Consensus Ratings: Reconceptualizing Additive Bias

Stephen L. France
University of Wisconsin – Milwaukee
Sheldon B. Lubar School of Business
P. O. Box 742
Milwaukee, WI, 53201
france@uwm.edu

ABSTRACT

Cultural consensus theory (CCT) is a technique for information fusion that aggregates ratings from multiple raters and can account for rater competency and bias. CCT has been utilized in many domains, including anthropology, sociology, education, and psychology. In this paper, we extend the continuous CCT model to account for ratings given on a fixed interval $[L, U]$. With the basic continuous CCT model, additive biases can lead to model ratings outside of the interval $[L, U]$. We introduce the concept of minimum possible error and give a set of extended axioms for continuous CCT on an interval to ensure that model ratings do not fall outside of the interval. We describe a model for linearly weighted bias and prove that this model holds for the extended axioms. We analyze a set of educational essay grading data using both the basic and item easiness continuous CCT models with linearly weighted bias and show that using linearly weighted bias fits the data much better than using scalar additive bias.

Categories and Subject Descriptors

G.3 [**PROBABILITY AND STATISTICS**]: [statistical computing, multivariate statistics]

; I.5.1 [**PATTERN RECOGNITION**]: Models—*statistical*; I.5.4 [**PATTERN RECOGNITION**]: Applications—*signal processing*

General Terms

Theory

Keywords

CCT, Evaluation, Bias, Maximum Likelihood

1. INTRODUCTION

The problem of aggregating a set of ratings for a set of items is a common one and can be found in a variety of domains including educational testing, consumer reviews, and multi-sensor data fusion. A common solution would be to use the arithmetic mean of the ratings. However, utilizing the arithmetic mean has several disadvantages. Some raters may be more competent than other raters and giving ratings from less competent raters equal weight to other ratings could introduce significant error. Raters may have inbuilt biases. For example, in educational evaluation applications, some raters may be harsher than others, and some raters may utilize a wider range of ratings or grades than others. These problems are significant, particularly when there are only a few ratings for each item.

Cultural consensus theory (CCT) [3, 6, 7, 21] is a technique for information pooling and ratings aggregation that both weights ratings by user competency and accounts for rater bias. The original application for CCT was the analysis of folk media beliefs in Guatemala [21], which is an example of simple dichotomous CCT [3, 6]. Females in a Guatemala village were given a list of diseases and were surveyed on whether they believed that each disease was "contagious or non-contagious" and whether it should be treated with a "hot" or a "cold" remedy. Both rater competencies and an overall aggregate solution or answer key were estimated for the CCT model. In CCT, the answer key measures "cultural knowledge" rather than general "scientific knowledge". For example, in the medical example, the aggregate cultural knowledge of the Guatemalan females may differ from scientific knowledge for certain diseases. The competence of a respondent measures how knowledgeable he or she is relative to this shared cultural knowledge. CCT has been widely utilized in the social and behavioral sciences. Other applications of CCT include the aggregation of on-line product reviews [13], analyzing "don't know" answers in social surveys [20], and deciding on ties in social networks [4].

In this paper, we extend and adapt the method of continuous CCT. Continuous CCT [7] has been successfully utilized on a range of data including essay rating [12] and movie review data [13]. A Bayesian variant of consensus analysis [14] has been utilized to analyze and aggregate ratings for peer grading in MOOCs (massive open on-line courses). While continuous CCT has been designed to analyze continuous data, it has been found to give good results for Likert scale data [13] when category intervals are approximately equidistant. Empirical work has shown that most Likert scale questionnaire data scales have latent continuous scales with equidistant categories [19], thus allowing the data to be treated as interval scale. Researchers in computer science [17, 22] and psychology [9, 11, 15] have found that treating Likert scale data as continuous and applying continuous/parametric statistical techniques to the data gives results that are often better than the results gained from applying ordinal/rank order techniques to the data.

Continuous CCT can model both additive bias (a shift factor) and multiplicative bias (a scaling factor). In this paper, we present a methodology for modeling additive bias when ratings data are constrained to lie in an interval $[L, U]$. The basic continuous CCT model assumes that both ratings and aggregate ratings are measured on the range $[-\infty, \infty]$.

However, most ratings are naturally constrained. For example, most educational essays or assignments are rated either on an ordinal (e.g., letter grading A-F) or a near continuous (e.g, 1-100) grading scale. Similarly, on-line reviews are usually collected on a Likert scale, either for a single attribute or for multiple attributes.

