Reweighting Data in the Spirit of Tukey: Using Bayesian Posterior Probabilities as Rasch Residuals for Studying Misfit

Educational and Psychological

Measurement

2016, Vol. 76(1) 88–113

© The Author(s) 2015

Reprints and permissions:

sagepub.com/journalsPermissions.nav

DOI: 10.1177/0013164415583351

epm.sagepub.com

William R. Dardick¹ and Robert J. Mislevy²

Abstract

A new variant of the iterative “data = fit + residual” data-analytical approach described by Mosteller and Tukey is proposed and implemented in the context of item response theory psychometric models. Posterior probabilities from a Bayesian mixture model of a Rasch item response theory model and an unscalable latent class are expressed as weights for the original data. The data are weighted by the units’ posterior probabilities for the unscalable class and used for further exploration of structures. Factor analysis models are compared with the original data and data as reweighted by the posterior probabilities for the unscalable class. In comparing two weighted data sets, Rasch-weighted data and data considered unscalable, differences were evident. Pattern types are detected for the Rasch baseline with patterns that are different patterns from random or systematic contamination. Rasch baseline patterns are strongest near item difficulties closest to the mean generating value of θ s. Patterns in baseline conditions are weaker as they depart from an item difficulty of zero and move toward extreme values. Random contamination patterns are typically flat and near zero regardless of item difficulty. Systematic contamination using reversed Rasch-generated data produced alternate patterns to the Rasch baseline condition and in some conditions showed an opposite effect from Rasch patterns. Differences could be detected within residually weighted data between the Rasch-generated subtest and contaminated subtest. Rasch subtest often had Rasch patterns while contaminated subtest had random/flat or systematic/reversed pattern.

¹The George Washington University, Washington, DC, USA

²Educational Testing Service, Princeton, NJ, USA

Corresponding Author:

William R. Dardick, The George Washington University, 2134 G Street NW#105, Washington, DC 20052, USA.

Email: wdardick@gwu.com

Keywords

Bayesian mixture models, mixture item response theory, Rasch misfit

In educational and psychological assessments, there are expected patterns in the data that form the basis of our hypothesis testing. We draw inference from patterns we can anticipate; however, other patterns exist in the data that may be expected or unexpected depending on the nature of the pattern. These other patterns may arise through systematic or idiosyncratic approaches of the respondents. An examination of model residuals can help elucidate some of these other patterns.

Data = Fit + Residual

In their foundational book, *Data Analysis and Regression*, Mosteller and Tukey (1977) provide a roadmap for the study of data interpretation by expressing data as the degree of fit plus residual and then exploring the pattern of residuals. For most predictive statistical analysis such as a regression model, one can obtain a value that is indicative of the fit to that model, and anything left over, positive or negative, is the residual for that case. In their closing chapter, Mosteller and Tukey posit a set of guidelines for examining regression residuals that can be very helpful in exploring if a better fit exists. This method is found more explicitly in *Understanding Robust and Exploratory Data Analysis* (Hoaglin, Mosteller, & Tukey, 1983). Their central concept is that the model is fit to the data and that the patterns associated with this model can be removed by examining the residuals. In removing the model fit, that portion of the data is stripped away, allowing reexploration of what is left over in the residual portion, and making it easier to examine the remainder of misfit when the portion of the data that is fit to the model is removed. Mosteller and Tukey continue to explore the data in an iterative process, examining and searching for patterns and occasionally observing recognizable patterns that can further explain what is left over.

Similar to the approach marshaled by Mosteller and Tukey (1977), the current project investigates this notion that the fitted portion of the data can be removed in order to explore the portion of the data that does not fit the model. It proposes, however, a different approach to “removing the fitted model from the data.” The approach is illustrated with an example from psychometrics, namely item response theory (IRT).

The Context of the Illustration: Misfit in Item Response Theory

IRT models the probability of a person’s response to an item as a function of one or more parameters for the person’s ability and one or more parameters for characteristics of the item, such as its difficulty and sensitivity to ability. The general form of a one-parameter IRT model also known as the Rasch model:

$$P(x_{ij}|\theta_i, b_j) = \frac{\exp[x_{ij}(\theta_i - b_j)]}{1 + \exp(\theta_i - b_j)}, \quad (1)$$

where x_{ij} is subject i 's response to item j , θ_i is the ability parameter associated with subject i , item parameter b_j associated with item j .

In IRT, the main trait(s) or factor(s) in a model ideally accounts for the responses one would give to items that should measure that ability (Hambleton, Swaminathan, & Rogers, 1991). The probability that an examinee will respond correctly to an item increases as the ability increases, as represented by that factor. There is always some level of departure from the unidimensionality assumption, as psychological and educational research does not take place in a vacuum.

Useful analogs to Rasch's model have previously been presented. Extending work by Goodman (1975), Dayton and Macready (1980) proposed a latent class model with intrinsically unscalable subjects. Andreassen, Woldbye, Falck, and Andersen (1987) used a similar strategy in a medical diagnosis example in which the states of the latent "disease" variable included several possible states of the suspected diseases, including a "normal" state, and an "unknown" state with independent probabilities. Patients with pathologies other than those anticipated would have high posterior probabilities in the unknown class. Yamamoto (1987, 1989, 1995) used a HYBRID model with a discrete latent class model where the two-parameter logistic (2-PL) holds until a subject switches to a guessing strategy, where a constant conditional probability is then used for the remaining random responses. An additional method for examining Rasch mixtures was proposed by Rost (1990). Rost's work with the Rasch model used latent class analysis (LCA) to conceptually split data sets and permit different parameters for a Rasch model within each latent class. The mixed Rasch model gives an alternative to testing the fit of the Rasch model. Using latent class analysis as this conceptual split for Rasch permits the latent classes to form the basis of comparison regarding parameters instead of the more typical use in differential item functioning (Holland & Wainer, 1993) where an available variable (such as score, age, or gender) may not make a meaningful qualitative difference for research into response processes.

The practical advantages of IRT hold when the assumption of local independence is approximately satisfied (Hambleton et al., 1991). Testing the fit of the model is crucial to ensure that violations of the underlying assumptions are not severe. Goodness-of-fit studies (Divgi, 1986; Rogers & Hattie, 1987) were flawed considering their sensitivity to sample size (Hambleton et al., 1991). Hambleton et al. (1991) discuss an approach to assessing data model fit as "(a) designing and conducting a variety of analyses designed to detect expected types of misfit." Useful mechanisms that can produce misfit have been developed within this literature of departure from Rasch data.

The statistical properties, algorithms, and statistical fit of data to the Rasch model have been explored (Kelderman, 1984). Misfit of the Rasch model often involves distributional mixing or multidimensional disturbance with some similar pattern

types for departure from the model (Mead, 1976; Smith, 1986, 1991; Wright, 1995). Similar types of patterns for departure from the Rasch model include guessing, startup (fumbling), plodding, content interaction, sloppiness (carelessness), reanalysis, and random responses. Guessing is a very common type of misfit modeled throughout Rasch research (Mislevy & Verhelst, 1990; Wainer & Wright, 1980). Effects of change on parameters, such as time effect or curriculum effect as shifts in θ or difficulty have been conceptualized (Mead, 1976). Changing θ or changing difficulty can be shown to be the same mathematically if done systemically. If done for the same population or subpopulation, a positive one standardized shift in item difficulty is equivalent to a negative one shift in θ .

