Title:

High-dimensional Explanatory Random Item Effects Models for Rater-mediated Assessments

Authors and Affiliations:

Ben Kelcey
University of Cincinnati
ben.kelcey@gmail.com

Shanshan Wang
University of Cincinnati

Kyle cox
University of Cincinnati

Background / Context:

Valid and reliable measurement of unobserved latent variables is essential to understanding and improving education (Raudenbush & Sadoff, 2008). A common and persistent approach to assessing latent constructs in education is the use of rater inferential judgment (e.g., Eckes, 2009). In rater-mediated assessments, raters conduct evaluations by interpreting evidence (e.g., responses, behaviors) using their trained, but subjective, judgments. For this reason, the use of raters to assign scores has been described as an indirect or rater-mediated process because measurements are not directly observed but rather inferred through raters' judgments (Bejar, Williamson & Mislevy, 2006).

An important assumption underlying meaningful comparisons in rater-mediated assessments is that measurement is invariant across raters. Measurement invariance across raters suggests that raters use items similarly so that the relationships between a latent trait and the manifest items with which it is measured do not depend upon which rater conducted an evaluation. When items function differently across raters, ratings no longer preserve a common meaning and basis for comparison across raters because scales are rater-specific. In this way, the extent to which a common scale can be formed across raters depends largely on the extent to which raters share a common basis for assigning scores.

Research has shown that a significant source of construct-irrelevant variation in many rater-mediated assessments arises from differences among raters in how they apply the standards established by an instrument (e.g., Hill, Charalambous, & Kraft, 2012). Such variations often arise from consistent differences among raters in how they apply standards across all items and participants but they also arise in more complex ways through, for example, multifaceted interactions among raters, items, and participants. Although findings of rater differences are not surprising, the complexity, inconsistency across items and participants, and magnitude relative to construct-relevant variance found by recent reports have demonstrated just how critical of an issue rater variability can be and raises questions about the degree to which scores from different raters are on commensurate scales (Kelcey, McGinn, & Hill, 2014). Despite extensive and consistent evidence of rater differences across a broad array of assessments, scores from different raters are routinely treated as if they were exchangeable across raters and are often used to make high-stakes comparative decisions (e.g., Baumgartner & Steenkamp, 2001).

To address these shortcomings, recent work has developed flexible methods to accommodate measurement noninvariance in rater-mediated assessments through cross-classified random item effects models (Kelcey, McGinn, & Hill, 2014). By leveraging empirical estimates of rater-specific deviations in the item parameters, these methods allow for the empirical identification and direct adjustment for noninvariance. In turn, a common, inter-rater scale can often be established to facilitate comparisons across units assessed by different raters.

Purpose / Objective / Research Question / Focus of Study:

The purpose of this study is to develop high-dimensional explanatory random item effects models designed for rater-mediated assessments. The models are built to address three specific issues. First, an important limitation of the use of cross-classified random item effects models in rater-mediated assessments is that the number of latent dimensions increases quickly with the number of items and facets. In turn, estimation of the latent dimensions quickly becomes computationally challenging and near impossible with a more than a handful of items and facets. For these reasons, the proposed models intentionally draw on the recently developed Metropolis-Hastings Robbins-Monro algorithm to estimate parameters (Cai, 2010).

Second, an important limitation of previous work is that it has not considered the potential for interactions among facets and between facets and items. More specifically, previous work in this area has only considered the potential for noninvariance along a single facet (i.e., noninvariance among raters). However, it is plausible that for many types of assessments, noninvariance may simultaneously exist along multiple facets as well as along interactions among those facets. Our formulation of the proposed model incorporates an n-level crossed random effects approach (i.e., any number of [crossed] levels) that allows for interactions among facets and with items so as to track and accommodate complex sources of noninvariance.