While continuous CCT has been successful applied to constrained ratings data and experiments have shown that adding additive bias to the basic continuous CCT model significantly increases model fit [12, 13], the combination of either a consensus rating close to U and a user with high positive bias or consensus rating close to L and a user with high negative bias, leads to expected scores that are outside of $[L, U]$. This invariably leads to error in the consensus model. In this paper, we develop a model for bias-adjusted consensus analysis on constrained ratings intervals. We introduce the concept of the minimum possible error, which is the minimum model error for a user with a given additive bias b_{Ai} and a score z_k where $|b_{Ai}| \leq U - L$ and $L \leq z_k \leq U$. We extend several of the axioms for continuous CCT to constrained ratings intervals to ensure that the minimum possible error is always 0 by allowing biases to be weighted by the values of z_k. We introduce a linearly weighted bias scheme. We show that the extended axioms hold for this scheme and describe a maximum-likelihood estimation procedure for fitting this scheme. We show that for a set of essay rating data, using this scheme gives a much better model fit than using a simple scalar additive bias.

2. CONSENSUS MODEL AND AXIOMS

The theory for continuous CCT is built up using a series of axioms or assumptions. Let $\mathbf{X}$ be an n rater $\times$ k item matrix of ratings, where x_{ik} is the rating given by rater i for item k, $\mathbf{d}$ be an n rater $\times$ 1 column vector of rater competencies, where d_i is the competency for rater i, and $\mathbf{z}$ be a $1 \times k$ item row vector of consensus aggregate ratings, where z_k is the consensus rating for item k. In consensus analysis terminology, $\mathbf{z}$ is often referred to as the answer key. In some work, for example [14], user competency is referred to as user precision. The basic axioms for continuous CCT are given below.

AXIOM 1. *Common Truth. For a single culture, there is a fixed answer key* $\mathbf{z}$*, where* $z_k \in (-\infty, \infty)$ *is the correct answer to question* k*.*

AXIOM 2. *Random Error. Let* x_{ik} *be the answer given by user* i *for item* k*, where* $x_{ik} \in (-\infty, \infty)$*. The answer is decomposed in* (1) *as the sum of the answer key value and random error.*

$$x_{ik} = z_k + \varepsilon_{ik}, \tag{1}$$

where the error values ϵ_{ik} *are normally distributed with expected value* $E(\epsilon_{ik}) = 0$*.*

AXIOM 3. *Local Independence. The values of* ϵ_{ik} *are mutually stochastically independent.*

AXIOM 4. *Inhomogeneous Variances. Let* $d_i \in (0, \infty)$ *be the competence or precision for user* i *and let* $d_i^{-1} = \sigma^2(\varepsilon_{ik})$*, where* $\sigma(\varepsilon_{ik})$ *is the error standard deviation.*

Given the Gaussian error in Axiom 2 and the independence of errors in Axiom 3, the likelihood function for the model parameters is given in (2).

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

Axiom 2 can be extended to account for both additive bias and multiplicative bias. Let $\mathbf{b}_A$ be an n rater $\times$ 1 column vector of additive biases, where b_{Ai} is the additive bias for rater i and $\mathbf{b}_M$ be an n rater $\times$ 1 column vector of multiplicative biases, where b_{Mi} is the multiplicative bias for rater i. If additive bias is included then (1) can be rewritten as (3). If multiplicative bias is included then (1) can be rewritten as (4). If both additive and multiplicative biases are included simultaneously then (1) can be rewritten as (5)

$$x_{ik} = z_k + b_{Ai} + \varepsilon_{ik} \tag{3}$$

$$x_{ik} = b_{Mi} z_k + \varepsilon_{ik} \tag{4}$$

$$x_{ik} = b_{Mi} z_k + b_{Ai} + \varepsilon_{ik} \tag{5}$$

With bias, (2) can be rewritten as (6). Setting $b_{Ai} = 0$ for all i gives a model without additive bias and setting $b_{Mi} = 1$ for all i gives a model without multiplicative bias. The parameters can be estimated by maximizing the log-likelihood function, which is given in (7).

$$L(\mathbf{d}, \mathbf{z}, \mathbf{b}_A, \mathbf{b}_M|\mathbf{X}) = \prod_{k=1}^{m} \prod_{i=1}^{n} \sqrt{\frac{d_i}{2\pi}} e^{\left(\frac{-d_i(x_{ik}-b_{Ai}-b_{Mi}z_k)^2}{2}\right)} \tag{6}$$

$$LL(\mathbf{d}, \mathbf{z}, \mathbf{b}_A, \mathbf{b}_M|\mathbf{X}) = \sum_{k=1}^{m} \sum_{i=1}^{n} \left[\begin{array}{l} \log\left(\sqrt{\frac{d_i}{2\pi}}\right) + \\ \left(\frac{-d_i(x_{ik}-b_{Ai}-b_{Mi}z_k)^2}{2}\right) \end{array} \right] \tag{7}$$