Gentner and Gentner (1983) describe models of erroneous knowledge that can serve as an inferential framework for a “reversal” effect. People are separated and taught different analogies to model knowledge. For example: Those using the flowing water model performed well on the battery section and poorly on the resistor section, whereas those using the moving crowd model performed better on resistor section and poorly on the battery section. The model of their performance shows a reversal effect. Given the same questions using different analogy, respondents differ depending on the interaction of the analogy and the model. The relative difficulty of items differs systematically with respect to which analogy respondents are using.

Bayesian Inference

Bayesian inference (Gelman, Carlin, Stern, & Rubin, 2004) is a process that fits a probability model to data sets whose results are summarized by probability distributions on both parameters of the model and unobserved quantities. Inference comes from the data using probability models to quantify uncertainty for observed and unobserved information. Probability statements are made about parameter θ (Gelman et al., 2004) and are conditional on observed values of y : $p(\theta | y)$. The joint probability distribution for θ and y allows for probability statements about θ given y . Density functions are the prior distribution $p(\theta)$ and the sampling distribution $p(y | \theta)$:

$$p(\theta, y) = p(\theta)p(y | \theta). \quad (2)$$

Conditioning on the known data y :

$$p(\theta | y) = \frac{p(\theta)p(y | \theta)}{p(y)}. \quad (3)$$

Omitting $p(y)$ yields the unnormalized posterior density:

$$p(\theta | y) \propto p(\theta)p(y | \theta). \quad (4)$$

In contrast to likelihood functions or estimation equations, Bayesian estimation using Markov-Chain Monte-Carlo (MCMC) estimation iterates through many draws in model parameter space (Rost 1990). Priors in conjunction with the algorithm are

used for each parameter to estimate the posterior. MCMC permits extensions of more complex IRT models, as the number and estimation of parameters is not limited to the more conventional likelihood function.

The goal of the current investigation is to introduce a technique to examine residuals from a Bayesian Rasch IRT model for patterns of responses that do not fit the model, due to utilization of different underlying strategies used by respondents. Notably, “the residual” is a reweighted facsimile of the original data set: the same response vectors, but with cases weighted in proportion to the degree to which they do not fit the posited model. We thus use the Rasch model as a filter and obtain residuals to reweight the data. The residually reweighted data are examined for what is left behind once a model which is unsatisfactory for the complete data set (Mislevy & Verhelst, 1990) has been used.

Method

The current study examined the use of posterior residuals from a Bayesian mixture model composed of a Rasch and an unscalable latent class expressed as weights for the original data to explore factors that may still exist. Such factors include the use of multiple strategies to answer survey questions. The posterior residuals are used to reweigh data to determine if these weights can distinguish significant differences in factor patterns between a Rasch and the unscalable class. In the current investigation, data were generated in accordance with an IRT model. The baseline condition data will be generated strictly in accordance with the Rasch model, whereas in other conditions data will be contaminated in some way, as suggested by earlier studies of person fit analysis. Data are generated into two subtests: a Rasch generated subtest and contaminated subtest. Posterior probabilities from a Bayesian mixture model of a Rasch and an unscalable latent class are expressed as weights for the original data. The data weighted by the unscalable class is used for exploration of further structure, in a new variant of the iterative “data = fit + residual” data-analytical approach described by Mosteller and Tukey. Exploratory factor analysis (EFA) models using tetrachoric correlations are first used to determine if the contamination can be detected in the original generated data. Factor structures are evaluated to determine if, on average, systematic differences are still manifest after the data have been weighted by the unscalable class or “residually reweighted” data set.

The Model

Simulated data were generated to provide fit and intentional misfit (i.e., contamination) to a Rasch model. Using a Bayesian procedure via the Winbugs computer program (Spiegelhalter, Thomas, Best, & Lunn, 2003), a mixture Rasch model was fit to each data set. Each simulated subject had a posterior probability that the case accords with a class defined by the Rasch model response process and a class defined by a residual represented by the independent product of 0.5 probabilities—in essence, a

class of unscalables (Dayton & Macready, 1980). An IRT mixture model was fit to all data conditions, in which one class consists of examinees responding in accordance with the Rasch model, and the other class represents examinees responding randomly. The following expressions are adapted from Mislevy and Verhelst (1990):

The two-class mixture model:

$$P(x_i|\xi) = \sum_{k=1}^2 P(x_i|\phi_k = 1, \xi) \pi_k, \quad (5)$$

where π is the class proportion, ϕ indicates subject parameter to indicate k strategy usage, and ξ is the full vector of item parameters (defined below). The mixture model for estimation will contain two classes. The first class defined by the Rasch model for Equation (1) and the second by Random guessing.

Rasch model:

$$P(x|\phi_1 = 1, b_1, \dots, b_J) = \int \prod_j \frac{\exp[x_j(\theta - b_j)]}{1 + \exp(\theta - b_j)} p(\theta) d\theta, \quad (6)$$

where b_j is the difficulty parameter for item j under the Rasch model.

Random guessing:

$$P(x|\phi_2 = 1, c_1, \dots, c_J) = \prod_j c_j^{x_j} (1 - c_j)^{1-x_j}, \quad (7)$$

where c_j is a prespecified guessing parameter for item j .

It follows from the definition of the model that

$$\begin{aligned} P(x_i|\xi) &= P(x|\phi_1 = 1, b_1, \dots, b_J) \pi_1 + P(x|\phi_2 = 1, c_1, \dots, c_J) \pi_2 \\ &= \left[\int \prod_j \frac{\exp[x_j(\theta - b_j)]}{1 + \exp(\theta - b_j)} p(\theta) d\theta \right] \pi_1 + \left[\prod_j c_j^{x_j} (1 - c_j)^{1-x_j} \right] \pi_2. \end{aligned} \quad (8)$$

The posterior probability of a given subject with response vector x_i belonging to each of the two classes is thus calculated as follows:

Rasch class:

$$= \frac{\left[\int \prod_j \frac{\exp[x_j(\theta - b_j)]}{1 + \exp(\theta - b_j)} p(\theta) d\theta \right] \pi_1}{\left[\int \prod_j \frac{\exp[x_j(\theta - b_j)]}{1 + \exp(\theta - b_j)} p(\theta) d\theta \right] \pi_1 + \left[\prod_j c_j^{x_j} (1 - c_j)^{1-x_j} \right] \pi_2}. \quad (9)$$

Random class:

$$= \frac{\left[\prod_j c_j^{x_j} (1 - c_j)^{1-x_j} \right] \pi_2}{\left[\int \prod_j \frac{\exp[x_j(\theta - b_j)]}{1 + \exp(\theta - b_j)} p(\theta) d\theta \right] \pi_1 + \left[\prod_j c_j^{x_j} (1 - c_j)^{1-x_j} \right] \pi_2}. \quad (10)$$

The posterior probability estimate can be used to determine in which class each response pattern likely belongs, Rasch or non-Rasch, and to what degree each response pattern likely belongs. In this model, we can interpret an examinee's posterior probability of belonging to the Rasch class as his fit to the Rasch model, and the posterior probability of belonging to the random guessing class as a residual. Each subject's posterior probability of being in the unscalable class were used to reweight the original data as the "residual" data set. Responses that fit perfectly to the Rasch class will have a weight near one, whereas those patterns that accord very poorly with the Rasch model will have a weight nearer to zero. All other response patterns will have a value between zero and one from the posterior distribution representing the degree of fit with the Rasch model. Those that fit well with the Rasch model should be those respondents that were generated to conform to the Rasch model. Those that differ are less like the theoretical model, and exploring this residual information can yield reasons for differences.

