Maximum Likelihood Item Easiness Models for Test Theory Without an Answer Key

Educational and Psychological

Measurement

2015, Vol. 75(1) 57–77

© The Author(s) 2014

Reprints and permissions:

sagepub.com/journalsPermissions.nav

DOI: 10.1177/0013164414527448

epm.sagepub.com

Stephen L. France¹ and
William H. Batchelder²

Abstract

Cultural consensus theory (CCT) is a data aggregation technique with many applications in the social and behavioral sciences. We describe the intuition and theory behind a set of CCT models for continuous type data using maximum likelihood inference methodology. We describe how bias parameters can be incorporated into these models. We introduce two extensions to the basic model in order to account for item rating easiness/difficulty. The first extension is a multiplicative model and the second is an additive model. We show how the multiplicative model is related to the Rasch model. We describe several maximum-likelihood estimation procedures for the models and discuss issues of model fit and identifiability. We describe how the CCT models could be used to give alternative consensus-based measures of reliability. We demonstrate the utility of both the basic and extended models on a set of essay rating data and give ideas for future research.

Keywords

cultural consensus theory (CCT), essay grading, maximum likelihood

¹University of Wisconsin–Milwaukee, Milwaukee, WI, USA

²University of California, Irvine, CA, USA

Corresponding Author:

Stephen L. France, Lubar School of Business, University of Wisconsin–Milwaukee, 3202 N. Maryland Avenue, Milwaukee, WI 53201-0742, USA.

Email: france@uwm.edu

Introduction

The problem of analyzing ratings from multiple raters is an established problem in psychometrics and in educational testing research. Consider a simple educational scenario, where n students write an essay and are graded on the content of the essay. Grading for essays is not an exact science, so to increase reliability, each essay is graded by multiple raters. Research in this area has traditionally concentrated on measuring the reliability of the ratings; see, for example, Raykov and Marcoulides (2011, pp. 147-181).

In this article, we take a slightly different approach. We consider each rater to have some underlying latent competency, which is an inverse measure of the rater's error variance. The aggregate rating for an item is not calculated as a simple average of the scores of the multiple raters, but as a weighted average, where each score is weighted by the rater's competency. The primary advantage of this approach is that the effect of inconsistent/poor quality raters can be minimized when calculating the aggregate score, which is calculated endogenously by the model. Thus, the overall reliability of the score is increased. The models described in this article come under the banner of cultural consensus theory (CCT). CCT is a technique with roots in scientific anthropology and was created to help aggregate and define cultural concepts using the idea of a common culture. In statistical terms, CCT is a method for information pooling or data aggregation. The original application for CCT was the analysis of folk medical beliefs (Romney, Weller, & Batchelder, 1986), which used simple dichotomous CCT (Batchelder & Romney, 1986; Batchelder & Romney, 1988). Modern reviews of CCT include Batchelder and Anders (2012) and Oravecz, Anders, and Batchelder (2013).

In this article, we examine and extend the continuous response CCT model introduced in Batchelder and Romney (1989). Given a set of user ratings, a competency parameter is calculated for each user and a competency weighted consensus aggregate rating is calculated for each item. An extension of the model allows for bias parameters to be calculated for each user. The authors describe a fixed point optimization procedure for the model. The basic model is built on several axioms. The axioms include common truth (Axiom 1), random error (Axiom 2), local independence (Axiom 3), and inhomogeneous variance (Axiom 4). Axiom 2 can be adapted to account for additive bias and multiplicative bias. Full technical details of the basic model and axioms are described in the appendix. This description follows the treatment of Batchelder and Romney (1989).

In the subsequent sections of this article, we describe several extensions to the Batchelder and Romney (1989) model. We describe two methods for modeling the rating easiness/difficulty for individual items. We describe the advantages and disadvantages of different approaches to optimizing both the original and the extended models. We show how the continuous CCT models described in this article can be used to produce competency adjusted reliability metrics. We describe a methodology for ensuring the identifiability of model parameters. We illustrate the usefulness of both the basic model and the model extensions using an illustrative grading application and then give suggestions for future work.

Cultural Consensus Analysis and Item Easiness/Difficulty

In rating/grading tasks, some items may be harder to rate than other items. The traditional method of modeling item difficulty in exam scoring models is the Rasch model (see Fischer & Molenaar, 1995). Given multiple items, multiple subjects, and simple dichotomous data, fitting the Rasch model gives both subject score and item difficulty parameters. A version of the Rasch model for continuous data is given in Müller (1987). A CCT model incorporating item difficulty has been developed for dichotomous CCT (Batchelder & Romney, 1988; Karabatsos & Batchelder, 2003), where item difficulty is conceptualized in a similar manner to the Rasch model. Overall, the underlying Rasch principle of calculating both the aggregate item difficulty and the average subject score within a single procedure provides a parsimonious representation for a range of problems. In a standard examination setting, the items are actual “questions” and the subjects are exam takers. In an essay evaluation setting, the items are the student’s essays and the subjects are the exam raters. The idea of item difficulty can be expanded to a general rating setting, where each item is assigned both an item score and an item difficulty and each rater is assigned a competency.

We propose two methods of incorporating variable item difficulty/easiness into the standard consensus model. Both these methods split the measure of competency into rater competency and item easiness. We use “easiness” rather than “difficulty,” as easiness is consistent with competency when modeling the error variance.

Method 1: Multiplicative Scaling Factor Model

Let $\mathbf{X}$ be an n rater $\times m$ item matrix of item ratings or scores, $\mathbf{d} \in \mathbb{R}^{n+}$ be an n rater $\times 1$ vector of user competencies, and $\mathbf{z} \in \mathbb{R}^m$ be a $1 \times m$ item vector of aggregate ratings, which is denoted as the answer key. For each item k , Let $\beta_k \in \mathbb{R}^+$ be an item easiness scaling factor, where a value of $\beta_k < 1$ indicates that ratings will have a smaller error variance than $\sigma_i^2(\varepsilon_{ik}) = d_i^{-1}$ and a value of $\beta_k > 1$ indicates that ratings will have a larger error variance than $\sigma_i^2(\varepsilon_{ik}) = d_i^{-1}$. Formally, Axiom 4 from the basic model can be rewritten as follows:

Axiom 4a: Multiplicative Item Easiness: For each rater i , define the rater competence $d_i \in \mathbb{R}^+$ as the inverse error variance and for each item k , define a scaling factor $\beta_k \in \mathbb{R}^+$, so that $\sigma_{ik}^2(\varepsilon_{ik}) = (d_i \beta_k)^{-1}$.