The use of bias helps minimize the inner squared term $(x_{ik} - b_{Ai} - b_{Mi}z_k)$, which is negative in the log-likelihood function. For a given rater i, additive bias can be thought of as a shift parameter and multiplicative bias can be thought of as a stretch parameter. For an educational evaluation application, a rater with a high additive bias would be considered to be an "easy" grader, while a rater with a low additive bias would be considered to be a "hard" grader. Likewise, a rater with an multiplicative bias greater than 1 would have a more spread out or "stretched" grading distribution than average, while a rater with an multiplicative bias less than 1 would have a more condensed or "narrow" grading distribution than average.

In ratings applications, some items are harder to rate than others. In [12], multiplicative and additive easiness parameters are added to the basic consensus model to account for the easiness/difficulty of rating items. The item easiness models were tested empirically on a set of essay ratings data. The data were gathered from raters who each graded 50 essays on 6 different attributes. In particular, the multiplicative item easiness model gave good results and the model parameters were found to have strong face validity.

The multiplicative item easiness model is defined by altering Axiom 4 as follows:

AXIOM 4a. *Multiplicative Item Easiness: For each rater* i*, define the rater competence* $d_i \in \mathbb{R}^+$ *and for each item* k*,*

define a scaling factor $\beta_k \in \mathbb{R}^+$*, so that the inverse error variance* $\sigma_{ik}^2(\varepsilon_{ik}) = (d_i\beta_k)^{-1}$.

The resulting log-likelihood function is defined in (8). Bias can be incorporated into the model by adding bias parameters to the inner squared term, as per (7). In [12], it is shown that the precision can be scaled between 0 and 1 so that the transformed multiplicative model parameters are equivalent to parameters from a Rasch model with user score p_i and item difficulty q_k, where $log(d_i) = p_i$ for all i and $log(\beta_k) = -q_k$ for all k.

$$LL(\mathbf{d}, \mathbf{z}, \boldsymbol{\beta} | \mathbf{X}) = \sum_{k=1}^{m}\sum_{i=1}^{n}\left[\begin{array}{l}\log\left(\sqrt{\frac{\beta_k d_i}{2\pi}}\right) \\ +\left(\frac{-\beta_k d_i (x_{ik}-z_k)^2}{2}\right)\end{array}\right] \quad (8)$$

3. BIAS RELATED EXTENSIONS

Empirical work has shown that the basic and bias adjusted models perform well on a range of ratings data. In fact, on the previously described essay ratings data, adding either additive or multiplicative bias to the consensus model, increased the value of the Akaike Information Criterion (AIC) [1], which is a measure of the overall log-likelihood, adjusted for the number of parameters. Adding additive bias gave a slightly better result than adding multiplicative bias and adding both types of bias simultaneously gave a further improvement in the value of the AIC. These results are validated by experiments on generated data described in [13], which show that adding bias to consensus models significantly increases model fit. However, if additive bias parameters are utilized for ratings given on a bounded ratings scale, then this can lead to problems with model fit, which are described below.

Most ratings data are constrained; i.e., there are minimum and maximum possible values of the ratings. For example, the IMDB.com movie website allows users to rate movies on a scale of 1-10. Some movies have ratings towards the top of the range and others have ratings towards the bottom of the range. For example, "The Godfather" has a rating of 9.1/10, while "Superbabies: Baby Geniuses 2" has a rating of 1.9/10. In a similar fashion, in the essay grading example described previously, when using the additive scores, the overall rating scale ranged from 6 to 36. The "I hate computers" essay had a consensus score of 6, while several of the essays had scores close to 36. Additive bias is modeled as a simple constant applied to a rater's score and is independent of the item being rated. In fact, an examination of the residuals for the essay grading example showed that much of the overall model error was due to cases either where a rater had a high positive bias and the item had a consensus rating close to U or the rater had a high negative bias and the item had an overall consensus score close to L. For item k, each rating is in the interval $[L, U]$, so that $L \leq x_{ik} \leq U$. In the basic model, though the input ratings are constrained within a fixed interval, there is no assumption of a fixed interval range within the model. However, for many consensus applications, a fixed range is also assumed for the answer key, so that $L \leq z_k \leq U$. For example, consider an essay that is graded on a scale of 1 to 10. Even if there is a situation where an essay has a consensus score of 10/10 and several raters have positive biases, the essay cannot be given a score greater than 10.