Third, although previous research has examined the presence of noninvariance in rater-mediated assessments, it has not examined the extent to which characteristics of each facet systematically explain variability in measurement. For instance, do raters with more years of experience demonstrate more lenient applications of rating standards or, similarly, are some items systematically more difficult for lower versus upper elementary students? Further, do complex interactions produce additional noninvariance—for example, are more experienced raters more lenient with lower elementary students? To address these explanatory questions we further introduce structural links between random and fixed facet-specific effects. In turn, our formulation draws on a latent regression framework to investigate correlates of noninvariance.

Setting & Participants:

To illustrate the proposed method, we applied it to a longitudinal study of students' academic engagement in elementary school. Two cohorts of students were rated about up to three times per year for three years across grades kindergarten to second or third to fifth. Across the study there were about 6000 students measured about 17,000 times. As a result, there are multiple concurrent facets that may contribute to the observed variation in engagement ratings. For the purposes of our application, the object of measurement or target of our inference is students' persistent engagement across those assessments. As a result, variance owing to other facets is viewed as construct-irrelevant variance. For instance, variance originating from consistent rater differences, short-term temporal instabilities, differences among items, and their interactions is construct-irrelevant variance. In our application, we assess the extent to which each of these facets contributes to measurement noninvariance.

Significance / Novelty of study:

Although strict measurement invariance across raters and other facets is optimal, the reality is that it will rarely hold in complex rater-mediated assessments. Developing measurement models that are more tightly attuned to the types of measurement errors present in rater-mediated assessments is likely to improve the validity and comparability of scores across raters and other sources of construct-irrelevant variation. The proposed method relaxes assumptions of measurement invariance in n-level cross-classified rater-mediated assessments by introducing random item effects to test for noninvariance in each facet and their interactions and empirically construct an inter-rater scale. More conceptually, the approach helps to identify and accommodate differences in how items function within and across facets so as to place measurements from different contexts on a similar scale.

Statistical, Measurement, or Econometric Model:

We explicate an example model based on the aforementioned data structure of the student engagement measure. Let Y_{isr} be the rating for item i measured at time t for student s by rater r. Using a dichotomous (dis)agree rating scale, our formulation begins with the standard two-parameter logistic item response model

$$\text{logit}(Y_{itsr} = 1) = a_i(\theta^*_{ts} - b_{isr}) \quad (1)$$

where a_i represents the discrimination parameter for item *i*, b_{isr} represents the threshold or difficulty of item *i*, and θ^*_{ts} represents the level of academic engagement by student *s* at time *t*.

Next, we expand the scope of this model to take into account construct-irrelevant variance originating from different facets. First, let us expand on the person-side variance by decomposing θ^*_{ts} into a component attributable to student *s*'s persistent level of engagement across time (θ_s), a time-specific component (θ_{ts}) that captures temporal instability or short-term deviations or differences, a rater-specific component (θ_r) that captures the extent to which a rater is uniformly more severe across all items relative to other raters, and a rater-by-student interaction component (γ_{r-s}) capturing the extent to which a rater is uniformly more severe with particular groups of students across all items. Assume that these components have independent normal distributions and $\theta_{ts} \sim N(0,1)$. Equation (1) becomes

$$\theta^*_{ts} = \theta_{ts} + \theta_s + \theta_r + \gamma_{r-s} \quad (2)$$

Next, expand on the item-side variance by tracking how the difficultly parameter (b_{isr}) varies. First, we establish an intercept (b_i^0), then decompose difficulty into a rater-specific component (b_{ir}) that quantifies the extent to which obtaining a positive rating is more difficult when being rated by rater r on item i (this differs from θ_r because b_{ir} captures item-specific differences in severity whereas θ_r captures the average severity differences across all items), a student-specific component (b_{is}) that captures the extent to which obtaining a positive rating on item i is more difficult for student *s* than for an average student (suggests ratings function differently for different students—e.g., those in upper v. lower elementary), and finally a rater-by-student-by-item interaction component (δ_{i-r-s}) that captures the potential for obtaining a positive rating to be more difficult for certain students when rated by rater *r* on item *i* (e.g., a rater is more severe on a particular with upper elementary kids). Assume that these components have independent normal distributions. Equation (1) is further expanded so that

