Operating Characteristics and Test Information of the Ages and Stages Questionnaire

Adon Rosen

December 14, 2022

Introduction

Developmental and behavioral difficulties (DBDs) cost society an estimated $250 billion per year (Bradley, 2003). With DBD prevalence on the rise, the economic gradient is steepening (Boyle et al., 2011). When unnoticed and untreated, the price of DBDs and the number of ensuing negative impacts increases. Early interventions have proven effective at remediating and preventing many DBDs, but resources for early identification are limited (Rosenberg, Zhang, & Robinson, 2008). Screening tools such as the Ages and Stages Questionnaire (ASQ) are frequently employed to identify individuals at-risk of DBDs (Bricker & Squires, 1989). Screening tools, such as the ASQ, seek to balance ease of implementation, as well as sensitivity to the presence of DBDs in order successfully identify children at-risk of DBDs. Psychometric validation is commonly used on screener tools to validate the sensitivity of these tools; however, these validation techniques typically employ factor analytic strategies which ignore the implementation of the tools, or use reliability assessments which are insensitive to the tails of the distribution (children at-risk of having DBDs). This study seeks to explore the psychometric properties of the ASQ by proposing a technique which is sensitive to the scale's ability to identify individual's who are at-risk of possessing DBDs.

The ASQ is comprised of five main subscales evaluating development in Communication, Gross Motor, Fine Motor, Problem Solving, and Personal-Social abilities (Squires, Twombly, Bricker,

& LaWanda, 2009). Within each subscale, there are 20 versions of the ASQ, the version a child receives is determined by the child's age in months. Every version has six partial credit questions. As intentioned, these screening items target behaviors and skills at the middle to low-end of each ability scale. Most current psychometrics are based on a sample of 15,138 unique children who completed 18,572 questionnaires (3,434 individuals completed 1+). Two-week test-retest correlations on the order of .75 to .82 (n=145), inter-observer correlations ranging from 0.43 to 0.69 (n=107), and internal consistency correlations between 0.60 and 0.84 support reliability. There is also relatively good support for construct validity of the measure's categorical ratings of risk with sensitivity and specificity estimates exceeding 0.83 when discriminating at-risk samples from the general population. Additional explorations of the ASQ have also used measurement models such as factor analysis to confirm the performance of the tool.

When exploring the factor structure of the ASQ commonly exploratory factor analytic techniques have been applied. The goal of applying an exploratory model tests the belief that every factor scale will map onto the subscale of interest. Recent explorations of the factor structure of the ASQ has supported the stability of the items within the subscale and over time (Olvera Astivia, Forer, Dueker, Cowling, & Guhn, 2017). While the application of an exploratory model confirms the items load onto the latent trait, a more applicable model for how the ASQ is being implemented would be a confirmatory factor model (CFA). Through a CFA framework, it enforces the items to act onto the latent trait they are being to used to test. A popular CFA model to test the quality of dichotomous or graded response items, such as those that the ASQ employs, is an IRT model (Embretson & Reise, 2000).

An IRT model can be described as:

$$P_i(\theta) = \frac{1}{1 + e^{-a_i(\theta - b_i)}}$$

In this formula, P_i represents the probability an item is endorsed, θ is the ability estimate of an individual, a_i represents the discrimination estimate for item i, and b_i represents the difficulty estimate for item i. The above characteristics can be used to map the item characteristic curve (ICC) a graphical representation of the item's characteristics. An IRT model which includes partial credit

is typically called a graded response model (GRM). The GRM extends the formulation of the binary IRT model by including probabilities for each potential endorsement:

$$P(x_i = k|\theta) = \frac{1}{1 + e^{-a_i(\theta - b_{ik})}} - \frac{1}{1 + e^{-a_i(\theta - b_{i(k+1)})}}$$

This formula now includes a difficulty parameter and endorsement probability for specific response values, k for every item i. By estimating the characteristics using this GRM framework both the implementation, and the underlying theory of the ASQ are more closely adhered. Through the model estimation, by estimating only a discrimination parameter across the entire scale this mimics the sum-score approach employed in the ASQ (McNeish & Wolf, 2020). By using the CFA framework, it also adheres to the fact that every item only maps onto a single subscale. Finally, the greatest benefit of applying an IRT estimation framework is the ability to estimate quality of an item and a scale for an indivdual.