Given the scaling factor β_k , the variance is now indexed across i and k . The log-likelihood (LL) function for the multiplicative scaling factor model is given in Equation (1). Bias corrected versions of the model can be created by adding bias to the $(x_{ik} - z_k)$ term, as per Equations (A5), (A7), and (A9).

$$LL(\mathbf{d}, \mathbf{z}, \boldsymbol{\beta} | \mathbf{X}) = \sum_{k=1}^m \sum_{i=1}^n \left[\log \left(\sqrt{\frac{\beta_k d_i}{2\pi}} \right) + \left(\frac{-\beta_k d_i (x_{ik} - z_k)^2}{2} \right) \right]. \quad (1)$$

This multiplicative model can be connected to the Rasch model. The precision for i and k can be defined as the inverse of the variance, so that the precision is $d_i\beta_k$. In Equation (2), $\phi_k \in [0, 1)$ is defined as the rescaled precision, so that when $d_i\beta_k = 0$, $\phi_k = 0$, and as $d_i\beta_k$ becomes larger, $\phi_k \rightarrow 1$.

$$\phi_{ik} = \frac{d_i\beta_k}{1 + d_i\beta_k}. \quad (2)$$

As $\phi_{ik} \in [0, 1)$, it can be decomposed into a user score term p_i and an item difficulty term q_k using the standard Rasch formulation for the probability of a correct answer, given in Equation (3):

$$\phi_{ik} = \frac{e^{(p_i - q_k)}}{1 + e^{(p_i - q_k)}}. \quad (3)$$

If the difficulty parameter q_k is replaced with the negative of an “easiness” parameter e_k , then equating Equation (3) with Equation (2) gives $e^{(p_i + e_k)} = d_i\beta_k$ or $p_i + e_k = \log(d_i) + \log(\beta_k)$. Thus, if the precision values are transformed using Equation (2), the multiplicative scaling factor model parameters are the log transformed Rasch model parameters.

Method 2: Additive Scaling Factor Model

Additive models have been used to decompose and examine rater effects. For example, de Gruijter (1984) gives an additive model for examining rater effects for overlapping raters. Raymond and Viswesvaran (1993) give a range of least squares fitting techniques for additive rater models. The additive scaling factor model follows in this tradition. If it is assumed that the error variance is proportional to both the inverse of the sum of the user competency and the item easiness, Axiom 4 can be rewritten as follows:

Axiom 4b: Additive Item Easiness: For each rater i , define $d_i \in \mathbb{R}^+$ as the rater competency and for each item k , define $\beta_k \in \mathbb{R}^+$ as the item easiness. The inverse error variance has both a rater competency component and an item easiness component, so that $\sigma^2(\varepsilon_{ik}) = (d_i + \beta_k)^{-1}$.

The log-likelihood function for the additive scaling factor model is given in Equation (4). Bias corrected versions of the model can be created by adding bias to the $(x_{ik} - z_k)$ term as per Equations (A5), (A7), and (A9).

$$LL(\mathbf{d}, \mathbf{z}, \boldsymbol{\beta} | \mathbf{X}) = \sum_{k=1}^m \sum_{i=1}^n \left[\log \left(\sqrt{\frac{d_i + \beta_k}{2\pi}} \right) + \left(\frac{-(d_i + \beta_k)(x_{ik} - z_k)^2}{2} \right) \right]. \quad (4)$$

Model Fitting and Optimization

Each of the models described in this article can be fit by maximizing its log-likelihood function using continuous optimization. We have expanded the basic methodology in Batchelder and Romney (1989) and have created a flexible set of optimization routines that can deal both with constrained and unconstrained models. In particular, we have extended existing work by developing a methodology for examining and dealing with model identifiability in continuous CCT models.

We have created a software library in MATLAB (France, 2013) for both the basic continuous CCT model and the two item easiness extensions. Three optimization methods are implemented in the software. These are (a) fixed point optimization, (b) gradient-free optimization, and (c) first-order gradient optimization. The fixed point optimization procedure is implemented in previous work on continuous CCT (Batchelder & Romney, 1989). The gradient-based optimization and derivative-free optimization algorithms are standard nonlinear optimization techniques implemented by MATLAB and are described in Bazaraa, Sherali, and Shetty (2006). The fixed point optimization procedure, where each parameter vector (e.g., $\mathbf{z}$, $\mathbf{d}$, $\mathbf{b}_A$, $\mathbf{b}_M$, $\mathbf{\beta}$) is calculated sequentially using first-order conditions until convergence, is by far the fastest, most efficient optimization scheme, but does not converge for every single model. For example, the additive item easiness factor model requires both the values of $(d_i + \beta_k)$ and the values of the individual d_i and β_k parameters to be constrained to be positive and cannot be fit using the unconstrained fixed point method.

Implementation Issues

In initial experiments, all three optimization methods converged to good solutions. In cases with only a few raters, one rater may become dominant and that rater's competency will go toward infinity. A similar problem can occur for item easiness values. The problem of dominant competency and item easiness values can be mitigated by setting bounds on the allowed competency and item easiness parameters. Empirical work in France and Batchelder (2013) shows that as the number of items increases, the chance of dominant competencies decreases. Based on empirical experiments on generated data (France & Batchelder, 2013), a competency bound of around $d_i = 5$ gives good results, though the exact optimal value may be dependent on the data set, so some tuning may be necessary.

The CCT models described in this article can include parameters for item answers, additive biases, multiplicative biases, and item easiness values. Fitting multiple sets of parameters together can lead to a loss of model identifiability. In initial experiments, we found that there is no identifiability problem when fitting the models without bias, that is, the optimal values of $\mathbf{d}$ and $\mathbf{z}$ are unique. However, introducing bias, can lead to a loss of identifiability. Consider a situation where there is a $1 \times m$ answer vector $\mathbf{z}$, an $n \times 1$ competency vector $\mathbf{d}$, and an $n \times 1$ additive bias vector $\mathbf{b}_A$, which together give a solution with the optimal log-likelihood for some data matrix $\mathbf{X}$. Replace $\mathbf{z}$ with $\mathbf{z}^* = \mathbf{z} - \mathbf{1}$ and replace $\mathbf{b}_A$ with $\mathbf{b}_A^* = \mathbf{b}_A + \mathbf{1}$. Consider the

likelihood function given in Equation (A5) for the modified parameters, that is, $LL(\mathbf{d}, \mathbf{z}^*, \mathbf{b}_A^* | \mathbf{X})$. Each of the log terms is unchanged, as $\mathbf{d}$ is unchanged. Each of the squared terms can be expressed as Equation (5), which is also unchanged, so $LL(\mathbf{d}, \mathbf{z}^*, \mathbf{b}_A^* | \mathbf{X}) = LL(\mathbf{d}, \mathbf{z}, \mathbf{b}_A | \mathbf{X})$:

$$\frac{-d_i(x_{ik} - b_{Ai}^* - z_k^*)^2}{2} = \frac{-d_i(x_{ik} - (b_{Ai} + 1) - (z_k - 1))^2}{2} = \frac{-d_i(x_{ik} - b_{Ai} - z_k)^2}{2}. \quad (5)$$