Using the residual weights from the posterior distribution of the Bayesian Rasch model, the original data were reexamined using factor analysis. Tetrachoric correlations were used instead of Pearson's correlation coefficient to accommodate dichotomous data. Three EFA models were constructed:

1. *Unweighted*: The first exploratory model examined the unweighted data set. It was considered useful to return to this unweighted model after examining the following weighted models.
2. *Rewighted with posterior probabilities from the unscalable class*: The second EFA model is the model of interest for examining response patterns to the extent that they are not in accord with the expected one-dimensional model. This model used reweighted data, where the weights represented posterior probabilities of the unscalable class, that is, a weighting of the data that best represents misfit to the Rasch model.
3. *Rewighted and estimated to represent fit to the Rasch model*: This model uses reweighted data, where the weights for each case are posterior probabilities of belonging to the Rasch class, that is, a weighting of the data that best represents fit to the Rasch model.

Rasch and Misfit Data

Data were generated in the current investigation to conform to three different strategies: Rasch, random, and a Rasch reversal effect. The Rasch model represented the

expected response pattern. Randomly generated responses represented the first patterns of misfit from the Rasch. These two strategies are commonly found in mixture IRT simulations (e.g., Smith, 1986, 1988, and Mislevy & Verhelst, 1990). The random effect misfit was characterized by random chance regardless of the difficulty of the item; the reversal effect represents special training, course of study, or analogy difference. The reversal alternative strategy is based on the reversal effect shown by Gentner and Gentner (1983) mentioned previously. The key idea here is that one kind of task is relatively easier for students approaching tasks one way, but another kind of task is relatively easier for students approaching tasks in a different way. Based on analogy or special knowledge, the reversal effect was simulated by reversing the difficulties of the items in one of the subtests for those in the contaminated condition.

The first strategy was represented by the Rasch model:

$$P(X_i = 1|\theta) = \frac{e^{\theta - b_i}}{1 + e^{\theta - b_i}}. \quad (11)$$

The second strategy will be represented by generating random responses. This “Random” strategy is common and sometimes based on issues such as the lack of care from a participant, rushing through responses, lack of knowledge, or some other underlying concept that would manifest an apparently random set of responses:

$$P(X_i = 1|\theta) = .25. \quad (12)$$

The probability, P , in the current investigation is set to .25 because the data were generated assuming that four responses were possible for each item, as is typical for most multiple choice exams. If a participant responds randomly to an item with C possible response categories, the resulting probability of guessing is

$$P(X_i = 1|\theta) = \frac{1}{C}. \quad (13)$$

Selecting 0.25 as the unscalable class is analogous to research by Andreassen et al. (1987) for diseases that are unknown. When looking at diseases in a Bayesian system Andreassen et al. (1987) introduce a state of “other.” This “other” condition helps avoid strong favorable statements for any of the set of prespecified disease states, when cases fit poorly. Instead of placing the case in the least poor fitting, prespecified condition, the network places this case in the “other” condition with a high probability. This is similar to the unscalable class in the current investigation. It is not that such a response pattern fits the unscalable condition but that it does not fit the posited model structure. This can be a useful indicator that anomalous structure not accounted for by the hypothesized model(s) is present.

The third strategy we present here is represented by generating responses affected by special knowledge or analogy. The reversal effect, similar to what Gentner and Gentner (1983) found, was simulated by reversing the difficulties on the generating parameters.

Design of the Simulation Study

Fixed Parameters. Theta for the Rasch portions of the mixture had a normal distribution with a mean of zero and a constant standard deviation of one. The sample size was fixed to 500 simulees. The number of replications per cell was set to 50. The number of items was held constant at 40 with evidence supporting this number of items or less from previous research (e.g., Smith 1988, and Wright & Tennant, 1996).

The researchers' preference to select precise points instead of having them randomly generated for the test permitted the evaluation of the study to have multiple overlapping item difficulties so that comparisons within and between subtests of the test could be made clear. To evaluate common structure, patterns were examined individually and aggregated by these common difficulties within subtests.

Manipulated Conditions. There were a total of four manipulated conditions in the current investigation: type of misfit or contamination, proportion of misfit or contamination, scaling factor of the test, and size of subtest. There were two sets of cells for the no contamination condition: no random contamination and no reverse contamination are the same effective cell and only replicated once. The number of cells replicated was the multiplication product of the number of levels for each manipulated condition, $2 \times 4 \times 4 \times 2 + 8 = 72$ cells in total.

The first of the manipulated parameters, type of misfit to the Rasch model, involves departure from Rasch generating data. The data were generated by altering the proportion of strategies being mixed and the strength of the models. The Rasch versus non-Rasch (random, reverse effect) condition represented key difference in cells. These non-Rasch conditions were considered contamination in the investigation.

The second manipulated parameter, the mixture proportions of expected and alternative models, will have five levels. The amount of non-Rasch, contaminated responses mixed in with the expected Rasch responses were 0, 5%, 20%, 50%, and 100%. The mixing proportion here is for the total data set, not within an individual's response pattern. This condition for a given set of responses was considered Rasch or contaminated between case proportions. Based on this first condition, responses were all Rasch or contaminated within a given response pattern.

The third manipulated parameter is the scaling factor of the test difficulties. In determining the range of the difficulties, several articles were reviewed (Chen & Davison, 1996; Chyn, Tang, & Way, 1994; Mislevy & Verhelst, 1990; Kingston & Dorans, 1982). In the sensitivity analysis (Hambleton & Rovinelli, 1973; as cited in Hambleton et al., 1991) carried out on the goodness-of-fit statistics to determine sample size, item difficulties were chosen to those commonly found in practice, between -2 and $+2$. Items in the current investigation, taking into account previous research and practical considerations, were constrained to have a base range of 2 through -2 . This item range interacted with a change in scaling factor specifically considered to interact with the contaminated conditions.

The scaling factors of the item difficulties in the test took on four conditions: The base range of the exam extends from ± 2 . This scaling factor was used to create the

four conditions where the base range of items is multiplied by (1) 1 for both subtests, (2) 3 for both subtests, (3) 1 for the Rasch-only subtest and by 3 for the contaminated subtest, and (4) 3 for the Rasch-only subtest and by 1 for the contaminated subtest. This provided a variety of test ranges from ± 2 for all items in both subtests to a relatively extreme condition with a range of ± 6 for items on subtests with interaction effect of those ranges also being examined.

The fourth and final manipulated parameter was the sizes of subtests. The simulated test had 40 items total and is divided into two subtests, a Rasch subtest which is generated as Rasch data, and a contaminated subtest generated with Rasch, random, or reversed-Rasch data depending on the generating parameter for type of mixing. In the first condition, the Rasch-only subtest had 30 items generated Rasch only, with the remaining 10 items in the contaminated subtest changing based on the mixing parameters. The second subtest condition for this parameter was equal subtests of 20 items each.

The random proportion was envisioned as the lack of care, laziness, or even an alternate rushing factor. In conforming to this notion, randomness was adjusted within a response pattern to represent different levels of carelessness. The reverse contamination was generated to conform to the analogy or alternative knowledge models. These two represent contamination in the second subtest. The contaminated responses were generated within the second subtest. This subtest represents not caring responses or differencing inferential models within the non-Rasch response subtest. All individuals conformed to the Rasch model for the first subtest. This manipulation was the amount of randomness within each response pattern.