$$b_{isr} = b_i^0 + b_{ir} + b_{is} + \delta_{i-r-s} \quad (3)$$

The model thus far has outlined person-side random effects (2) and item-side random effects (3) but we have not incorporated explanatory components. Each parameter can be associated with fixed effects through models that provide structural links. These structural links attempt to model why we observe variation originating from each facet (e.g., why are some raters more severe in rating engagement. To introduce explanatory components, we expand so that

$$\theta_{ts} = \beta_0^{\theta_{ts}} + \sum_{n=1}^{N} \beta_n^{\theta_{ts}} X_{nts} + \theta_{ts}^{\varepsilon} \qquad \theta_s = \beta_0^{\theta_s} + \sum_{j=1}^{J} \beta_j^{\theta_s} W_{js} + \theta_s^{\varepsilon}$$
$$\theta_r = \beta_0^{\theta_r} + \sum_{k=1}^{K} \beta_k^{\theta_r} Z_{kr} + \theta_r^{\varepsilon} \qquad \gamma_{r-s} = \beta_0^{\gamma_{r-s}} + \sum_{m=1}^{M} \beta_k^{\gamma_{r-s}} (WZ)_{msr} + \gamma_{r-s}^{\varepsilon} \quad (4)$$

Here the β_0's are the intercepts, $\beta_\cdot$'s are the latent regression coefficients capturing the extent to which a predictor (denoted as X, W, and Z), such as student gender, is associated with variation in a specific facet, and $\theta_\cdot^{\varepsilon}$'s (and $\gamma_{r-s}^{\varepsilon}$) are the unexplained residual variation in the components with continued independent normal distributions. Let X_{nts} be time-varying student predictors (e.g., age a time of assessment), W_{js} be time-invariant student predictors (e.g., gender of student), Z_{kr} be rater predictors (e.g., gender of rater), and $(WZ)_{msr}$ are the interactions formed by the product of time-invariant student predictors and rater predictors.

Next introduce item-side explanatory structural models for each random item effect associated with the difficulty parameter. We have

$$b_i^0 = \beta_0^{b_0} + \sum_{l=1}^{L} \beta_l^{b_0} F_{li} + b_i^{\varepsilon} \qquad b_{is} = \beta_0^{b_{is}} + \sum_{q=1}^{Q} \beta_q^{b_{is}} W_{qs} + b_{is}^{\varepsilon}$$
$$b_{ir} = \beta_0^{b_{ir}} + \sum_{p=1}^{P} \beta_p^{b_{ir}} Z_{pr} + b_{ir}^{\varepsilon} \qquad \delta_{i-r-s} = \beta_0^{\delta_{i-r-s}} + \sum_{u=1}^{U} \beta_u^{\delta_{i-r-s}} (WZ)_{usr} + \delta_{i-r-s}^{\varepsilon} \quad (5)$$

Here the β_0's are the intercepts, $\beta_{\cdot}$'s are the latent regression coefficients capturing the extent to which a predictor (denoted as F, W, and Z), such as student gender, is associated with variation in a specific facet, and $b_{\cdot}^{\varepsilon}$'s (and $\delta_{i-r-s}^{\varepsilon}$) are the unexplained residual variation in the difficulty components with continued independent normal distributions. Let F_{li} be item predictors (e.g., negatively worded item or other characteristic of item), W_{qs} be time-invariant student predictors (e.g., gender of student), Z_{pr} be rater predictors (e.g., gender of rater), and $(WZ)_{msr}$ are the interactions formed by the product of time-invariant student predictors and rater predictors. Substituting equations leaves us with

$$\text{logit}(Y_{itsr}=1) = a_i([\theta_{ts}^{\varepsilon} + X\beta_i] + [\theta_s^{\varepsilon} + W\beta_s] + [\theta_r^{\varepsilon} + Z\beta_r] + [\gamma_{r-s}^{\varepsilon} + WZ\beta_{r-s}] - \{[b_i^{\varepsilon} + F\beta_i] + [b_{ir}^{\varepsilon} + Z\beta_{ir}] + [b_{is}^{\varepsilon} + W\beta_{is}] + [\delta_{i-r-s}^{\varepsilon} + WZ\beta_{i-r-s}]\}) \quad (6)$$

To estimate the model, we drew on the Metropolis-Hastings Robbins-Monro algorithm detailed in (Cai, 2010).