The MATLAB software developed to implement the models in this article contains a set of fix parameters to deal with identifiability issues. In the implemented software, users can specify an input vector with the entries $[Fixz, Fixd, Fixb_A, Fixb_M]$, where each entry of the vector can be 0, 1, or -1 . Here, 0 indicates no fix, 1 a partial fix (one variable), and -1 a total fix (all variables). The fixed values are set by running a restricted, identified version of the model and then fixing a single value (the one with the maximum partial log-likelihood) or values from this restricted model. Based on initial results, fixing the value of z_k for a single item k gives a fully identified model when bias is added to the model. The differences between fixing a single value (a partial fix) and all the variables (a full fix) are both practical and philosophical. A partial fix still fits the full model. For example, for the basic model with additive bias and the fix vector $[1, 1, 0, 0]$, bias is a component of the overall cognitive model and the variable fixes do not alter the overall log-likelihood of the model. A full fix essentially prioritizes the fitting of the subset of variables from the restricted model. For example, given the basic model with additive bias and the fix vector $[-1, -1, 0, 0]$, the values of $\mathbf{z}$ and $\mathbf{d}$ are calculated for the basic model without bias and then $\mathbf{b}_A$ is calculated for the model with bias with $\mathbf{z}$ and $\mathbf{d}$ fully fixed. Here the bias is a residual measure of bias for the model without bias.

Cultural Consensus Theory and Reliability Theory

How does consensus theory fit in with traditional psychometric methods of reliability for multiple raters? There are several methods for calculating reliability for continuous data. We describe one of these methods and discuss how it can be adapted to fit the consensus paradigm and related models. Most methods for calculating agreement use some measure of variance or correlation. Berry and Mielke (1988) extend the kappa measure (Cohen, 1960) to continuous multivariate data and use Euclidean distances to measure differences between scores. Janson and Olson (2001) use squared Euclidean distances rather than standard Euclidean distances. We describe this formulation, as it fits well with the squared Euclidean distance terms in the CCT likelihood functions. The original Janson and Olson (2001) agreement measure allows for each item to have multiple traits. For simplicity, and to fit with the consensus model, we only describe a single trait version of the measure. The expected disagreement D_e is defined as the average squared Euclidean distance between a rater's score for one

item k and any other rater's score for any other item. The observed disagreement D_o is defined as the average squared Euclidean distance between a rater's score for one item k and any other rater's score for the same item.

$$D_e = \frac{2}{m^2 n(n-1)} \sum_{i=1}^{n-1} \sum_{j=i+1}^n \sum_{k=1}^m \sum_{l=1}^m (x_{ik} - x_{jl})^2 \quad (6)$$

$$D_o = \frac{2}{mn(n-1)} \sum_{i=1}^{n-1} \sum_{j=i+1}^n \sum_{k=1}^m (x_{ik} - x_{jk})^2 \quad (7)$$

The overall level of agreement between raters is denoted by ι (ι), which is defined in Equation (8) and has a maximum value of 1 when there is no observed disagreement, that is, $D_o = 0$.

$$\iota = 1 - \frac{D_o}{D_e} \quad (8)$$

The value of ι can be taken as a measure of the reliability of an overall score, based on the agreement between raters.

For a CCT answer key, outlying raters with low agreement contribute less to the overall score than raters with high agreement. This should boost the overall level of agreement of the aggregate score. For each k , calculating the value of $\partial f / \partial z_k$ for the simple model with no bias from Equation (A3) gives Equation (9) and setting this value to zero gives the first-order conditions shown in Equation (10).

$$\frac{\partial f}{\partial z_k} = \sum_{i=1}^n [d_i (x_{ik} - z_k)] \quad (9)$$

$$z_k = \frac{\sum_{i=1}^n d_i x_{ik}}{\sum_{i=1}^n d_i} \quad (10)$$

Thus, the contribution to $\mathbf{z}$ for user i is directly relative to the value of the competency of i , d_i . Accounting for competency, Equations (6) and (7) can be rewritten as Equations (11) and (12):

$$D_e = \frac{2}{m^2 n(n-1)} \frac{\sum_{i=1}^{n-1} \sum_{j=i+1}^n \sum_{k=1}^m \sum_{l=1}^m d_i d_j (x_{ik} - x_{jl})^2}{\sum_{i=1}^{n-1} \sum_{j=i+1}^n \sum_{k=1}^m \sum_{l=1}^m d_i d_j} \quad (11)$$

$$D_o = \frac{2}{mn(n-1)} \frac{\sum_{i=1}^{n-1} \sum_{j=i+1}^n \sum_{k=1}^m d_i d_j (x_{ik} - x_{jk})^2}{\sum_{i=1}^{n-1} \sum_{j=i+1}^n \sum_{k=1}^m d_i d_j}. \quad (12)$$

Inaccuracies can occur in correlational measures such as Cohen's kappa when users have intrinsic biases (Di Eugenio & Glass, 2004). However, scores from bias corrected competency models can account for errors due to bias. In fact, the first derivatives of any model described in this article can be used to provide an adjusted measure of agreement. For example, the value of $\partial f / \partial z_k$ for the model with multiplicative item easiness and additive bias is given in Equation (13) and the associated first-order equation for z_k is given in Equation (14).

$$\frac{\partial f}{\partial z_k} = \sum_{i=1}^n [d_i \beta_k (x_{ik} - b_{Ai} - z_k)]. \quad (13)$$

$$z_k = \frac{\sum_{i=1}^n d_i \beta_k (x_{ik} - b_{Ai})}{\sum_{i=1}^n d_i \beta_k}. \quad (14)$$

The resulting disagreement values, accounting for competency, item difficulty, and bias are given in Equations (15) and (16):

$$D_e = \frac{2}{m^2 n(n-1)} \frac{\sum_{i=1}^{n-1} \sum_{j=i+1}^n \sum_{k=1}^m \sum_{l=1}^m d_i \beta_k d_j \beta_l ((x_{ik} - b_{Ai}) - (x_{jl} - b_{Aj}))^2}{\sum_{i=1}^{n-1} \sum_{j=i+1}^n \sum_{k=1}^m \sum_{l=1}^m d_i \beta_k d_j \beta_l}. \quad (15)$$

$$D_o = \frac{2}{mn(n-1)} \frac{\sum_{i=1}^{n-1} \sum_{j=i+1}^n \sum_{k=1}^m d_i \beta_k d_j \beta_k ((x_{ik} - b_{Ai}) - (x_{jk} - b_{Aj}))^2}{\sum_{i=1}^{n-1} \sum_{j=i+1}^n \sum_{k=1}^m d_i \beta_k d_j \beta_k}. \quad (16)$$

In a practical sense, the agreement measures can be used to calculate the level of "reliability" of the consensus answer key $\mathbf{z}$. Consider a situation where raters are highly competent, but have differing biases. The simple agreement measure given in Equation (8) will give a poor overall level of agreement, despite the high competency of the raters. A competency adjusted metric would give a better overall idea of the value of the information given by the raters.