Within the IRT framework, "quality" is determined by the amount of information an item possess for an individual. The amount of infomration an item possess is estimated by an indivudla's esitmated ability and an item's characteristics. These item characteristics are used to calculate an item's information at specific levels of theta (Louis, 1982; Monroe, 2019). The information for an item adheres to the following formula:

$$I_i(\theta) = a_i^2 P_i(\theta) Q_i(\theta)$$

Where $I_i(\theta)$ is the information produced by a item i, a_i is the item's discrimination, $P_i(\theta)$ is the probability for endorsement, and $Q_i(\theta)$ is the probability of non endorsement. This formulation indicates several characteristics about the amount of information an item possess: first, as a_i increases the amount of information increases, a larger discrimination parameter is almost always desirable, second, information is the greatest amount of information exists at the inflection points of an item's characteristic curve, this means that information is maximized when an examinee has a 50% chance of endorsement. Finally, information varies per person.

The test information function sums across every item's information function and takes the following

form:

$$I_t(\theta) = \sum_{i=1}^{N} I_i(\theta)$$

Where $I_t(\theta)$ represents the amount of information a test possess at a specific ability level. Finally, the total relevant information takes the integral of the test information function based on the cutoff values taken at the latent ability level. Motivated by methodology of the ASQ, the scoring cutoff explores any individual two standard deviations below the mean score, so the total relevant information for a high-risk classification from an ASQ administration takes the following form:

$$I_r = \int_{-\infty}^{-2} \sum_{i=1}^{N} I_i(\theta) d(\theta)$$

Test information is frequently used in computerized adaptive testing where the goal is to maximize the amount of information for an examinee by modifying the item bank for every individual (Moore et al., 2019). By extending this methodology to a classification problem, it allows insights into the tests ability to identify individual's within a predetermined θ range. This alternative estimation of test reliability mitigates the concerns of alternative reliability estimates which explore scale reliability over the entire θ range. By applying these methods, and estimating the total relevant information to an developmental screener it allows researchers an assessment to the quality of the tool, and the likelihood of misclassification of a DBD.

Methods

The goal of this study is to explore if total relevant information from the ASQ-3 could be used to identify high-risk classification inconsistencies among real children. In order to perform this, several tasks had to be completed. First, the reliability and intra-class correlation of ASQ administrations were explored as a function of time. Second, a IRT estimates were estimated following a GRM formulation for every version within each subscale of the ASQ-3. Third, the total relevant information was estimated using these GRM parameter estimates. Fourth and finally, relationships between total relevant information and classification inconsistencies are explored.

Participants

Parents receiving home-based parenting services were identified via the Oklahoma State Department of Health's Efforts-to-Outcomes administrative database and contacted using mailed letters, phone calls, in-person, and electronic communication procedures. Parents who wished to participate were provided with additional study information and research assistants scheduled a time to complete the interview. Surveys primarily took place in the home with the index child present. Upon participant request, some surveys were conducted in a private location other than the home, such as a library, or virtually due to COVID-19 restrictions. Research assistants obtained informed consent, then the survey was administered through REDCap, a secure web-based application for data collection (Harris et al., 2009). The survey consisted of a battery of measures aimed at parent and child health and well-being, parent-child interactions, and child development. Specific measures used for the present analyses are described in detail below. Participants' responses were confidential; however, research assistants were available to answer questions or help with survey administration if requested. Surveys lasted approximately 1-3 hours, and participants were compensated for their time.

ASQ reliability

The first set of analyses explores how consistent the reliability of the ASQ-3 is as the time between administrations increases. By exploring the variability of reliability as a function of time it can inform downstream analyses if the differences in identification are driven by diminishing reliability. In order to assess this, an initial caliper of 30 days was applied, any serial ASQ administrations that were performed within the caliper had a Krippendorff's α and an ICC calculated. The ICC was estimated assuming multiple random raters as the ASQ may have been administered to any of the child's care providers. The caliper was extended by increments of 5, ranging from 30 to a total of 200 days separation in the administration. This was performed separately for the z-scored ASQ scores as well as the risk-categories assigned in an ASQ administration. For the z-scored values the Interval approached was used to estimate the Krippendorff's α whereas the ordinal approach was used for the risk-category.