The minimum possible error (MPE) for rater i and item k defined as the minimum possible ϵ_{ik} for additive bias $|b_{Ai}| \leq U - L$ and answer value $L \leq z_k \leq U$. If $b_{Ai} > 0$ then $MPE_{ik} = z_k + b_{Ai} - U$. If $b_{Ai} < 0$ then $MPE_{ik} = L - z_k - b_{Ai}$. Thus, for MPE_{ik} to be 0 then the conditions in (9) must hold.

$$MPE_{ik} = 0 \rightarrow \left\{\begin{array}{l} U \geq z_k + b_{Ai},\ if\ b_{Ai} > 0 \\ L \leq z_k + b_{Ai},\ if\ b_{Ai} < 0 \end{array}\right. \quad (9)$$

To illustrate the concept of minimum possible error, consider a rating scale of 1-10 for essays and a rater i who rates essays rather harshly. This user has a negative bias of 3. A student produces a very poor essay, which has an aggregate z_k of 1.5 out of 10. Now, $z_k + b_{Ai} = -1.5$, but the minimum possible rating is 1. Thus, $MPE_{ik} = 1 - (-1.5) = 2.5$ and as as $x_{ij} \geq 1$,$\epsilon_{ik} \geq 2.5$. The weighted additive bias model introduced in this paper, replaces the simple scaler additive bias b_{Ai} with a weighted bias $w_k \times b_{Ai}$, where w_k depends on z_k and w_k is set to ensure that $MPE_{ik} = 0$ for all i, k.

3.1 Extended Model and Actions

A model for continuous CCT on a closed interval $[L, U]$ with additive bias and $MPE_{ik} = 0$ for all i,k is defined by a modified version of the CCT axioms. The modified versions of Axioms 1 and 2 are given below. Axioms 3 and 4 are unchanged from the basic model.

AXIOM 1. *Common Truth. For a single culture, there is a fixed answer key* $\mathbf{z}$*, where* $z_k \in [L, U]$ *is the correct answer to question* k.

AXIOM 2. *Random Error. Let* x_{ik} *be the answer given by user* i *for item* k*, where* $x_{ik} \in [L, U]$*. The answer is decomposed in* (10) *as the sum of the answer key value, a weighted bias, and random error.*

$$x_{ik} = z_k + w_k b_{Ai} + \varepsilon_{ik}, \quad (10)$$

where the error values ϵ_{ik} *are normally distributed with expected value* $E(\epsilon_{ik}) = 0$ *and* w_{ik} *is defined so that the following conditions hold:*

I The minimum error MPE_{ik} *should be equal to 0, i.e.,* (11) *must hold.*

$$L \leq w_k b_{Ai} + z_k \leq U \quad (11)$$

II Consider any other item $k^* \neq k$.

i If $b_{Ai} < 0$ *and* $z_k^* \leq z_k$ *then* $w_k^* \leq w_k$.
ii If $b_{Ai} < 0$ *and* $z_k^* \geq z_k$ *then* $w_k^* \geq w_k$.
iii If $b_{Ai} > 0$ *and* $z_k^* \leq z_k$ *then* $w_k^* \geq w_k$.
iv If $b_{Ai} > 0$ *and* $z_k^* \geq z_k$ *then* $w_k^* \leq w_k$.
v If $b_{Ai} = 0$ *then* w_k *is undefined.*

The first condition ensures that the MPE is always 0. The second condition ensures that w_{ik} is monotone decreasing with the distance of z_k from each range boundary, thus axiomatizing the previously described idea that bias should decrease as the distance from the boundaries increases. The $b_{Ai} > 0$ and $b_{Ai} < 0$ cases are treated separately, as $b_{Ai} > 0$ affects the upper boundary of the interval and $b_{Ai} < 0$ affects the lower boundary of the interval. Given the revised axioms, the likelihood function with additive bias can be written as (12) and the definition of MPE is given in (13).

$$L(\mathbf{d}, \mathbf{z}, \mathbf{b}_A|\mathbf{X}) = \prod_{k=1}^{m}\prod_{i=1}^{n}\sqrt{\frac{d_i}{2\pi}}e^{\left(\frac{-d_i(x_{ik}-w_k b_{Ai}-z_k)^2}{2}\right)} \quad (12)$$

$$MPE = 0 \rightarrow \begin{cases} U \geq z_k + w_k b_{Ai},\ if\, b_{Ai} > 0 \\ L \leq z_k + w_k b_{Ai},\ if\, b_{Ai} < 0 \end{cases} \quad (13)$$