The metrics described by Berry and Mielke (1988) and Janson and Olson (2001) can be described in terms of the analysis of variance (ANOVA) framework and the

sum of squares decomposition of the rater answers. In fact, Barnhart, Haber, and Lin (2007) describe an agreement model that uses a two-way fixed effect model and splits the mean squared dissimilarity (MSD) of $\mathbf{X}$ into interrater MSD and intrarater MSD. The ANOVA framework can be used to help define hypotheses. For example, “did essay set one provide more reliable answers than essay set 2?” or “did the training session given to a subset of raters change their rating behavior?” Given the ANOVA framework and using a partial F statistic, the consensus adjusted reliability metrics could be used to help evaluate model fit in a similar manner to the model log-likelihood. Modern reliability analysis often uses generalizability theory (Raykov & Marcoulides, 2011; Shavelson, Webb, & Rowley, 1989), which uses an ANOVA decomposition of error to identify different sources of error. By using the framework in Barnhart, Haber, and Lin (2007), the competency adjusted sums of squares error could be decomposed in a similar manner.

Exam Grading Illustrative Example

Cultural Consensus Theory and Exam Grading

We demonstrate the use of the CCT models described in this article using an illustrative real world example. Though the models can be used with a wide range of rating/estimation data, educational rating/grading data are particularly relevant to CCT. Consider a situation where multiple raters grade a set of essays. To maintain fairness, there should be a high degree of consistency between raters. Traditionally, reliability analysis has been used to analyze the effectiveness of essay graders; see, for example, Braun (1988). However, reliability analysis gives limited information on the relative competency of the individual graders. Some reliability measures, for example, the ι measure described in Janson and Olsson (2001), allow for the weighting of items, but the weighting is a priori. When analyzing essay grading patterns, the items are the individual “essays” and it is difficult and time consuming to guess the easiness/difficulty of the essays. Applying the item easiness CCT models to the data gives a set of grader competencies, a set of competency weighted aggregate essay grades, and a set of item easiness factors. Information measures based on the model log-likelihood, for example, the Akaike information criterion (AIC; Akaike, 1974), or one of the adjusted reliability metrics described in the previous section, can be used to provide an overall aggregate measure of grader reliability.

We used data from an essay scoring data set, “The Hewlett Foundation: automated essay scoring” (Kaggle, Inc., 2012) data set. The data set contains sets of student essays for several different essay prompts. Each essay set contains the source essays and scores for each essay given by two expert graders. We take data from an essay set, which has the following prompt:

We all understand the benefits of laughter. For example, someone once said, “Laughter is the shortest distance between two people.” Many other people believe that laughter is an

important part of any relationship. Tell a true story in which laughter was one element or part.

A grading rubric is associated with the essay set. The essay grading rubric contains six different attributes: “Ideas and Content,” “Organization,” “Voice,” “Word Choice,” “Sentence Fluency,” and “Conventions.” Each essay is to be graded on an integer scale from 1 to 6 for each of the attributes. We chose 50 essays from the essay set. The essays were not chosen completely randomly. We ordered the essays into deciles by quality, using the scores given by the two expert graders. We then randomly selected 5 essays from each of the 10 deciles. This ensured that there was a good range of essay quality in the experimental data set. We brought in 12 student graders. The students were compensated for their time. The student graders were undergraduate juniors or seniors at the University of Wisconsin–Milwaukee. Each grader was given the training materials for the essay, which contained a detailed description of the grading rubric and four example graded essays. The students were tasked with reading each essay and grading it on each of the six attributes described in the grading rubric. The students were given 30 minutes to view the training material and a minimum of 120 minutes and a maximum of 150 minutes to grade the 50 essays (2.4–3 minutes per essay). The time per essay was chosen based on the standard time given to ETS exam graders. The minimum time was to ensure that students did not rush through the essay grading process without thought.

Results

The resulting data set from the experiment contains ratings for the 2 expert raters and the 12 student raters for each of the 50 chosen essays. The rubric guide recommends calculating an overall essay score by summing the constituent attribute scores for each essay. We took two approaches to calculating the overall essay scores. For the first approach, we summed all the attribute scores. As per classical test theory (Raykov & Marcoulides, 2011, pp. 115–136), we assumed that the overall additive scale corresponds to a continuous scale with even intervals between item values. For the second approach, we did not make any initial measurement assumptions regarding the attribute scores. Instead, we used multiple correspondence analysis (Greenacre & Blasius, 2006) to map the data onto a continuous scale. Multiple correspondence analysis creates chi-squared distances between data points from a table of associations between categorical variables. A singular value decomposition can then be used to map the data points to a continuous dimension. The resulting correspondence analysis scores had a Pearson correlation of .9783 with the additive scores, which suggests that the additive classical test theory model holds for the data.

The single culture CCT models described in this article are based on the assumption that the raters belong to a single culture. Given that in the example, graders are grading from a single rubric, one may be able to make the assumption of a single culture, but this is not a given as raters may interpret a single rubric in multiple ways.

Table 1. Results for Simple Model.

Model	Param	Data	PMax	LL*	AIC	Mind	Maxd
Basic	64	Add	0	−1922.3	3972.6	0.0284	0.1639
Basic+	78	Add	0	−1817.9	3791.8	0.0356	0.2046
Basic×	78	Add	0	−1827.7	3811.4	0.0322	0.1951
Basic+×	92	Add	0	−1776.77	3737.54	0.0408	0.2077
Basic	64	CA	0	−1952.1	4032.2	0.0246	0.1573
Basic+	78	CA	0	−1815.6	3787.2	0.0364	0.1967
Basic×	78	CA	0	−1837.6	3831.2	0.033	0.1917
Basic+×	92	CA	0	−1781.73	3747.46	0.0454	0.2089

Note. AIC = Akaike information criterion; CA, correspondence analysis; Param = number of parameters; PMax = proportion of user competencies and item easiness parameters that take the maximum value of 5; LL* = optimal log-likelihood.