Sample size for the current study was evaluated in preliminary investigations, and the results indicated that 500 people estimated parameters significantly better than 100 people but not significantly different from 2,000 people per replication. Overall there were 72 mixture response patterns generated with 500 subjects per replication.

Simulation Method and Preliminary Investigation

The model for consideration was constructed and run in Winbugs using MCMC estimation (Spiegelhalter et al., 2003). The measurement model was developed as a mixture Rasch model which fits a latent class model with the first class being the Rasch model and the second class being unscalable. Estimation of residual was performed using a Bayesian mixture Rasch model within each of the 72 cells. In preliminary investigations using Winbugs, the priors were set for each parameter and tested to make sure the priors did not overwhelm the data. In the current analysis, a normal distribution (e.g., Mislevy, 1986) with a weak prior (Rupp, 2003) is used for of item difficulties. Priors for item difficulty and theta were examined in preliminary analysis and were set here to $N(0, 4)$ for item difficulty and $N(0, \tau)$ for theta, where the prior distribution for tau was $\text{Gamma}(0.5, 1)$.

Data points were fixed to resolve the label switching indeterminacy with multiple classes (Chung, Loken, & Schafer, 2004). The data were fixed by the first replication

always being set to the first class. The model was set to remain as a fixed condition in the overall simulation. The preliminary study also set the number of burn-in and estimation cycles for the study. The two-class mixture model of Rasch and an unscalable class is used in all replications across all cells.

The burn-in cycles were set to 2,000 and the estimation cycles were set to 5,000. In the preliminary investigation, we follow recommendations by Gelman et al. (2004) to empirically evaluate burn-in and estimation cycles' burn-in, sample size, priors, and replications while debugging the model. Preliminary investigations evaluated burn-in cycles (iterations set to 50, 250, 500, 1,000, 2,000, 4,000, and 8,000). Burn-in cycles were evaluated using Gelman–Rubin (2004) statistics, with examination of density patterns and convergence of three chains showing very little if any improvement after 450 burn-in cycles. Estimation cycles' sizes of 2,500, 5,000, and 7,500 were tested with posterior means compared with generating parameters. Sample sizes of 100, 500, and 2,000 were investigated in preliminary studies and average difference recovered from generating parameters for item difficulties by sample was 100, $M = -0.23$ (standard error $[SE] = 0.08$); 500, $M = -0.08$ ($SE = 0.03$); 2,000, $M = -0.07$ ($SE = 0.02$).

Time was also a consideration in selecting sample size. By selecting sample size of 500 over 2,000, not only was the difference in recovery small (.01) but this was balanced with an average saving of 37 minutes per replication, which accounted for more than 3 months of simulation time. We chose to use 50 replications per cell, after considering time of convergence per replication and average percent classified into the contamination condition was shown to have an effect significantly different from a Rasch-only condition. A detailed account of the preliminary analysis and results, with graphics, tables, and outcomes for Gelman–Rubin statistics, examination of density patterns chain convergence, and results can be found in Dardick (2010).

Probabilities were generated in SAS and based on Rasch and misfitting, contaminated conditions using Equation (11) for Rasch, Equation (12) for random misfit and reversing the difficulties using Equation (11) for the reversal pattern.

Exploration of the Residuals in the Spirit of Tukey

In a similar context of data = fit + residual (Mosteller & Tukey, 1977), the data were explored using reweighted data. Summary statistics were calculated for 50 replications per cell for the posterior mixture model. After cells were simulated, three EFAs were conducted on samples of sufficient size (Gorsuch, 1983; Kline, 2013; MacCallum, Widaman, Zhang, & Hong, 1999), with weighted value of 500 simulees for each replication from each individual data set based on the posterior mixture model. The first factor model used the unweighted data, the second used Rasch-weighted data, and the third used residual-weighted data. The number of factors in a condition was considered and the structure of factors was evaluated visually for patterns after aggregating over the 50 replications. Loading patterns were aggregated across the 50 replications and compared through visual inspection and appear in the

“Results” section. There were two types of graphs constructed from the factor patterns. The first type of graph aggregates the loading patterns from the 50 replications within a cell and sums across all items within each subtest with the same difficulty. This graph compares the Rasch subtest with the contaminated subtest visually for aggregated factor loadings to compare patterns within subtests by item difficulty. In the current investigation, a random generation of 2% of the data was used for the Rasch-only condition to provide a baseline pattern for average factor loadings and the other conditions used the residually weighted data to explore the pattern of residuals in an attempt to distinguish patterns of factor loadings other than Rasch patterns. The second type of graph compares two sets of aggregated factor loadings across all 40 items: Rasch weighted factor loadings are visually compared with residually weighted factor loadings. The item loadings were aggregated across the 50 replications and the two sets of patterns were inspected.

To determine the number of factors for the unweighted factor models, 50 replications were used in Horns’s parallel analysis to stay consistent with the 50 replications in each cell. Visual inspection of factors was conducted visually to support the results of the parallel analysis. Each replication was run on random data with the same structure as the data used in the current analysis: 40 items, 500 simulees in the unweighted model, and the correct proportion per cell when data were weighted.

The initial examination of the conditions determined what proportion of times respondents who were generated to have misfit data were assigned to the residual class, answering the question, how well does the mixture model categorize these respondents into the unscalable class? Examinations of the residuals were compared with the baseline Rasch model. In the current investigation smaller residuals were explored, provided they worked with the factor analysis procedure and were reasonably large compared with the baseline Rasch models. Residuals, as they get extraordinarily small, start to account for less weight than two simulated people, 0.4% in the current investigation, making them difficult to interpret and often weighting on just one case. The smallest condition explored was 0.4%, because it was feasible for analysis. Preliminary investigations demonstrated that these values were outside 2 *SEs* from the Rasch preliminary investigation; values less than 0.4 represented so few cases that the investigation would be a case-level investigation which is not the intention of the current study. In the results, we evaluate size of the residuals and explore patterns in magnitude of detected residuals. We then explore patterns in the residuals through interpretation of graphical relationships of loadings of items on factors to help demonstrate patterns of Rasch and misfit when captured by a residual. To better understand what is left over in the unscalable class, the graphical comparison focuses on the residual-weighted data and compares the Rasch subtest with the contaminated subtest.

Results

The class proportions are presented in Table 1 as percentage classified into the unscalable condition. The largest of these unscalable class percentages for the Rasch-

Table 1. Average Percent Classified Into Unscalable Class.

