to be refit to obtain updated estimates of the parameters when new information is collected. However, one advantage of the Bayesian hierarchical MCMC sampling scheme is the richness of information provided in the parameter posterior distributions obtained from the MCMC samples. This article demonstrated how additional personalized assessments can be derived from the MCMC samples. In addition, this model can also be used to predict milestones that were not observed, by probabilistically imputing the estimates for these unobserved values based on the current population and individual parameter estimates.

Previous literature has advised that consideration should be taken in the choice of prior distribution when using Bayesian estimation for small sample sizes, as this can result in biased estimates when the prior is incorrectly specified (McNeish, 2016; Smid, McNeish, Miočević, & van de Schoot, 2020). This issue was addressed by using Bayesian hierarchical priors for the multivariate ability estimates and the discrimination parameter of the milestones. When using a Bayesian hierarchical prior, the hyperparameters of the first-level parameters (e.g., the children's abilities and the discrimination of the milestones) are not specified directly but are given prior distributions (König et al., 2020). This means that only the prior distributions for the hyperparameters are required to be specified, which alleviates the problem of having to specify informative priors when using a nonhierarchical model (König et al., 2020). Sheng (2012, p. 28) states that "hierarchical priors offer flexibility to specify weakly informative priors on the hyperparameters" and that using hierarchical priors allows for a more objective approach. Natesan et al. (2016) also recommends the use of a

hierarchical prior over informative priors for small sample sizes. Small samples and sparse data are often unavoidable in clinical settings, such as the application in this article, and this research has demonstrated additional support for the use of Bayesian hierarchical priors in situations where the sample is small and the data are sparse.

In summary, this research demonstrates a new approach to create personalized developmental profiles for children through the use of Bayesian hierarchical MIRT modeling. The use of the Bayesian hierarchical modeling framework overcame estimation problems that often occur when using small, sparse samples. In addition, this framework allowed for probabilistic statements to be made regarding both the milestone parameters and the children's abilities so that personalized and detailed information regarding each child's development could be provided. Through incorporating this information from the IRT modeling with the results of the clustering, detailed developmental profiles were able to be constructed. The developmental profiles produced from this research can be used to assist with identification and personalized early intervention for children who are showing signs of developmental delay. These developmental profiles complement developmental surveillance tools by offering personalized feedback on the developmental progress of a child, so that parental concerns can be addressed as soon as a delay is recognized.

Acknowledgments

The data used in this article were generously provided by The Developing Foundation.


Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research is jointly supported by a Queensland University of Technology (QUT) Postgraduate Research Award and an Australian Technology Network Industry Doctoral Training Centre Scholarship co-funded by QUT and The Developing Foundation.

ORCID iD

Patricia Gilholm  <https://orcid.org/0000-0001-8135-1520>

Supplemental Material

Supplemental material for this article is available online.

References

- Adams, R. J., Wilson, M., & Wang, W.-C. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement*, 21(1), 1-23. <https://doi.org/10.1177/0146621697211001>