Total Relevant Information for the ASQ-3

In order to obtain a domain and version specific total relevant information two tasks had to be performed. First a GRM parameters have to be estimated for every version and subscale, second, these GRM characteristics must be used to identify the test information functions, third and finally, these test information functions are used to identify the total relevant information. The estimation of an IRT model was motivated by the ASQ scoring practices. That is, because sum-scores are utilized in an ASQ, a fixed discrimination parameter was estimated within each version within each domain of the ASQ (McNeish & Wolf, 2020). Accordingly, when an IRT model was estimated a single discrimination parameter was estimated as well as two difficulty parameters for every item for a partial and complete endorsement of an item. This requires 12 difficulty parameters and a single discrimination parameter to be estimated. These models were estimated using the `mirt` (Chalmers, 2012) package in R version 4.2.1 (R Core Team, 2020). When no endorsements of an item were present, or complete full endorsement of an item was present, the item's difficulty was assigned to -4, or 4 respectively.

Next, the test information function is estimated using the parameter estimates identified within every domain and version. The test information function was calculated using `testinfo` function from the `mirt` package. The `testinfo` returns a series of point estimates of information at predetermined levels of theta, in order to calculate the integral of this function the `integrate` function was used from `R`. The integrated between $-\infty$ and -2 was calculated for every domain and version for these test information functions.

Modeling High-risk Inconsistencies Between Serial ASQ Administrations

In order to model the high-risk inconsistencies between serial administrations of an ASQ as a function of relevant information, several tasks had to be performed. First, participants who had received serial administrations of an ASQ within 100 days were identified. Next, the outcome, agreement in high-risk status, was created for every participant. An agreement was identified if both ASQ administrations identified the participant as either high-risk or not high-risk. That is, if an individual had a first administration ASQ risk category of normal development and the second administration they received a score within the monitoring range, this was classified as

an agreement between the two ASQ administrations. However, if the first ASQ administration scored the individual as having normal development and the second scored them as having a high-risk of developmental delay, this was classified as a disagreement. To test for agreement, a generalized linear mixed effects model was estimated with a logistic link function. The outcome agreement was modeled as a function of the total relevant information from the first and second ASQ administrations and the interaction of these two values. The model also controlled for the version of the first administration and the time between the administrations in days. A random intercept was included for every participant, as every participant had at least 5 agreement statuses, one for every ASQ domain. The model was trained using the `lme4` package (Bates, Mächler, Bolker, & Walker, 2015). The formula takes the following structural format:

$$Y_{st} = \beta_{0s} + \beta_1 \times X_{TRI1st} + \beta_2 \times X_{TRI2st} + \beta_3 \times X_{TRI1st*TRI2st} + \beta_c \times X_c + e_{st}$$

Where Y_{st} represents high-risk agreement, β_{0s} is an individual specific intercept term, β_1, β_2 are the fixed effects for the total relevant information, with X_{TRI1st}, X_{TRI2st} are the the total relevant information values for individual s at time t. The β_c and X_c represent fixed effects for covariates such as time between administration and child age at initial administration. Finally, e_{st} represents an individual specific error term.

Counter Factual Examination of at-risk Inconsistency

Finally, based on the results from the inconsistency analysis an optimal administration timeline will be explored. By using a counter factual exploration it will allow for the time between serial administrations dependent on child's age and an ASQ score to be estimated. The following structural model will be estimated

$$y^*_{st} = \lambda_v(\gamma' x + \zeta) + e_{st}$$

where y^*_{st} reflects the predicted probability of agreement between two serial administrations, λ_v reflects the discrimination parameters for a specific ASQ version, γ is a vector of regression coefficients, x_{st} is a vector of manifest variables, ζ_{st} is a disturbance term and e_{st} is a residual variable.

This model formulation adheres to a multiple indicator multiple causes model (Muthén, 1984). The goal of these analyses is to identity the optimal time between administrations which minimizes the risk of a false positive. In order to perform this, parameter estimates will be obtained using the entire population and individual traits will be modified for an individual's manifest variables including their ASQ score, and their time between administration. By varying an individual's ASQ score it allows for an examination into the relationship between the standard error of measurement and misclassification. Specifically, individual's who may have a version with lower information, and therefore a greater standard error of measurement, are more likely to exhibit missclassification when the score is closer to an at-risk threshold. By varying this in conjunction with the time between administration it will guide the best time to perform a follow up examination to confirm a at-risk status.

Anticipated Results

ASQ Reliability

As time between administrations of an ASQ increase, it is estimated that the reliability of the scale should decrease. Because the ASQ seeks to measure an individual's developmental status, and development is a moving target littered with individualized trajectories overtime. An additional concern for the reliability of the scale is the amount of information changes over ASQ versions. SO the quality of the exam will vary overtime. These things altogether contribute to a diminishing reliability of the ASQ as a function of time.