3.2 A Linear Weighted Bias Model

The two conditions given for Axiom 2 are not sufficient to define the actual values of the bias weights. In this section, we introduce a simple linear weighted bias model, where the weighted biases decline linearly to 0 as the value of z_k approaches the rating interval boundaries. The ratings are defined more formally in (14). The weighting is essentially the proportion of the interval from z_k to the boundary of the interval in the direction in which the bias could produce a value of $MPE_{ik} > 0$. E.g, in the previous example the harsh grader had a bias of -3, which combined with a low value of z_k, with $z_k - L < 3$, would lead to $MPE_{ik} > 0$.

$$w_k = \begin{cases} \frac{U-z_k}{U-L},\ if\, b_i > 0 \\ \frac{z_k-L}{U-L},\ if\, b_i < 0 \end{cases} \quad (14)$$

A visual example to help explain the weighting scheme is given in Figure 1. Here, the value of z_k is plotted on the horizontal axis and the overall weighted bias $b_i \times z_k$ is plotted on the vertical axis. Separate plots are given for bias values of 4 and -3. One can see that for negative biases, the weighted bias tends to 0 as z_k tends to L, and for positive biases, the weighted bias tends to 0 as z_k tends to U.

THEOREM 3.1. *The two conditions for Axiom 2, the random error axiom, hold for the linear weighted bias model, for all $|b_{Ai}| \leq U - L$ and $L \leq z_k \leq U$.*

PROOF. The first condition states that $MPE_{ik} = 0$ for all, i, k. Assume that there is some $|b_{Ai}| \leq U - L$ and some $L \leq z_k \leq U$ for which $MPE_{ik} > 0$. From (13), either i) $b_{Ai} > 0$ and $U < z_k + w_k b_{Ai}$ or ii) $b_{Ai} < 0$ and $L > z_k + w_k b_{Ai}$.

Take case i).

$$U < z_k + w_k b_{Ai} = z_k + \left(\frac{U - z_k}{U - L}\right) b_{Ai} \quad (15)$$

Subtracting Z_k from both sides of the equation and dividing both sides by $U - z_k$ gives (16).

$$1 < \frac{b_{Ai}}{U - L} \quad (16)$$

But, $b_{Ai} > 0$ and $|b_{Ai}| \leq U - L$, so the RHS of (16) is less than 1, which gives a contradiction.

Take case ii).

$$L > z_k + w_k b_{Ai} = z_k + \left(\frac{z_k - L}{U - L}\right) b_{Ai} \quad (17)$$

Subracting z_k from both sides of the equation and dividing both sides by $z_k - L$ gives (18).

$$-1 > \frac{b_{Ai}}{U - L} \quad (18)$$

But $b_{Ai} < 0$ and $|b_{Ai}| \leq U - L$, so the RHS of (18) is greater than -1, which also gives a contradiction, giving an overall proof by contradiction for both possible cases.

As the weighting function is linear, the monotone property is obvious. From (14), if $b_{Ai} > 0$ then z_k is negative in w_k, so w_k is monotone decreasing with z_k and if $b_{Ai} < 0$ then z_k is positive in w_k, so w_k is monotone increasing with z_k. □

3.3 Model Estimation

In [7, 12], i) fixed point optimization, ii) gradient based optimization, and iii) derivative free optimization procedures are implemented to estimate both the basic and item easiness models, with or without bias. All three procedures are implemented in MATLAB in a software package described in [14]. The gradient based optimization and derivative free optimization procedures are implemented using the standard MATLAB optimization procedures. The fixed point procedure is implemented by calculating each vector of parameters ($\mathbf{z}$, $\mathbf{d}$, $\mathbf{b}_A$, etc.) from first order conditions until the parameter values converge. On the whole, the fixed point procedure is up to 100 times faster than the gradient based procedure, but it is unable to estimate the additive item difficulty models. Given the large number of parameters, CCT models can have parameter identifiability issues. In [12], a method is described to mitigate this problem. First, an initial model with a fully identified subset of the parameters is fit to the data. Then, either a single parameter value is fixed or a set of parameter values are fixed in order to fully identify the model. The full model is then estimated with the fixed parameter values taken from the initial subset model. If a single parameter of a given type is fixed then the optimized model has the same likelihood as the optimized model with no fixed parameters. If all parameters of a given type are fixed then a parameter fitting hierarchy is introduced, with priority given to the fixed variables fit from the subset model. The user can supply 'a vector of fitting constants $[Fixz, Fixd, Fixb_A, Fixb_M]$, where each element of the vector can be 0, -1, or 1. A value of 0 indicates that no element is fixed, 1 indicates a partial fix and -1 indicates a full fix. We implemented all three optimization techniques for the linear weighted bias model. First order conditions for the fixed point optimization procedure are given in the Appendix.