Fortunately, most CCT models have the property that the interrater correlations across items exhibit an approximate one-factor structure as exhibited by Spearman’s law of tetrads, for example, Batchelder and Romney (1988), Batchelder and Anders (2012). One way of verifying that informants have a single culture is to perform a factor analysis on the interrater correlations and then examine the eigenvalues for evidence of a one-factor structure (e.g., Anders & Batchelder, 2012). For both the additive scores and the correspondence analysis scores, we calculated correlations between the 14 raters and performed a minimum residual factor analysis using the `fa()` function from the “psych” library in R (Revelle, 2013). The eigenvalues for the additive scores were [$\lambda_1 = 11.41, \lambda_2 = 0.62, \lambda_3 = 0.35, \lambda_4 = 0.31$] and the eigenvalues for the correspondence analysis scores were [$\lambda_1 = 11.60, \lambda_2 = 0.57, \lambda_3 = 0.33, \lambda_4 = 0.27$]. For both solutions, the first factor accounted for more than 99% of the solution variance. This strongly suggests a one-factor solution, supporting the single culture assumption of the models.

We ran each of the three models (the basic model and the two-item easiness models), setting a maximum competency and item easiness of 5, to ensure that no rater’s precision was completely dominant over the other raters’ precisions. We ran each model with no bias, additive bias (+), multiplicative bias (×), and simultaneous fitting of both additive and multiplicative bias (+×).

The results for the basic model are given in Table 1 and the results for the item easiness models (multiplicative and additive) are given in Table 2. The model name, the number of parameters, the data preprocessing method (additive or correspondence analysis [CA]), the optimal log-likelihood value (LL*), the AIC value associated with the optimal log-likelihood ($2 \times \text{No. Parameters} - 2 \times \text{LL}^*$), the minimum and maximum values of the answer, the competency, and the item easiness parameters are given for each model. The bias type (+, ×, or +×) is appended after the model name. The AIC value can be used as a general measure of model fit, relative to the number of parameters. Using the AIC rather than the log-likelihood helps

Table 2. Results for Item Easiness Models.

Model	Param	Data	PMax	LL *	AIC	Mind	Maxd	MinQE	MaxQE
IEMultip ⁺	114	Add	0.0469	-1845.09	3918.18	1.7103	5	0.0065	5
IEMultip [×]	128	Add	0	-1770.5	3797	0.1099	1.1531	0.1018	1.8688
IEMultip [×]	128	Add	0	-1766.3	3788.6	0.1161	1.2532	0.0885	8.8509
IEMultip ⁺⁺	142	Add	0	-1730.59	3745.18	0.1841	1.4911	0.0559	1.6961
IEAdditive ⁺	114	Add	0.0156	-1858.8	3945.6	0.002	0.1294	0.0053	5
IEAdditive ⁺	128	Add	0	-1782.1	3820.2	0.0048	0.1732	0.0001	0.2698
IEAdditive [×]	128	Add	0	-1778.2	3812.4	0.0042	0.1593	0.0004	3.1954
IEAdditive ⁺⁺	142	Add	0	-1741.65	3767.3	0.003	0.141	0.0006	0.5877
IEMultip ⁺	114	CA	0.016	-1872.4	3972.8	1.764	5	0.0077	5
IEMultip ⁺	128	CA	0	-1775.6	3807.2	0.1706	1.4098	0.0889	1.0513
IEMultip [×]	128	CA	0	-1776.5	3809	0.1533	1.3666	0.0849	6.7815
IEMultip ⁺⁺	142	CA	0	-1745.62	3775.24	0.2124	1.4818	0.0744	0.8485
IEAdditive ⁺	114	CA	0.016	-1883.4	3994.8	0.001	0.1184	0.0071	5
IEAdditive ⁺	128	CA	0	-1783.1	3822.2	0.0028	0.1549	0.0013	0.2087
IEAdditive [×]	128	CA	0	-1785.3	3826.6	0.0038	0.152	0.0008	2.6256
IEAdditive ⁺⁺	142	CA	0.016	-1781.05	3846.1	0.0029	0.1613	0.004	5

Note. AIC = Akaike information criterion; CA, correspondence analysis; IE, item easiness; Param = number of parameters; PMax = proportion of user competencies and item easiness parameters that take the maximum value of 5; LL = log-likelihood.

Figure 1. Plot of multiplicative bias versus competency (d).

protect against overfitting when testing and comparing models. In addition, the proportion of user competencies and item easiness parameters that take the maximum value of 5 is given as “PMax.” The value of PMax can be considered to be a quality criterion for the solution, with solutions with low values of PMax considered to be good quality. The rationale here is that if too many users have the maximum competency or too many items have the maximum item easiness then this will hinder the interpretability of the solution.

The results for the basic models are given in Table 1. For the basic model, adding bias significantly decreased the AIC. For both score types, adding additive bias decreased the AIC by more than adding multiplicative bias and simultaneously adding additive and multiplicative bias decreased the AIC further. The additive scores gave a slightly better fit than the scores calculated with correspondence analysis. All four basic models gave PMax values of 0.

The results for the item easiness models are given in Table 2. For both score types, for the models without bias, the multiplicative item easiness model gave the lowest AIC and had a single item with maximum competency. For both score types, the additive and multiplicative scaling factor item easiness models gave a score of PMax = 0 when bias was added. As per the basic model, adding bias reduced the AIC. The multiplicative item easiness model with both additive and multiplicative bias had the lowest AIC for both additive and correspondence analysis scores.

As an illustrative example, we give scatterplot representations of the parameters for one of the models, the multiplicative item easiness model with multiplicative bias. In the first scatter plot, Figure 1, competency is plotted against bias. The two

Figure 2. Scatter plot of item easiness versus score.

expert graders are indicated with an “E” and the student graders are indicated with an “S.” One can see that one of the expert graders has a much stronger competence than all the other graders. The second expert grader is part of a cluster of graders with slightly above average bias and competency. Among students, there is a wide variety of biases and competencies. In the second scatter plot, Figure 2, for each item k , the item (essay) easiness β_k is plotted against the answer key (essay) score z_k . As the values of β_k are multiplicative, they are mostly clustered around 1. There is a correlation of -0.3471 ($p = .0135$) between essay easiness values and essay score values. In particular, there is a cluster of essays with high scores and low item easiness values. This suggests that for this essay prompt, it was harder to evaluate essays with minor errors than essays with major errors. One essay has the maximum item easiness parameter of 5. The content of the essay in its entirety is “I don’t like computers.” Every rater graded the essay as 1 out of 6 on every attribute, leading to zero item error and thus an item easiness parameter which theoretically could go to infinity without affecting the likelihood function. The essay with the lowest easiness value contains musings by a student on social acceptance in a group of peers in a foreign country. The inventiveness and use of the English language in the essay is of a high level, obviously higher than that of most of the other essays. However, the narrative is somewhat fractured and happenstance and the essay is a little shorter than some of the other essays. One can see why this essay was hard to evaluate. Given these two examples, and examining other essays with either high or low easiness values, the item easiness parameters seem to possess face validity.