Percentage contaminated	Scaling factor	Subtest size	Random		Reverse	
			<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>
100	1 × 1	10	2.51	0.88	0.12	0.09
		20	6.35	1.71	0.16	0.15
	3 × 3	10	0.01	0.04	0.00	0.00
		20	0.05	0.08	0.00	0.00
	3 × 1	10	0.01	0.05	0.00	0.00
		20	0.05	0.07	0.00	0.00
	1 × 3	10	2.63	0.99	0.03	0.06
		20	6.59	1.52	0.01	0.03
	1 × 1	10	0.64	0.35	0.26	0.12
		20	1.92	0.78	0.53	0.24
50	3 × 3	10	0.01	0.03	0.01	0.04
		20	0.06	0.08	44.09	16.43
	3 × 1	10	0.01	0.02	0.00	0.00
		20	0.03	0.06	0.00	0.01
	1 × 3	10	1.14	0.51	0.56	0.33
		20	18.44	18.02	50.08	0.11
	1 × 1	10	0.30	0.19	0.38	0.22
		20	1.71	0.55	3.91	1.64
	3 × 3	10	0.01	0.02	0.01	0.04
		20	1.47	0.67	20.04	0.00
20	3 × 1	10	0.00	0.03	0.00	0.00
		20	0.01	0.04	0.03	0.04
	1 × 3	10	1.57	0.71	7.88	4.29
		20	19.18	0.51	20.07	0.04
	1 × 1	10	0.17	0.11	0.40	0.30
		20	0.83	0.18	2.50	0.63
	3 × 3	10	0.04	0.07	0.24	0.23
		20	3.47	0.58	5.01	0.00
	3 × 1	10	0.00	0.00	0.00	0.00
		20	0.01	0.04	0.10	0.12
5	1 × 3	10	1.25	0.48	4.65	0.37
		20	4.81	0.17	5.03	0.05
	1 × 1	10	0.14	0.12	—	—
		20	0.15	0.12	—	—
	3 × 3	10	0.00	0.00	—	—
		20	0.00	0.00	—	—
	3 × 1	10	0.00	0.01	—	—
		20	0.01	0.02	—	—
	1 × 3	10	0.02	0.04	—	—
		20	0.00	0.01	—	—
0	3 × 3	10	0.00	0.00	—	—
		20	0.00	0.00	—	—
	3 × 1	10	0.00	0.01	—	—
		20	0.01	0.02	—	—
	1 × 3	10	0.02	0.04	—	—
		20	0.00	0.01	—	—

Note. The dash “—” indicates Rasch-only condition not evaluated because of redundancy.

only conditions ($\bar{X} = 0.1482\%$, $S_x = 0.1203$) were considered as a baseline for the other 64 cells. Table 1 shows the average unscalable percentage classified into the unscalable class for random and reverse contamination conditions across all levels of

contamination, subtest size, and subtest variability. The results examine residual size, factors extracted, and graphical evidence. The most interesting cells in which to explore the residual data are those with a residual size of 0.4% or greater and have eigenvalues over the threshold. The 29 largest conditions were explored to determine if we could determine patterns of residual size and visually evaluate the residual. Given the expansive nature of the exploratory work only a subset of conditions are highlighted in the results indicating some of the more important and interesting findings. A more detailed exploration of results using statistical techniques, graphs, and tables for weighted, unweighted, and residually weighted data appear in Dardick (2010).

Residual Size in Simulated Conditions

The size of the contaminated subtest had a large impact on the magnitude of the residual, and in rank order, it was always larger as a residual for true contaminated conditions. When the scaling factors for the two subtests were equal, the size of the contaminated subtest affected the size of the residual. When the subtest holding the scaling factor constant had 10 items for reverse Rasch contamination in the 50%, 20%, and 5% conditions, only one of the six residuals were just large enough to meaningfully explore. In contrast, all 6 of the 20-item subtest reversed Rasch contamination condition holding the scaling factor constant were above the 0.4% threshold set for examination. The same pattern was evident for the 12 values in the random condition; one of the six 10-item conditions was just large enough to explore whereas five of the six 20-item conditions could be explored. In conditions with the smaller scaling factor of Rasch items and contamination, the residual was large enough to explore except in some reverse conditions. When the scaling factor switched, the contaminated subtest was too small to examine.

Relative to the size of the generated contamination, the success of the model was inversely proportional to the size of the residual. In the 100% condition, the largest residual was 6.59%; in the 50% contamination condition, three residuals stood out: 50.08%, 44.09%, and 18.44%. In the 20% contamination, three residuals were approximately 20% and one was 7.88%. In the 5% contamination condition, four residuals were approximately 5%, one was 3.47%, one was 2.50%, and one was 1.25%. Proportionally, as the size of the residual decreased, the proportion classified increased.

The size of the residual increased on average from the random condition $\bar{X} = 2.35$, $S_x = 4.67$ to the reversed $\bar{X} = 5.19$, $S_x = 12.12$ condition. As percent contamination condition decreased, the residual increased 100%, $\bar{X} = 0.01$, $S_x = 0.02$; 50%, $\bar{X} = 0.15$, $S_x = 0.32$, 20%; $\bar{X} = 0.24$, $S_x = 0.39$; 5% $\bar{X} = 0.36$, $S_x = 0.42$. The size of the subtest increases from $\bar{X} = 0.77\%$, $S_x = 1.55$ in the 10-item condition to $\bar{X} = 6.77$, $S_x = 11.14$ in the 20-item subtest condition. An increase in the scaling factor of the subtest increased the size from $\bar{X} = 0.72$, $S_x = 1.40$ when the range was -2 to $+2$ to $\bar{X} = 6.83$, $S_x = 12.31$ for the -6 to $+6$ condition. In general, the residual size was largest

proportionately in the 20-item subtest with a larger scaling factor. The one exception was the reversed-Rasch 100% contamination condition, which is duplicative of the Rasch baseline condition.

Residual Patterns for the Rasch Baseline Condition

There is predominantly one main factor in the baseline and the 100% contamination conditions with some minor secondary factors in the Rasch and reversed-Rasch condition. The reversed-Rasch scaled the difficulties backward for all simulees in the conditions, effectively mirroring the all-Rasch condition. The 100% reversed-Rasch is indistinguishable from the Rasch baseline. In the unweighted data for the reversed-Rasch contamination conditions for 50%, 20%, and 5%, and for the random contamination of 50% and 20%, two factors were present. In the 5% random contamination condition, only a second factor existed for the conditions with the increased subtest scaling factor.

Graphs in Figure 1 show residual patterns aggregated by difficulties within subtest based on patterns estimated after simulation as explained in the “Method” section. This type of figure is also used as part of a panel in Figures 2 through 5. The visual representation shows patterns on the Rasch subtest compared with the contaminated subtest. In Figure 1, the differences in patterns are because of the change in scaling factor for variance. The Rasch pattern is the same across all conditions from the same scaling factors. The scaling factor of 1 has a relatively flat set of patterns around 0.3, whereas the scaling factor of 3 has a mountain or wave-like pattern with the extreme values going to 0 and the items around 0 have patterns around 0.3. These two stable patterns can be used as a baseline to explore other conditions for departure from what might be expected.