3.4 Essay Grading

In data mining terminology, consensus analysis is an "unsupervised model", so cannot be tested in terms of direct feature recovery or prediction. However, there are several methods of testing the performance of unsupervised techniques. The techniques described in this paper are fit using a maximum likelihood methodology. A measure of statistical fit for the models is given by the log-likelihood (LL). As more parameters are added to a maximum likelihood model, so the model freedom increases and the log-likelihood decreases. Thus, for model comparison purposes, an information measure adjusted for the number of parameters, such as the Aikaike information criterion (AIC) [1], which is defined as $AIC = 2 \times No.Parameters - 2 \times LL$, can provide an overall measure of model fit. As the log-likelihood is negative in the AIC equation, the lower the value of the AIC, the better the model fit. In this example, we take a set of essay grading data described in [12] and analyze the model fit for continuous CCT models with no bias, with simple additive bias, and with weighted additive bias.

In [12], 14 graders each rated 50 essays from the "Hewlett Foundation: Automated essay scoring" dataset [18] on six

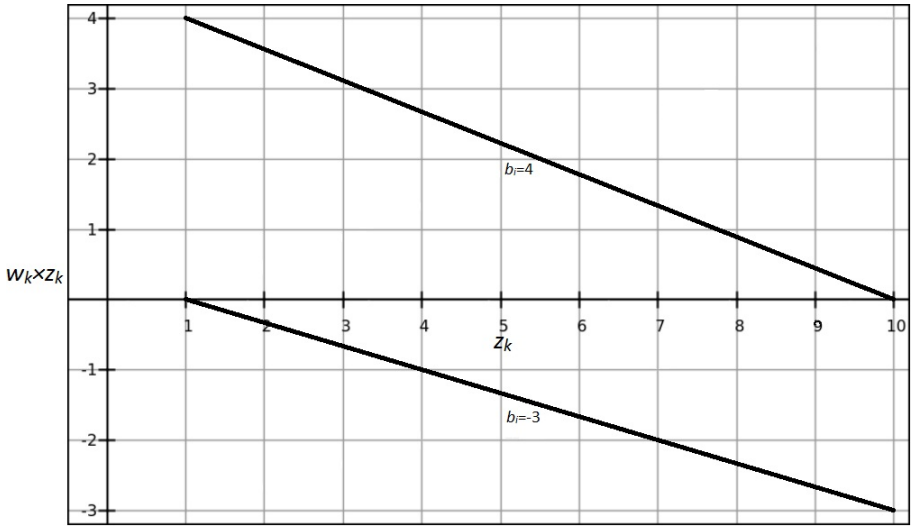

Figure 1: Linear Weighting Scheme Example

different attributes. The rated essays were written by school students for an essay prompt on the subject of laughter. The 14 graders included two expert graders and 12 student graders, who were trained using a set of materials that described the grading rubric for essays along with several graded example essays. The six attributes were "Ideas and Content", "Organization", "Voice", "Word Choice", "Sentence Fluency", and "Conventions". In [12], the attribute scores were either added together to give an assumed (by classical test theory) scale or were aggregated into a continuous scale using correspondence analysis [16]. For this experiment, we utilize the additive summated scale score.

When comparing the simple additive bias model with the weighted additive bias model, both the distribution of the scores and the overall scale of the scores are important. For example, based on the model mechanics, it is likely that the weighted additive bias model has relatively stronger performance than the simple additive bias model when more scores are closer to the scale boundary. For the educational data, the histograms of the scores for the individual attributes are given in Figure 2 and the histogram for the overall summed score is given in Figure 3. In each histogram, the mean value is shown as a dashed line. The majority of the scores are close to the center, indicating a distribution that is closer to the normal distribution than to the uniform distribution, but there are significant attribute scores and overall scores at the edge of the boundaries. The attributes scales are highly correlated, with inter-attribute correlations ranging from $\rho = 0.8058$ to $\rho = 0.9121$. This suggests that there are very few essays that have high scores in some attributes but low scores in others. This is substantiated by the overall score histogram, which has a slightly more central distribution than the individual histograms, but has thick tails, with essays that score low or high on most attributes.