Thus far, we have discussed the meaning of the different output parameters for CCT and how these parameters are fit from the initial ratings data. There are several ways that the parameters could be used to help inform decisions in essay grading applications. Consider a situation where graders are tested on a set of example evaluation essays or exam questions. Each grader can be assigned a competency. These competencies could be used to select graders from a pool of potential graders or to help assign graders to exams. Rather than assigning a fixed number of graders to an essay, the number of graders assigned could be selected based on the competencies of the graders. For example, one essay could be assigned to two high competency graders while another could be assigned to three lower competency graders. When calculating aggregate scores, scores could be calculated using the competencies gained from the example evaluation essays. In the case of the item easiness parameters, it is important to note that in the context of essay rating, “easiness” refers to the ease of grading an answer to a specific essay rather than the overall ease of grading for a certain essay prompt. In an actual test environment, the item easiness parameter could be used as a measure of answer score accuracy. If an essay has a low “easiness parameter” then additional graders could be assigned. For example, in the data set described in this section, if the “I don’t like computers” essay gained the same minimum score from all the initial raters then the “easiness” parameter would be high and additional raters would not be required. If the “exchange student” essay gained several conflicting ratings and a low “easiness” parameter then additional raters could be added to improve accuracy.

Discussion and Future Work

We have introduced a set of CCT models for analyzing rating data. The models take an n user $\times m$ item input matrix $\mathbf{X}$ and maximize a log-likelihood criterion function by fitting as output parameters an $n \times 1$ vector of user competencies $\mathbf{d}$, an $m \times 1$ answer key vector $\mathbf{z}$ of competency weighted aggregate ratings, and optionally an $n \times 1$ vector of additive biases $\mathbf{b}_A$, an $n \times 1$ vector of multiplicative biases $\mathbf{b}_M$, and a $1 \times m$ vector of item easiness values $\boldsymbol{\beta}$. We have implemented algorithms for fitting the models in MATLAB. The fixed-point optimization method was by far the fastest, most scalable optimization method, but constrained optimization techniques, implemented using MATLAB’s inbuilt optimization routines, were needed for the additive item easiness models.

CCT provides a cognitively based aggregation scheme that can correct for different rater biases in the aggregation process. However, when bias parameters are added to the model, model identifiability issues can occur. We implemented a method of fixing certain model parameters. In this method, the fixed values are taken from an optimized model with a fully identified subset of the model parameters. Constraining a single parameter value gives the same optimal solution as the unconstrained solution. Fixing all parameters of a given parameter type essentially introduces a

parameter fitting hierarchy. Further analytical and empirical work on analyzing identifiability issues in CCT models could be useful.

CCT models are applicable to a wide range of rating/questionnaire data from the social and behavioral sciences. Applications include anthropological analysis of cultural knowledge (Romney et al., 1986), cross-cultural analysis (Lamm & Keller, 2007), determining ties in a social network (Batchelder, Kumbasar, & Boyd, 1997), analyzing social survey data (Oravecz, Faust, & Batchelder, in press), grammaticality aggregating judgments of sentences in linguistics (Anders & Batchelder, 2013) and aggregating product reviews (France & Batchelder, 2013). In this article, we gave a specific example in the area of education. We demonstrated the utility of CCT models on a set of essay scoring data. CCT models are models of shared culture and shared belief. In this sense, when we used CCT to analyze essay scoring data, we were analyzing the culture of the essay graders. The graders with the highest competency were the graders closest to the center of the culture. In the example given in this article, graders were given a training rubric and a set of graded examples. If implemented properly, exam rubrics can help increase consistency among graders (Moskal & Leydens, 2000). In the context of CCT, the rubrics form the basic cultural knowledge. However, individual graders may stray from the rubric or may interpret the rubric differently than one another. The “culture” of the raters is defined by their actual essay rating behavior and may or may not reflect overall adherence to the grading rubric. For example, there may be a situation where there is a group of experienced essay graders who have previously used an old version of the rubric and a smaller group of new graders who have only ever used the current version of the rubric. If the majority of the raters are still influenced by the old rubric then these raters will be considered more competent than those strictly following the current rubric. In this case, a multi-culture (Anders & Batchelder, 2012) or clusterwise (France & Batchelder, 2013) CCT model may be required to split the raters into groups.

In the experiment, we implemented three different models, with four bias variants for each model. We gained parsimonious model information for all the models tested. All three models showed a noticeable improvement in fit as bias parameters were incorporated into the model. However, the results are only for one data set and the differences in the AIC values between the additive and multiplicative models and between the additive and multiplicative biases are small, so the best model may depend on the best interpretation for the task at hand.

The CCT models described in this article assume continuous input data. However, most ratings data are measured on an ordinal scale. The data set analyzed in the experiment had six different traits, each measured on an ordinal scale from 1 to 6. We calculated an overall scale value from the six traits using an additive model, as per classical test theory, and using multiple correspondence analysis. Correspondence analysis is a technique that can be used to approximate a continuous scale from ordinal data. In the case of ordinal data, it is more appropriate to calculate an overall continuous scale using correspondence analysis rather than factor analysis, as correspondence analysis is designed specifically to analyze ordinal and nominal

categorical variables. We found that the additive scores were very strongly correlated to the continuous correspondence analysis scores. This suggests that at least in the case of the essay grading, the additive scores can be taken as continuous. In fact, Kennedy, Riquier, and Sharp (1996) examine a range of Likert-type scale ratings data and show that when mapped to a continuous scale using correspondence analysis, the differences between Likert-type category values are almost even, allowing the data to be approximated as continuous. In addition, experiments in Batchelder, Strashney, and Romney (2010), which describes a model where ordinal boundaries are mapped onto a continuous scale, show that a continuous model gives a good approximation to Likert-type scale data.

For future work, explicit modeling of ordinal variables could be incorporated into the CCT optimization routine. The update equations for the fixed point technique for $\mathbf{z}$ minimizes the squared Euclidean distances between $\mathbf{z}$ and $\mathbf{x}_i$, for each rater i . Rather than explicitly minimizing the squared distance, an algorithm could be designed to minimize the deviation in rank order between $\mathbf{z}$ and $\mathbf{x}_i$. For example, Kruskal's (1964) up-down algorithm for minimizing rank deviation has been used to implement nonmetric (rank order) variants of multidimensional scaling. A similar approach could be used for the CCT procedures described in this article. An alternative approach to dealing with categorical data is to map the data to a fixed continuous scale and then determine the position of the boundaries between categories (Batchelder et al., 2010). A comparison between this approach and the approach described in this article of assuming an underlying continuous scoring scale could prove useful.