Highlighted Residual Patterns

The focus of this section is to investigate some of the more interesting residuals with something left over when the Rasch partitioning of the data is removed. The random contamination residuals met the criteria for detailed exploration in three conditions where the test range was ± 2 for the uncontaminated subtest and ± 6 for the contaminated subtest, when the subtest of size was 20 for the 50%, 20%, and 5% contamination for the first factor. Figure 2 highlights this importance with two examples, in the 50% and 5% conditions. In the random contamination condition, the interaction between a subtest of 20 items and an increase in scaling factor of the contaminated subtest was associated to the residual containing a factor large enough to explore. The contaminated subtest in Figure 2 had smaller patterns close to zero whereas the Rasch subtest has a pattern similar to those observed in the Rasch baseline condition. In the bottom two panels of Figure 2, we examine the difference between Rasch and residually weighted data where the first 20 items are the Rasch subtest and the second 20 items contain some percent of contamination. The differences in the 50%

Figure 1. Rasch data panels. Three panels of aggregated patterns for the Rasch-only condition each with a 20-item subtest and scaling factors from top left to bottom right: 1×1 , 3×3 , 1×3 . The data are residual factor patterns aggregated by difficulties within each subtest. The two subtests are labeled Rasch subtest for item difficulties in the Rasch subtest and misfit subtest for item difficulties in the potentially contaminated subtest. The values from both subtests are shown on the same scale from -2 to $+2$. The visual representation shows patterns on the Rasch subtest compared with the contaminated subtest. These three examples highlight all-Rasch conditions. Patterns are generated using a rand 2% condition. Conditions with greater percentages conform to the same pattern.

contamination between the Rasch and residually weighted data are noticeable in both subtests whereas in the 5% contamination conditions the difference is more extreme in the contaminated portion of the data. The patterns in the 5% contaminated panels, residually weighted data show that same Rasch pattern for the Rasch generated subtest, Items 1 through 20, and show patterns hovering around zero in the contamination portion of the subtest. In the residual data this is evidently a non-Rasch pattern for the contaminated subtest and a Rasch pattern for the first 20 items. This is the random pattern expected in the residual data when a pattern would exist.

Systematically contaminated residuals were interesting to explore in the following conditions: The six conditions for 50%, 20%, and 5% where the 20-item contaminated subtests had a range of ± 6 , the two conditions for the 20% and 5% where the 10-item contaminated subtests range was ± 2 for the uncontaminated subtest and ± 6 for the contaminated subtest, the one condition for the 5% condition where the 20-item contaminated subtests had a range of ± 2 for the uncontaminated subtest and ± 6 for the contaminated subtest.

Figure 2. Panels for random 50% and 5% (left to right) contaminated data. Rasch and contaminated 20-item subtests with associated Rasch and residually weighted data with a scaling factor of 1×3 . These two conditions represent some of the more interesting and important findings. In the top two panels we observe separation between the Rasch and residual subtest with the residual patterns close to zero. The bottom two panels express the difference between Rasch and residually weighted data where the first 20 items are the Rasch subtest and the second 20 items contain contamination, 50% and 5%, respectively.

Figures 3 through 5 provide highlights for the 50%, 20%, and 5%, systematically reversed-Rasch contaminated conditions. Figure 3 highlights residual patterns for the 50% reversed contamination for Factors 1 and 2, respectively. Figure 4 reveals an interesting finding in residual patterns for the 20% reversed contamination in Factors 1 and 2, respectively. Figure 5 represents some of the more important residual patterns for the 5% reversed contamination by Factors 1 and 2, respectively.

We see slight departure from Rasch patterns in Figure 3. The residuals are Rasch-like and might be starting to reveal the reversed-Rasch characteristics in the 1×3 condition for these figures. It is not clear in this 50% contamination condition but will be more evident in Figures 4 and 5. The two common themes in patterns highlighted for the 20% (Figure 4) and 5% (Figure 5) conditions for the first factor were crossing pattern and Rasch-like patterns. This pattern shows what we will call a reversal effect and a distinct pattern in most cases from the Rasch model. We consider the majority of these first factors a reversed-Rasch factor. In the comparison model, the crossing pattern looks a lot like steps or mirrored patterns. In the 1×3 20% conditions in Figure 4, the first factor is the crossing pattern of reversed Rasch and indicative of the systematic contamination whereas a Rasch-like pattern is more prevalent in the second factors. Some crossing is apparent in all the 5% conditions, but only one of

Figure 3. Fifty percent factor pattern example for reversed contamination condition with a scaling factor of 1×3 . The two panels on the left represent the first and second factor patterns for the highlighted 50% reversed contamination condition. The right two panels express the difference between Rasch and residually weighted data where the first 20 items are the Rasch subtest and the second 20 items contain contamination, for Factors 1 and 2, respectively.

Figure 4. Twenty percent factor pattern example for reversed contamination condition. The two panels on the left represent the first and second factor patterns for the highlighted 20% reversed contamination condition. The right two panels express the difference between Rasch and residually weighted data where the first 30 items are the Rasch subtest and the second 10 items contain contamination, for Factors 1 and 2, respectively.

Figure 5. Five percent factor pattern example for reversed contamination condition. The two panels on the left represent the first and second factor patterns for the highlighted 5% reversed contamination condition. The right two panels express the difference between Rasch and residually weighted data where the first 20 items are the Rasch subtest and the second 20 items contain contamination, for Factors 1 and 2, respectively.

them has a strong enough second factor to highlight. When a pattern was detected as the second factor in the reversed-Rasch contaminated conditions, it was very similar to the Rasch baseline patterns.

Discussion

The goal of the current investigation was to examine residuals from a Bayesian Rasch IRT model for patterns of responses that do not fit the model, due to utilization of different underlying strategies used by respondents. Notably, “the residual” was a reweighted facsimile of the original data set: the same response vectors, but with cases weighted in proportion to the degree to which they do not fit the posited model. The investigation used the Rasch model as a filter and obtained residuals to reweight the data. The residually reweighted data were then examined for what is left behind once a model, which is unsatisfactory for the complete data set in the sense of Mislevy and Verhelst (1990), has been used.

Residual Size

Separating out some information from the Rasch class into an unscalable class was successful under certain conditions. The residuals were large enough to explore in 29

of those conditions. "Large enough to explore" was operationally defined by a threshold of 0.4% or the equivalent of at least two simulees. In the 100% contamination condition, the random contamination produced a searchable residual four out of eight times. The reversed contamination condition did not produce anything to explore. The reversed-Rasch scaled the difficulties backward for all simulees in the conditions, effectively mirroring the all-Rasch condition. The 100% reversed subtest was Rasch with items in the opposite order for the subtest and should be treated as such. The residuals that are left looked identical in both magnitude and relative pattern as the Rasch baseline condition. Seventy-two conditions were explored with 8 Rasch and 8 were of the 100% reversed-Rasch subtest, leaving 56 conditions to explore further, 29 of which had large residuals for this study.

The model selected was unable to acquire a residual when the Rasch subtest had the scaling factor of 3 and/or the contaminated condition had a scaling factor of 1. The Rasch subtest with a wide range of ± 6 seemed to saturate the model. It is likely that the larger scaling factor overwhelmed the model with the broad range in the Rasch subtest. The contaminated subtest with a range of ± 2 was relatively small compared with the Rasch subtest with a range of ± 6 . This accounts for 12 cells not overlapping with the two previously mentioned Rasch conditions.

The opposite overall effect was true for the conditions where the Rasch subtest had a scaling factor of 1 and the contamination condition had a scaling factor of 3. In all cases, except for the Rasch-only and the 100% reversed-Rasch condition, a residual was considered large enough to reevaluate the weighted data. The Rasch data in these conditions for the Rasch subtest were restricted to ± 2 range. Here the truncated Rasch was in the Rasch-only subtest and the contaminated subtest had potential for large discrepancy in its mixing subtest. The ± 6 and ± 3 item difficulties had a large difference in the contaminated condition from the unscalable value. In the Rasch and reversed-Rasch conditions some very extreme discrepancies occurred when ± 6 and ± 3 difficulties were reversed for the same set of items.