For the additive essay rating data, for both the basic and item easiness variants of the consensus model, we compared the models with no bias, simple additive bias, and weighted bias. This gave a total of six models. For the bias methods, we used the previously described "partial fix" when running the model. In order to test whether or not the difference

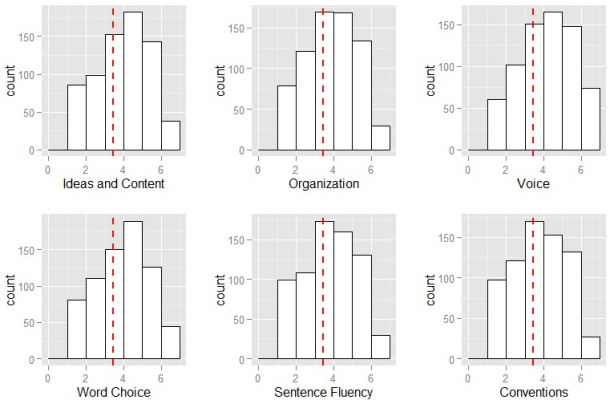

Figure 2: Attribute Score Histograms

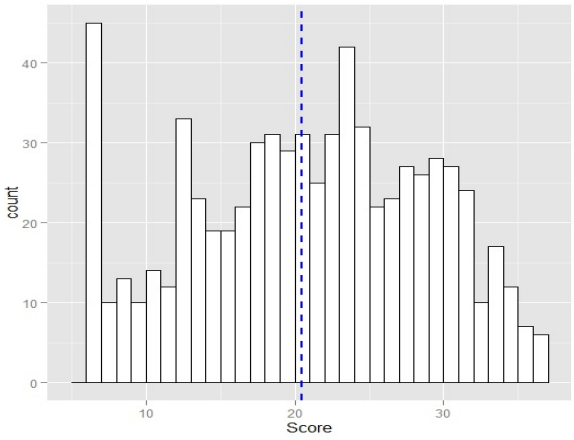

Figure 3: Overall Score Histogram

in performance between the weighted additive bias and the simple additive bias was due to size and granularity of the

scale, we created summated scales for every possible combination of attributes (including single attribute scales). For the six attributes this gave a total of $2^6 - 1 = 63$ possible scales. We optimized the models and calculated AIC values from the optimal log-likelihood values. The results, summarized by the number of attributes, are given for the basic model in Table 1 and for the item easiness model in Table 2.

Table 1: AIC values for basic models

No. Att	No Bias	Simple Bias	Weighted Bias
1	1748.3835	1606.3782	1594.8557
2	2564.1290	2403.0307	2388.3201
3	3069.9430	2900.0774	2883.6062
4	3439.5761	3264.5649	3246.7570
5	3731.4040	3553.0088	3533.9920
6	3972.6799	3791.8770	3771.9457

Table 2: AIC values for item easiness models

No. Att	No Bias	Simple Bias	Weighted Bias
1	1762.4229	1633.3617	1623.4503
2	2558.8114	2420.3363	2405.7195
3	3053.2057	2912.0755	2895.5383
4	3414.7207	3273.3753	3255.7925
5	3700.3412	3559.6947	3540.9898
6	3936.6527	3796.9912	3777.9122

In all instances, adding simple additive bias to a model decreases the AIC and adding weighted additive bias decreases the AIC still further. This suggests that for the essay rating data, the weighted bias model is more appropriate than the simple bias model. To examine the results further, for the case with all six attributes selected, we plotted combined scatterplot/histogram graphics for both the item parameters (Figure 4) and the rater parameters (Figure 5) for the item easiness model with weighted bias. The results are similar

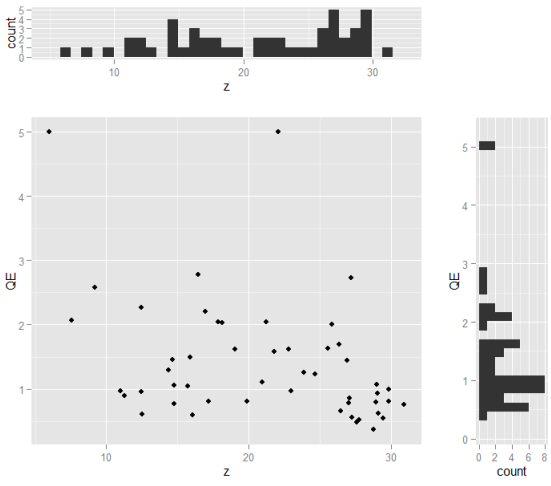

Figure 4: Item Scatterplot and Histograms

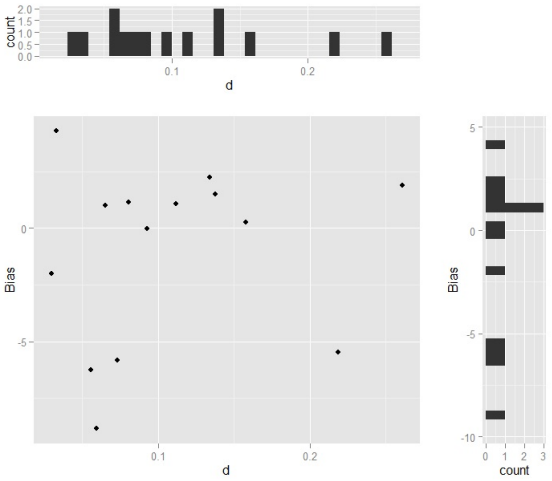

Figure 5: Rater Scatterplot and Histograms

to the results of the analysis carried out in [12]. The item answer key scores ranged from 6 to 30.91. The essay with the lowest value of 6, is the essay with the largest possible easiness value of 5. This is an essay where the student had ignored the lesson prompt and had written "I hate computers" as the entirety of the essay. This essay was graded as 6 (1 on each attribute) by every single rater. The essays with lower item easiness values were essays that were harder to grade. Overall, the results posses good face validity.