Findings / Results

Although there are several important differences among the models in terms of invariance, fit, and the tenability of assumptions, we simply highlight relative model fit and the relative magnitude of the variance originating from each facet. Our results indicated that there was significant variability across most facets including variability in the item parameters across raters and students. Most notably, measurement noninvariance captured by the item-side variance was a massive 31% (Table 2). More specifically, when an instrument is invariant across facets, the item side variation should be near zero. However, our analyses suggested that 21% of the observed variance was attributable to rater-by-item specific differences and an additional 10% was attributable to student-by-item specific differences. Put differently, students' academic engagement scores heavily depend on who rated them and scores from different raters for different students are not fundamentally on different scales.

Our preliminary analyses indicated the model that best balanced parsimony with fit based on the information criteria was the model that included random effects for items, time points, students, raters, students-by-raters, items-by-raters, and items-by-students (Table 1). Put differently, our analyses suggested that: (1) items differed in their difficulty, (2) students' engagement levels varied across time points, (3) persistent engagement levels varied across students, (4) raters differed in the severity with which they applied the scale (on all items), (5) raters differed in how they applied the scale to different students, (6) individual items functioned differently across raters, and (7) items functioned differently across students.

Conclusions

The comparability of rater-mediated scores is a well-known and complex problem because raters may vary in how they interpret and score observations. In this study, we proposed a new approach to address measurement noninvariance in raters but also across additional facets. Evidence from our initial case study suggests the feasibility and promise of n-level random item effect models to address measurement non-invariance in rater-mediated assessments. However, the potential value of this method needs to be carefully studied to understand the extent to which random item effect models to can effectively address non-invariant conditions.

Appendices

Appendix A. References

Baumgartner, H., & Steenkamp, J. B. E. (2001). Response styles in marketing research: A cross-national investigation. Journal of Marketing Research, 38(2), 143–156.

Bejar, I. I., Williamson, D. M., & Mislevy, R. J. (2006). Human scoring. In D. M. Williamson, R. J. Mislevy & I. I. Bejar (Eds.). Automated scoring of complex tasks in computer-based testing (49-82). Mahwah, NJ: Lawrence Erlbaum Associates.

Cai, L. (2010). Metropolis-Hastings Robbins-Monro Algorithm for Confirmatory Item Factor Analysis. *Journal of Educational and Behavioral Statistics*, 35, 307-335.

Eckes, T. (2009). On common ground? How raters perceive scoring criteria in oral proficiency testing. In A. Brown & K. Hill (Eds.), *Tasks and criteria in performance assessment: Proceedings of the 28th Language Testing Research Colloquium* (pp. 43–73). Frankfurt, Germany: Lang

Hill, H., Charlambous, C., & Kraft, M. (2012). When rater reliability is not enough: teacher observation systems and a case for the generalizability study. Educational Researcher, 41, 2, 56-64.

Kelcey, B., McGinn, D., & Hill, H. (2014). Approximate Measurement Invariance in Cross-classified Rater-mediated Assessments. Frontiers in Quantitative Psychology and Measurement.

Raudenbush, S., & Sadoff, S. (2008). Statistical inference when classroom quality is measured with error. *Journal of Research on Educational Effectiveness*, 1, 138-154.

Appendix B. Tables and Figures

Table 2: Relation between value-added and classroom observation scores

	AIC	BIC
Full model	195,330	195,364
Without three-way interaction (students-raters-items)	193,340	193,386
With only main facets and rater-item interaction	206,404	206,466
With only main facets (items, students, raters)	211,388	211,434

Table 1: Variance in item parameters across raters (total variance=1)

Facet	Variance
Students	0.22
Items	0.02
Time	0.02
Raters	0.21
Students-by-raters	0.22
Items-by-raters	0.21
Students-by-items	0.10