Factor Patterns

When the data were weighted by the residual class, the reversed condition showed clearer factors than the random condition. In the exploration of residuals with factors, the reversed-Rasch contamination often showed up in the residual and a secondary factor was also present in several conditions as a Rasch pattern. In the random contamination, most of the time the patterns were scattered around zero.

Differences in the patterns for random and systematic contamination subtest differed from the all-Rasch subtests. The Rasch patterns had a shape that was flat with a slight rounding pattern around 0.3 on average for the subtest with a range from ± 2 or an upside down "V" shape peaking around 0.3 or 0.4 for the ± 6 subtests. The pattern for the ± 2 condition is a less extreme version of the ± 6 condition. On the first factor for the conditions with a range of item difficulties of ± 6 : the extremely ± 6 item difficulties have pattern values around 0, the moderate ± 3 difficulties had

patterns around 0.3 and the 0 item difficulties had values around 0.5. As the items approached zero the patterns in the factor were stronger. This was seen in a less extreme condition in the ± 2 condition. The ± 2 item difficulties have values close to 0.2 patterns whereas the difficulties around 0 reach a maximum around 0.4. This means that items which were closer to the average ability were the strongest patterns. As items diverged from the average simulees' θ value of 0 to more extreme values, the patterns converge toward 0. The patterns seem to reach a maximum value where maximum information occurs.

The current underlying theory for Rasch-only factor patterns is that maximum patterns are displayed at maximum information and the minimum patterns are displayed at minimum information. The secondary factor pattern is just a weak copy of the first pattern.

In the residual of the random contamination condition when a factor is determined to exist, the pattern that occurs is a separation between the Rasch and random patterns. The random subtest has patterns around zero whereas the Rasch subtest has patterns that match the Rasch baseline patterns, with the largest patterns closest to the mean generating value of zero being largest and those with more extreme difficulties being weaker. This instance shows that random contamination may still be present and a residual could possibly be extracted from the data even when the unweighted model looks like a Rasch pattern.

In the reverse conditions, the most interesting patterns had a crossing or reversal pattern. The reversal pattern, when it is most present in the residual data, yields a pattern different from the Rasch maximum information pattern. The reversed pattern had large positive average factor patterns for large positive item difficulties, small patterns for item difficulties near zero, and large negative pattern values for large negative item difficulties. The unweighted data are distinct from both the residual and Rasch-weighted data. When the contaminated subtest is explored in the residual-weighted data, a reversed or crossing pattern exists.

Implications

When the data were truly Rasch under the eight tested conditions and the eight conditions fully-Rasch with reversed patterns, the Rasch model was classified correctly. One could use the model as part of a battery of examinations to determine if anything out of the ordinary can be detected with an unusual size residual. The amount detected did not indicate the amount contaminated but for future research in theoretical and applied work could act as a flag for detection types of competing contamination to the Rasch model. Although a small residual would not be enough indication that there was not contamination of the type described in this investigation, a positive find would certainly warrant further investigation of residual patterns.

In cases in which all respondents are equally misinformed or contaminated, the model may not pick up or detect anything unusual, as they may be similar to the reversed-Rasch condition or random condition. Where some students are functioning Rasch and others are functioning in some alternative method, this may be detected

by a model as simple as adding an unscalable class to use a residual information. The residual can then be explored.

Just because a residual was detected did not mean it had pattern distinct from the Rasch data. Sometimes the residual was extracted and it looked just like the Rasch-weighted data. Other times it looked like random scatter around zero. This technique could serve as a tool to determine first if there is a residual class of sufficient size to be considered as a flag for departure from the expected model, in this case the Rasch model. Next, the residual-weighted data can be compared with the Rasch-weighted data to determine if differences exist. If differences exist between the two patterns, the residually weighted data can be explored in more detail to determine if the factor patterns match some form of known contamination, such as the reversed-Rasch or random contamination, or if the contamination is something else yet to be identified.

Novelty of Study

This investigation differed from the Mosteller and Tukey approach in its notion of misfit. The Rasch model may be fit to the data but a known value to generate a residual for an individual's ability does not exist. In the current article, we examined the model through a mixture perspective using the structure of a latent class model. An unscalable class is used as a type of residual catchall. Mosteller and Tukey (1977) provided a road map for exploring residuals, particularly those in the regression context. The model will fit the data to some degree regardless of the underlying distributions. The fit of the model can be extracted and what is left behind can be explored. These ideas run parallel to this study. Data are transformed and reevaluated to explore for additional anomalous and probable patterns. Mosteller and Tukey explore the residual to determine what type of transformation of the original data will simplify and clarify the analysis. The differences arise in this investigation when data are reweighted by the posterior probability of class membership—specifically by the probability of a residual class rather than a class defined by the “model” class corresponding to the Rasch model—and then transformed and explored. Each respondent receives a weight from 0 to 1 and the new data are a weighted copy of the original data set. The same data vectors exist but each individual case represents only a proportion of the original case. In the Mosteller and Tukey method, data are transformed and a new data set is manifest with each case still representing the same proportion as in the original data, while each data point is the difference between the original data point and the value predicted by the model. Their transformation changes the data values while maintaining case proportionality. This investigation changes case proportionality and maintains the original data values.

Future Research

The general technique to examine residuals from a Bayesian Rasch model can be extended to other IRT conditions, logistic models, and classes beyond the two

presented here. The focus of this investigation was introduction to the technique by using an unscalable class to investigate divergence from an expected model, in this case the one-parameter logistic (1-PL) model. The approach is by no means limited to the Rasch model. Additional models could be checked iteratively using the unscalable class mixture after the first model demonstrated contamination (e.g., 2-PL, 3-PL). It would be interesting to use various IRT models together. Models could incorporate different logistic parameter models, such as 2-PL and 3-PL models, or covariates, each as its own class with its own weighted value. A Bayesian model with three classes, one Rasch, one unscalable, and the last one a variant of the hyperbolic cosine model (Andrich & Lou, 1993) could be examined. This would be particularly useful in extending this research to areas of surveys of agreement. Depending on attitude, some respondents may answer survey items which fit a cumulative model while others respond in a way more consistent with an unfolding model. Data can be contaminated using the unfolding approach as a starting point for what a distinct systematic form of contamination would look like.

It would be useful to explore the baseline 1-PL model as well as other models (e.g., 2-PL, 3-PL, hyperbolic cosine model), under additional conditions. In particular the interaction between θ and item difficulty should be explored to determine if factor pattern remains strongest when item difficulties and θ are close together as suggested in the current analysis. When values for item difficulty and person ability, θ , are close to each other information is higher than when further apart (Hambleton et al., 1991). It is interesting to see how this would be reflected in observed patterns under real testing conditions. Do the item difficulties that best target a sample also have the highest factor patterns? If so we might be able to determine the patterns that should be found in a testing condition and help detect systemic and random departure from expected patterns for the baseline model.

Conclusion

The current investigation explores original data using residuals to look for suspicious patterns that could indicate an alternative and better fit. Specifically, this approach was inspired by the perspective of mixture modeling. The data were reexpressed in terms of weights to examine the data set by splitting it into the weighted proportion that fits both the model and that which is misfit from the model. We examined residuals of a Rasch model by incorporating a mixture perspective with an unscalable class as a type of residual. We reexpressed the original data as residual fit to the Rasch class by using the unscalable class and examined what was left over by looking for possible patterns to explain the results. This regression-inspired Rasch residual permits exploration of misfit in a novel way. It was found that systematic contamination in smaller proportions of the sample size showed some clear departures from Rasch patterns. In many random conditions, the unscalable class was extracting contamination that had no distinguishable pattern. Several times an exceptional pattern was discovered. A Rasch pattern has been detected using factor analysis and is

identified as a maximum pattern at maximum information. When residuals are detected, it might benefit the researcher to investigate why this is happening. Some patterns might indicate guessing, whereas others like the crossing pattern in the reversed Rasch contamination might lead to investigating curriculum to determine if multiple methods are being taught.