Appendix

The Basic Model

This section contains an outline of the basic continuous CCT model. The notation and data parameters used to describe the CCT model are defined as follows. Let $\mathbf{X}$ be an n rater $\times m$ item matrix of item ratings or scores, $\mathbf{d} \in \mathbb{R}^{n+}$ be an n rater $\times 1$ vector of user competencies, and $\mathbf{z} \in \mathbb{R}^m$ be a $1 \times m$ item vector of aggregate ratings, which is denoted as the answer key. The model axioms, as per Batchelder and Romney (1989), are given below:

Axiom 1: Common Truth: For a single culture, there is a set of latent ratings or answer key $\mathbf{z}$, where $z_k \in \mathbb{R}$ is the consensus rating for item k .

Axiom 2: Random Error: Let $x_{ik} \in \mathbb{R}$ be the rating given by rater i for item k , z_k be the answer key value of item k , and ε_{ik} the error for rater i and item k . The rating x_{ik} can be decomposed as per (A1):

$$x_{ik} = z_k + \varepsilon_{ik}, \quad (\text{A1})$$

where the error variables ε_{ik} are distributed according to the Gaussian distribution and have expected value $E(\varepsilon_{ik}) = 0$.

Axiom 3: Local Independence: The values of ε_{ik} are mutually stochastically independent.

Axiom 4: Inhomogeneous Variance: For each rater i , define the rater competence or precision $d_i \in \mathbb{R}^+$ as the inverse error variance, so that $d_i^{-1} = \sigma_i^2(\varepsilon_{ik})$.

The likelihood function for the basic continuous CCT model is given in Equation (A2) and the log-likelihood function is given in Equation (A3):

$$L(\mathbf{d}, \mathbf{z} | \mathbf{X}) = \prod_{k=1}^m \prod_{i=1}^n \sqrt{\frac{d_i}{2\pi}} e^{\left(\frac{-d_i(x_{ik} - z_k)^2}{2} \right)}. \quad (\text{A2})$$

$$LL(\mathbf{d}, \mathbf{z} | \mathbf{X}) = \sum_{k=1}^m \sum_{i=1}^n \left[\log \left(\sqrt{\frac{d_i}{2\pi}} \right) + \left(\frac{-d_i(x_{ik} - z_k)^2}{2} \right) \right]. \quad (\text{A3})$$

The model given above does not account for rater bias. Rater bias has long been seen as a problem in psychometric modeling (Saal, Downey, & Lahey, 1980). Neglecting to include bias in models of rater behavior can lead to distorted effect sizes (Hoyt, 2000). To account for the problem of bias, bias parameters have been incorporated into several rater models. For example, the many-faceted Rasch model (Linacre, 1989) extends the Rasch model by adding rater “severity” or bias parameters.

The continuous CCT model described in Batchelder and Romney (1989) allows for rater bias. Both additive and multiplicative rater bias are compared and contrasted in France and Batchelder (2013). Either type of bias can be added on its own or both types of bias can be added simultaneously. Additive bias can be thought of as a shift parameter and multiplicative bias can be thought of as a scaling parameter. In the context of educational testing, a grader who grades more harshly than the average will have a negative additive bias, whereas a grader who grades more leniently will have a positive additive bias. A grader with a strong central tendency and with less sensitivity to quality differences than average will have a multiplicative bias less than 1. A grader with a more spread out distribution of grades and with more sensitivity to quality differences than average will have a multiplicative bias greater than 1. Let $\mathbf{b}_A \in \mathbb{R}^n$ be an n rater $\times$ 1 vector of additive user biases and $\mathbf{b}_M \in \mathbb{R}^{n+}$ be an n rater $\times$ 1 vector of multiplicative user biases:

Axiom 2a: Random Error with Additive Bias: Let $b_{Ai} \in \mathbb{R}$ be defined as the additive bias for rater i . The rating x_{ik} can be decomposed as per Equation (A4). The associated log-likelihood function is given in Equation (A5).

$$x_{ik} = z_k + b_{Ai} + \varepsilon_{ik}. \quad (\text{A4})$$

$$LL(\mathbf{d}, \mathbf{z}, \mathbf{b}_A | \mathbf{X}) = \sum_{k=1}^m \sum_{i=1}^n \left[\log \left(\sqrt{\frac{d_i}{2\pi}} \right) + \left(\frac{-d_i(x_{ik} - b_{Ai} - z_k)^2}{2} \right) \right]. \quad (\text{A5})$$

Axiom 2b: Random Error with Multiplicative Bias: Let $b_{Mi} \in \mathbb{R}^+$ be defined as the multiplicative bias for rater i . The rating x_{ik} can be decomposed as per Equation (A6). The associated log-likelihood function is given in Equation (A7).

$$x_{ik} = z_k b_{Mi} + \varepsilon_{ik}. \quad (\text{A6})$$

$$LL(\mathbf{d}, \mathbf{z}, \mathbf{b}_M | \mathbf{X}) = \sum_{k=1}^m \sum_{i=1}^n \left[\log \left(\sqrt{\frac{d_i}{2\pi}} \right) + \left(\frac{-d_i(x_{ik} - b_{Mi}z_k)^2}{2} \right) \right]. \quad (\text{A7})$$

Axiom 2c: Random Error with Simultaneous Bias: Let $b_{Ai} \in \mathbb{R}$ and $b_{Mi} \in \mathbb{R}^+$ be defined respectively as the additive and multiplicative biases for rater i . The rating x_{ik} can be decomposed as per Equation (A8). The associated log-likelihood function is given in Equation (A9).

The formulation is a generalization of previous formulations. Fixing $b_{Ai} = 0$ for all i removes additive bias from the model and fixing $b_{Mi} = 1$ for all i removes multiplicative bias from the model.

$$x_{ik} = z_k b_{Mi} + b_{Ai} + \varepsilon_{ik}. \quad (\text{A8})$$

$$LL(\mathbf{d}, \mathbf{z}, \mathbf{b}_A, \mathbf{b}_M | \mathbf{X}) = \sum_{k=1}^m \sum_{i=1}^n \left[\log \left(\sqrt{\frac{d_i}{2\pi}} \right) + \left(\frac{-d_i(x_{ik} - b_{Ai} - b_{Mi}z_k)^2}{2} \right) \right]. \quad (\text{A9})$$

The CCT models can be fit by maximizing the log-likelihood functions. This can be achieved (Batchelder & Romney, 1989; France & Batchelder, 2013) by calculating the first-order derivatives for the log-likelihood function and then using a fixed point estimation procedure.