Modeling High-risk Inconsistencies Between Serial ASQ Administrations

It is anticipated that as the difference between the total relevant information between serial examinations increases, the probability of an inconsistency will increase. As the total relevant information is an assessment of scale quality, when the difference between the quality of the scale increases it becomes more likely that the scale will identify differences.

Counter Factual Examination of at-risk Inconsistency

The counter factual examination seeks to identify the likelihood an individual is at-risk based on their score. In order to perform this the time between administration will be varied as well as an individual's ASQ score. The motivation of these analyses is to increase the confidence in an individual's at-risk assessment while also lowering the burden in downstream referral to developmental promotion providers. It is theorized that when an individual receives a score further from the at-risk threshold (i.e. $\theta < -2$) than their time between administration to confirm the are at-risk of DBDs. However, when an individual is closer to the threshold then a serial assessment may reveal they graduate to a lower risk assessment. The anticipated results will reveal a inverse quadratic relationship, as an individual is further from the threshold serial assessments to confirm the original status will need to be closer and when an individual is closer to the threshold serial assessments should be further separated to allow time for development between assessments.

Bibliography

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, *67*, 1–48. doi:10.18637/jss.v067.i01

Boyle, C. A., Boulet, S., Schieve, L. A., Cohen, R. A., Blumberg, S. J., Yeargin-Allsopp, M., … Kogan, M. D. (2011). Trends in the prevalence of developmental disabilities in US children, 1997-2008. *Pediatrics*, *127*(6), 1034–1042. doi:10.1542/peds.2010-2989

Bradley, S. J. (2003). Handbook of early childhood intervention. Second edition. *The Canadian Child and Adolescent Psychiatry Review*, *12*(4), 122. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2533836/

Bricker, D., & Squires, J. (1989). The Effectiveness of Parental Screening of At-Risk Infants: The Infant Monitoring Questionnaires. *Topics in Early Childhood Special Education*, *9*(3), 67–85. doi:10.1177/027112148900900306

Chalmers, R. P. (2012). mirt: A Multidimensional Item Response Theory Package for the R Environment. *Journal of Statistical Software*, *48*, 1–29. doi:10.18637/jss.v048.i06

Embretson, S. E., & Reise, S. P. (2000). *Item response theory*. New York: Psychology Press. doi:10.4324/9781410605269

Harris, P. A., Taylor, R., Thielke, R., Payne, J., Gonzalez, N., & Conde, J. G. (2009). Research electronic data capture (REDCap)–a metadata-driven methodology and workflow process for providing translational research informatics support. *Journal of Biomedical Informatics*, *42*(2), 377–381. doi:10.1016/j.jbi.2008.08.010

Louis, T. A. (1982). Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, *44*(2), 226–233. Retrieved from https://www.jstor.org/stable/2345828

McNeish, D., & Wolf, M. G. (2020). Thinking twice about sum scores. *Behavior Research Methods*, *52*(6), 2287–2305. doi:10.3758/s13428-020-01398-0

Monroe, S. (2019). Estimation of Expected Fisher Information for IRT Models. *Journal of Educational and Behavioral Statistics*, *44*(4), 431–447. doi:10.3102/1076998619838240

Moore, T. M., Calkins, M. E., Satterthwaite, T. D., Roalf, D. R., Rosen, A. F., Gur, R. C., & Gur, R. E. (2019). Development of a computerized adaptive screening tool for overall psychopathology ("p"). *Journal of Psychiatric Research*, *116*, 2633.

Muthén, B. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika*, *49*(1), 115–132. doi:10.1007/BF02294210

Olvera Astivia, O. L., Forer, B., Dueker, G. L., Cowling, C., & Guhn, M. (2017). The Ages and Stages Questionnaire: Latent factor structure and growth of latent mean scores over time. *Early Human Development*, *115*, 99–109. doi:10.1016/j.earlhumdev.2017.10.002

R Core Team. (2020). *R: A language and environment for statistical computing*. Vienna, Austria. Retrieved from https://www.R-project.org/

Rosenberg, S. A., Zhang, D., & Robinson, C. C. (2008). Prevalence of developmental delays and participation in early intervention services for young children. *Pediatrics*, *121*(6), e1503–1509. doi:10.1542/peds.2007-1680

Squires, J., Twombly, E., Bricker, D., & LaWanda, P. (2009). *ASQ®-3 user's guide*. Brookes. Retrieved from https://products.brookespublishing.com/ASQ-3-Users-Guide-P571.aspx