- Ayoub, C. C., & Fischer, K. W. (2006). Developmental pathways and intersections among domains of development. In K. McCartney, & D. Phillips (Eds.), *Handbook of early child development* (pp.62-81). Blackwell. <https://doi.org/10.1002/9780470757703.ch4>
- Betancourt, M., & Girolami, M. (2015). Hamiltonian Monte Carlo for hierarchical models. *Current Trends in Bayesian Methodology With Applications*, 79(30), 2-4. <https://doi.org/10.1201/b18502-5>
- Bürkner, P.-C. (2019). *Bayesian item response modelling in R with brms and Stan*. arXiv preprint arXiv:1905.09501. <https://arxiv.org/pdf/1905.09501.pdf>
- Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., & Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, 76(1). <https://doi.org/10.18637/jss.v076.i01>
- Chen, C.-Y., Xie, H., Clifford, J., Chen, C.-I., & Squires, J. (2018). Examining internal structures of a developmental measure using multidimensional item response theory. *Journal of Early Intervention*, 40(4), 287-303. <https://doi.org/10.1177/1053815118788063>
- Cheng, Y.-Y., Wang, W.-C., & Ho, Y.-H. (2009). Multidimensional Rasch analysis of a psychological test with multiple subtests: A statistical solution for the bandwidth–fidelity dilemma. *Educational and Psychological Measurement*, 69(3), 369-388. <https://doi.org/10.1177/0013164408323241>
- Curtis, S. M. (2010). BUGS code for item response theory. *Journal of Statistical Software*, 36(1), 1-34. <https://doi.org/10.18637/jss.v036.c01>
- de Ayala, R. J. (2013). *The theory and practice of item response theory*. Guilford Press.
- de la Torre, J., & Hong, Y. (2010). Parameter estimation with small sample size a higher-order IRT model approach. *Applied Psychological Measurement*, 34(4), 267-285. <https://doi.org/10.1177/0146621608329501>
- de la Torre, J., & Patz, R. J. (2005). Making the most of what we have: A practical application of multidimensional item response theory in test scoring. *Journal of Educational and Behavioral Statistics*, 30(3), 295-311. <https://doi.org/10.3102/10769986030003295>
- de Lourdes Drachler, M., Marshall, T., & De Carvalho Leite, J. C. (2007). A continuous-scale measure of child development for population-based epidemiological surveys: A preliminary study using item response theory for the Denver Test. *Paediatric and Perinatal Epidemiology*, 21(2), 138-153. https://doi.org/10.1111/j.1365-2214.2007.00774_6.x
- Demiriz, A. (2004). Enhancing product recommender systems on sparse binary data. *Data Mining and Knowledge Discovery*, 9(2), 147-170. <https://doi.org/10.1023/b:dam.0000031629.31935.ac>
- The Developing Foundation. (2017). *Help your child develop important milestones*. (<https://www.developingfoundation.org.au/milestone-tracking-enhancement.html> [Accessed: 11 .06.2020])
- Finch, W. H., & French, B. F. (2012). Parameter estimation with mixture item response theory models: A Monte Carlo comparison of maximum likelihood and Bayesian methods. *Journal of Modern Applied Statistical Methods*, 11(1), 14. <https://doi.org/10.22237/jmasm/1335845580>
- Fischhoff, B., & Davis, A. L. (2014). Communicating scientific uncertainty. *Proceedings of the National Academy of Sciences of the U S A*, 111(Suppl. 4), 13664-13671. <https://doi.org/10.1073/pnas.1317504111>
- Fox, J.-P. (2004). Applications of multilevel IRT modeling. *School Effectiveness and School Improvement*, 15(3-4), 261-280. <https://doi.org/10.1080/09243450512331383212>