The goal of the current investigation was to introduce this technique to examine residuals from a Bayesian Rasch IRT model and although the unscalable residual may have been a crude proxy for a residual, it still achieved this goal to further examine reweighted data when contamination was suspect. Further development of a more sophisticated model rather than just a 1-PL model with an unscalable residual would greatly aid in detection of contamination.

Some patterns exist for the Rasch baseline conditions explored which seem to be driven by item difficulties' proximity to the average person ability θ . In the residual-weighted data, patterns different from Rasch patterns were present. The random contamination tended to yield a weak flat pattern. The systematic contamination showed an interesting and opposite effect to the Rasch baseline. This contamination is easier to detect in the unweighted data when systematic contamination exists as some adulteration of the data can be found.

Declaration of Conflicting Interests

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

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

References

- Andreassen, S., Woldbye, M., Falck, B., & Andersen, S. K. (1987). MUNIN: A causal probabilistic network for interpretation of electromyographic findings. *Proceedings of the 10th International Joint Conference on Artificial Intelligence* (pp. 366-372). Milan, Italy: Kaufmann.
- Chen, T.-H., & Davison, M. L. (1996). A multidimensional scaling, paired comparisons approach to assessing unidimensionality in the Rasch model. In M. Wilson & G. Engelhard (Eds.), *Objective measurement: Theory into practice* (Vol. 3, pp. 309-334). Stanford, CT: Ablex.
- Chung, H., Loken, E., & Schafer, J. (2004). Difficulties in drawing inferences with finite-mixture models: A simple example with a simple solution. *The American Statistician*, 58, 152-158.
- Chyn, S., Tang, K. L., & Way, W. D. (1994). *An investigation of IRT-based assembly of the TOEFL test* (ETS Research Report TR-09). Princeton, NJ: Educational Testing Service.
- Dardick, W. R. (2010). *Reweightings data in the spirit of Tukey: Using Bayesian posterior probabilities as Rasch residuals for studying misfit* (Doctoral dissertation). Retrieved from Digital Repository at the University of Maryland: <http://hdl.handle.net/1903/11096>

- Dayton, C. M., & Macready, G. B. (1980). Scaling model with response errors and intrinsically unscalable respondents. *Psychometrika*, 45, 343-356.
- Divgi, D. R. (1986). Does the Rasch model really work for multiple choice items? Not if you look closely. *Journal of Educational Measurement*, 23, 283-298.
- Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2004). *Bayesian data analysis* (2nd ed.). Boca Raton, FL: Chapman & Hall.
- Gentner, D., & Gentner, D. R. (1983). Flowing waters or teeming crowds: Mental models of electricity. In D. Gentner & A. L. Stevens (Eds.), *Mental models* (pp. 99-129). Hillsdale, NJ: Lawrence Erlbaum.
- Goodman, L. A. (1975). A new model for scaling response patterns: An application of the quasi independence concept. *Journal of the American Statistical Association*, 70, 755-768.
- Gorsuch, R. L. (1983). *Factor analysis* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.
- Hambleton, R. K., Swaminathan, H., & Rogers, J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage.
- Hoaglin, D. C., Mosteller, F., & Tukey, K. W. (Eds.). (1983). *Understanding robust and exploratory data analysis*. New York, NY: Wiley.
- Holland, P. W., & Wainer, H. (Eds.). (1993). *Differential item functioning*. Hillsdale, NJ: Lawrence Erlbaum.
- Kelderman, H. (1984). Loglinear Rasch model tests. *Psychometrika*, 49, 223-245.
- Kingston, N. M., & Dorans, N. J. (1982). *The effect of the position of an item within a test on item responding behaviors: An analysis based on item response theory* (GRE Board Professional Report GREB 79-126P and ETS Research Report 82-22). Princeton, NJ: Educational Testing Service.
- Kline, R. B. (2013). Exploratory and confirmatory factor analysis. In Y. Petscher & C. Schatsschneider (Eds.), *Applied quantitative analysis in the social sciences* (pp. 171-207). New York, NY: Routledge. Retrieved from <http://psychology.concordia.ca/fac/kline/library.html>
- MacCallum, R. C., Widaman, K. F., Zhang, S., & Hong, S. (1999). Sample size in factor analysis. *Psychological Methods*, 4, 84-99.
- Mead, R. (1976). *Fit of data to the Rasch model through the analysis of residuals* (Unpublished doctoral dissertation). University of Chicago, Chicago, IL.
- Mislevy, R. J. (1986). Bayes model estimates in item response models. *Psychometrika*, 51, 177-195.
- Mislevy, R. J., & Verhelst, N. (1990). Modeling item responses when different subjects employ different solution strategies. *Psychometrika*, 55, 195-215.
- Mosteller, F., & Tukey, J. W. (1977). *Data analysis and regression: A second course in statistics*. Reading, MA: Addison-Wesley.
- Rogers, H. J., & Hattie, J. A. (1987). A Monte Carlo investigation of several person and item fit statistics for item response models. *Applied Psychological Measurement*, 11, 47-57.
- Rost, J. (1990). Rasch models in latent classes: An integration of two approaches to item analysis. *Applied Psychological Measurement*, 14, 271-282.
- Rupp, A. (2003). Item response modeling with BILOG-MG and MULTILOG for Windows. *International Journal of Testing*, 3, 365-384.
- Smith, R. M. (1986). Person fit in the Rasch model. *Educational and Psychological Measurement*, 46, 359-372.
- Smith, R. M. (1988). The distributional properties of Rasch standardized residuals. *Educational and Psychological Measurement*, 48, 657-667.

- Smith, R. M. (1991). The distributional properties of Rasch item fit statistics. *Educational and Psychological Measurement*, 51, 541-565.
- Spiegelhalter, D. J., Thomas, A., Best, N. G., & Lunn, D. (2003). *WinBUGS Version 1.4: User manual*. Cambridge, England: Cambridge Medical Research Council Biostatistics Unit. Retrieved from <http://www.mrc-bsu.cam.ac.uk/software/bugs/>
- Wainer, H., & Wright, B. (1980). Robust estimation of ability in the Rasch model. *Psychometrika*, 45, 373-391.
- Wright, B. D. (1995). Diagnosing person misfit. *Rasch Measurement Transactions*, 9, 430-431.
- Wright, B. D., & Tennant, A. (1996). Sample size again. *Rasch Measurement Transactions*, 9, 468.
- Yamamoto, K. (1987). *A model that combines IRT and latent class models* (Unpublished doctoral dissertation). University of Illinois, Champaign-Urbana, IL.
- Yamamoto, K. (1989). *Hybrid model of IRT and latent class models* (ETS Research Report RR-89-41). Princeton, NJ: Educational Testing Service.
- Yamamoto, K. (1995). *Estimating the effects of test length and test time on parameter estimation using the HYBRID model* (ETS Research Report RR-95-2). Princeton, NJ: Educational Testing Service.