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) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The second author gratefully acknowledges support from research grants from the Army Research Office (ARO) and the Oak Ridge Institute for Science and Education (ORISE).

References

- Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19, 716-723.
- Anders, R., & Batchelder, W. H. (2012). Cultural consensus theory for multiple consensus truths. *Journal of Mathematical Psychology*, 56, 452-469.
- Anders, R., & Batchelder, W. H. (2013). Cultural consensus theory for the ordinal data case. *Psychometrika*. Advance online publication. doi:10.1007/s11336-013-9382-9
- Barnhart, H. X., Haber, M. J., & Lin, L. I. (2007). An overview on assessing agreement with continuous measurements. *Journal of Biopharmaceutical Statistics*, 17, 529-569.
- Batchelder, W. H., & Anders, R. (2012). Cultural consensus theory: Comparing different concepts of cultural truth. *Journal of Mathematical Psychology*, 56, 316-332.
- Batchelder, W. H., Kumbasar, E., & Boyd, J. P. (1997). Consensus analysis of three-way social network data. *Journal of Mathematical Sociology*, 22, 29-58.
- Batchelder, W. H., & Romney, A. K. (1986). The statistical analysis of a general Condorcet model for dichotomous choice situations. In B. Grofman & G. Owen (Eds.), *Information pooling and group decision making: Proceedings of the second University of California, Irvine conference on political economy* (pp. 103-112). Greenwich, CT: JAI Press.
- Batchelder, W. H., & Romney, A. (1988). Test theory without an answer key. *Psychometrika*, 53, 71-92.
- Batchelder, W. H., & Romney, A. K. (1989). New results in test theory without an answer key. In E. E. Roskam (Ed.), *Mathematical psychology in progress* (pp. 229-248). Heidelberg, Germany: Springer-Verlag.
- Batchelder, W. H., Strashny, A., & Romney, A. K. (2010). Cultural consensus theory: Aggregating continuous responses in a finite interval. In S.-K. Chai, J. J. Salerno, & P. L. Mabrey (Eds.), *Social computing, behavioral modeling, and prediction 2010* (pp. 98-107). New York, NY: Springer-Verlag.
- Bazaraa, M. S., Sherali, H. D., & Shetty, C. M. (2006). *Nonlinear programming: Theory and algorithms* (3rd ed.). Hoboken, NJ: Wiley.
- Berry, K. J., & Mielke, P. W. (1988). A generalization of Cohen's kappa agreement measure to interval measurement and multiple raters. *Educational and Psychological Measurement*, 48, 921-933.
- Braun, H. I. (1988). Understanding scoring reliability: Experiments in calibrating essay readers. *Journal of Educational and Behavioral Statistics*, 13, 1-18.
- Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20, 37-46.
- de Gruijter, D. N. M. (1984). Two simple models for rater effects. *Applied Psychological Measurement*, 8, 213-218.
- Di Eugenio, B., & Glass, M. (2004). The kappa statistic: A second look. *Computational Linguistics*, 30, 95-101.
- Fischer, G. H., & Molenaar, I. W. (Eds.). (1995). *Rasch models: Foundations, recent developments and applications*. New York, NY: Springer-Verlag.
- France, S. L. (2013). *MLCCT Software* (2013). Retrieved from <https://sites.google.com/site/psychminegroup/>
- France, S. L., & Batchelder, W. H. (2013). *Unsupervised consensus analysis for on-line review and questionnaire data* (Working paper). Unpublished manuscript. University of Wisconsin-Milwaukee.

- Greenacre, M., & Blasius, J. (2006). *Multiple correspondence analysis and related methods* (1st ed.). Boca Raton, FL: Chapman & Hall.
- Hoyt, W. T. (2000). Rater bias in psychological research: When is it a problem and what can we do about it? *Psychological Methods*, 5(3), 64-86.
- Janson, H., & Olsson, U. (2001). A measure of agreement for interval or nominal multivariate observations. *Educational and Psychological Measurement*, 61, 277-289.
- Kaggle, Inc. (2012). *The Hewlett Foundation: Automated essay scoring*. Retrieved from <http://www.kaggle.com/c/asap-aes>
- Karabatsos, G., & Batchelder, W. (2003). Markov chain estimation for test theory without an answer key. *Psychometrika*, 68, 373-389.
- Kennedy, R., Riquier, C., & Sharp, B. (1996). Practical applications of correspondence analysis to categorical data in market research. *Journal of Targeting, Measurement, and Analysis for Marketing*, 5, 56-70.
- Kruskal, J. B. (1964). Nonmetric multidimensional scaling: A numerical method. *Psychometrika*, 29, 115-129.
- Lamm, B., & Keller, H. (2007). Understanding cultural models of parenting: The role of intracultural variation and response style. *Journal of Cross-Cultural Psychology*, 38, 50-57.
- Linacre, J. M. (1989). *Many-faceted Rasch measurement* (2nd ed.). Chicago, IL: MESA Press.
- Moskal, B. M., & Leydens, J. A. (2000). Scoring rubric development: Validity and reliability. *Practical Assessment, Research & Evaluation*, 7(10), 71-81.
- Müller, H. (1987). A Rasch model for continuous ratings. *Psychometrika*, 52, 165-181.
- Oravecz, Z., Anders, R., & Batchelder, W. H. (2013). Hierarchical Bayesian modeling for test theory without and answer key. *Psychometrika*. Advance online publication. doi:10.1007/s11336-013-9379-4
- Oravecz, Z., Faust, K., & Batchelder, W. H. (in press). An extended cultural consensus theory model to account for cognitive processes in decision making in social surveys. *Sociological Methodology*.
- Raykov, T., & Marcoulides, G. A. (2011). *Introduction to psychometric theory* (1st ed.). New York, NY: Taylor & Francis.
- Raymond, M. R., & Viswesvaran, C. (1993). Least squares models to correct for rater effects in performance assessment. *Journal of Educational Measurement*, 30, 253-268.
- Revelle, W. (2013). *psych: Procedures for Personality and Psychological Research* (Version = 1.3.10). Evanston, IL: Northwestern University. Retrieved from <http://CRAN.R-project.org/package=psych>
- Romney, A. K., Weller, S. C., & Batchelder, W. H. (1986). Culture as consensus: A theory of culture and informant accuracy. *American Anthropologist*, 88, 313-338.
- Saal, F. E., Downey, R. G., & Lahey, M. A. (1980). Rating the ratings: Assessing the psychometric quality of rating data. *Psychological Bulletin*, 88, 413-428.
- Shavelson, R. J., Webb, N. M., & Rowley, G. L. (1989). Generalizability theory. *American Psychologist*, 44, 922-932.