- Fox, J.-P. (2010). *Bayesian item response modeling: Theory and applications*. Springer Science & Business Media.
- Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis*. CRC Press.
- Gilholm, P. (2020). *Supplementary MIRT developmental surveillance*. (https://github.com/TrishG89/supplementary_MIRT_developmental_surveillance)
- Gilholm, P., Mengersen, K., & Thompson, H. (2020). Identifying latent subgroups of children with developmental delay using Bayesian sequential updating and Dirichlet process mixture modelling. *PLOS ONE*, 15(6), Article e0233542.
- Guralnick, M. J., & Bruder, M. B. (2019). Early intervention. In J. L. Matson (Ed.), *Handbook of intellectual disabilities* (pp. 717-741). Springer.
- Haley, S. M., Coster, W. J., Kao, Y.-C., Dumas, H. M., Fragala-Pinkham, M. A., Kramer, J. M., Ludlow, L. H., & Moed, R. (2010). Lessons from use of the Pediatric Evaluation of Disability Inventory (PEDI): Where do we go from here? *Pediatric Physical Therapy*, 22(1), 69-75. <https://doi.org/10.1097/PEP.0b013e3181cbfbf6>
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning (Vol. 112)*. Springer.
- König, C., Spoden, C., & Frey, A. (2020). An optimized Bayesian hierarchical two-parameter logistic model for small-sample item calibration. *Applied Psychological Measurement*, 44(4), 311-326. <https://doi.org/10.1177/0146621619893786>
- Lancaster, G. A., McCray, G., Kariger, P., Dua, T., Titman, A., Chandna, J., McCoy, D., Abubakar, A., Hamadani, J. D., Fink, G., Tofail, F., Gladstone, M., & Janus, M. (2018). Creation of the WHO indicators of Infant and Young Child Development (IYCD): Metadata synthesis across 10 countries. *BMJ global health*, 3(5), e000747. <https://doi.org/10.1136/bmjgh-2018-000747>
- Malika, C., Ghazzali, N., Boiteau, V., & Niknafs, A. (2014). NbClust: An R package for determining the relevant number of clusters in a data set. *Journal of Statistical Software*, 61, 1-36. <https://doi.org/10.18637/jss.v061.i06>
- McNeish, D. (2016). On using Bayesian methods to address small sample problems. *Structural Equation Modeling: A Multidisciplinary Journal*, 23(5), 750-773. <https://doi.org/10.1080/10705511.2016.1186549>
- Monnahan, C. C., Thorson, J. T., & Branch, T. A. (2017). Faster estimation of Bayesian models in ecology using Hamiltonian Monte Carlo. *Methods in Ecology and Evolution*, 8(3), 339-348. <https://doi.org/10.1111/2041-210x.12681>
- Murtagh, F., & Legendre, P. (2014). Ward's hierarchical agglomerative clustering method: Which algorithms implement Ward's criterion? *Journal of Classification*, 31(3), 274-295. <https://doi.org/10.1007/s00357-014-9161-z>
- Natesan, P., Nandakumar, R., Minka, T., & Rubright, J. D. (2016). Bayesian prior choice in IRT estimation using MCMC and variational Bayes. *Frontiers in Psychology*, 7, 1422. <https://doi.org/10.3389/fpsyg.2016.01422>
- Ntzoufras, I. (2011). *Bayesian modeling using winbugs (Vol. 698)*. Wiley.
- Powell, A., Davies, H., & Thomson, R. (2003). Using routine comparative data to assess the quality of health care: Understanding and avoiding common pitfalls. *BMJ Quality & Safety*, 12(2), 122-128. <https://doi.org/10.1136/qhc.12.2.122>
- R Core Team. (2019). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from <https://www.R-project.org/>

- Reeve, B. B., & Fayers, P. (2005). Applying item response theory modeling for evaluating questionnaire item and scale properties. *Assessing Quality of Life in Clinical Trials: Methods of Practice*, 2, 55-73.
- Roberts, J., & Gough, P. (2016). Communicating statistical uncertainty to non-expert audiences: Interactive disease mapping. In *2016 Big data visual analytics (BDVA)* (pp. 1-3). <https://doi.org/10.1109/bdva.2016.7787045>
- Rose, N., von Davier, M., & Nagengast, B. (2015). Commonalities and differences in IRT-based methods for nonignorable item nonresponses. *Psychological Test and Assessment Modeling*, 57(4), 472-498.
- Scarborough, A. A., Lloyd, E. C., & Barth, R. P. (2009). Maltreated infants and toddlers: Predictors of developmental delay. *Journal of Developmental & Behavioral Pediatrics*, 30(6), 489-498. <https://doi.org/10.1097/dbp.0b013e3181c35df6>
- Sheng, Y. (2012). An empirical investigation of Bayesian hierarchical modeling with unidimensional IRT models. *Behaviormetrika*, 40(1), 19-40. <https://doi.org/10.2333/bhmk.40.19>
- Sheng, Y., & Wikle, C. K. (2007). Comparing multiunidimensional and unidimensional item response theory models. *Educational and Psychological Measurement*, 67(6), 899-919. <https://doi.org/10.1177/0013164406296977>
- Sheng, Y., & Wikle, C. K. (2008). Bayesian multidimensional IRT models with a hierarchical structure. *Educational and psychological measurement*, 68(3), 413-430. <https://doi.org/10.1177/0013164407308512>
- Smid, S. C., McNeish, D., Miočević, M., & van de Schoot, R. (2020). Bayesian versus frequentist estimation for structural equation models in small sample contexts: A systematic review. *Structural Equation Modeling: A Multidisciplinary Journal*, 27(1), 131-161. <https://doi.org/10.1080/10705511.2019.1577140>
- Stan Development Team. (2020). *RStan: The R interface to Stan*. Retrieved from <http://mc-stan.org/> (R package version 2.19.3)
- Zhang, J. (2012). Calibration of response data using MIRT models with simple and mixed structures. *Applied Psychological Measurement*, 36(5), 375-398. <https://doi.org/10.1177/0146621612445904>