4. CONCLUSIONS AND FUTURE WORK

Utilizing continuous CCT with simple scalar bais on ratings gathered in a bounded interval $[L, U]$ can result in model estimates of x_{ik} outside of that interval. Using weighted biases that conform to the extended CCT axioms given in this paper ensures that $MPE_{ik} = 0$ for all i, k and thus that the model estimates of x_{ik} are within $[L, U]$. We have introduced a specific linear weighting model and showed that this weighting model performed well on a set of educational testing data.

The weighted bias model is defined by the CCT axioms and the conditions that the $MPE_{ik} = 0$ for all i,k and that the weighting function is monotonic with respect to z_k for both $b_{Ai} < 0$ and $b_{Ai} > 0$. In this paper, we have introduced a specific weighting model, the linear weighted bias model. However, the weighting conditions do not require that the weighting function is linear. In fact, if $|b_{Ai}| = U - L$, the graph of $w_k b_{Ai}$ plotted for z_k, as in Figure 1, defines the line for which $MPE_{ik} = 0$. Any weighting scheme that gives a graph "underneath" the $MPE_{ik} = 0$ line across z_k for all i and is monotonic with respect to z_k is permissible. CCT models have been developed where bias is conceptualized as a logit function over a $[0, 1]$ interval [8] and where bias is added to the error adjusted answer score [5]. To help determine the utility and usage scenarios for different bias models, comparative experiments could be run for different ratings scenarios.

At a slightly deeper level, an interesting topic that could be explored further is the boundary between ordinal and continuous data. For example, if an essay is graded on a 3-point like Likert scale as "poor", "average", or "good" then

most people would consider the data gathered using this scale as ordinal. However, as the number of scale points increases, the data is harder to categorize as strictly ordinal. If an essay is graded on an integer scale of 1 to 100 then the resulting data are still ordinal in terms of measurement properties as scores between the integer values are not defined. However, most people would utilize the data as continuous data. To give a real world example, on-line review data are usually gathered using an ordinal Likert like scale; for example, Amazon.com has consumers rate items from 1 to 5 stars. The data are usually considered to be continuous and the mean score is usually reported rather than the median score. Recommendation system researchers often treat these data as continuous and in a method of predicting reviews called collaborative filtering [10], continuous correlations between items and between reviewers, calculated from Likert scale review data, have been found to have more predictive value than the equivalent Spearman rank correlations [17].

The models presented in this paper assume a single culture and a single set of cultural knowledge. This may not always be the case. For example, when examining cross cultural differences, each culture may have a different set of ratings or knowledge. Several models have been developed that account for multiple cultures. A model for analyzing multiple cultures for dichotomous CCT is given in [2]. A clusterwise implementation of continuous CCT has been developed [13], which simultaneously clusters subjects into cultural groups and calculates the consensus parameters for each cultural group. The models in this paper can easily be integrated into the clusterwise CCT framework. As mentioned previously in this paper, one would expect essay/question raters to consist of a single culture defined by a set of instructions or grading rubric. However, in some situations, for example, in the scenario given earlier in the paper, more than one grading culture may develop. In this case, clusterwise CCT could be a useful framework for determining if more than one grading culture exists, which in exam grading situations is a situation that may have to be rectified.

5. REFERENCES

[1] H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, December 1974. ID: 1.

[2] R. Anders and W. H. Batchelder. Cultural consensus theory for multiple consensus truths. *Journal of Mathematical Psychology*, 56(6):452–469, 12 2012.

[3] W. Batchelder and A. Romney. Test theory without an answer key. *Psychometrika*, 53(1):71–92, 1988. 10.1007/BF02294195.

[4] W. H. Batchelder, E. Kumbasar, and J. P. Boyd. Consensus analysis of three-way social network data. *The Journal of Mathematical Sociology*, 22(1):29–58, 03/01; 2014/05 1997. doi: 10.1080/0022250X.1997.9990193; 31.

[5] W. H. Batchelder, Z. Oravecz, and R. Anders. Cultural consensus theory for continuous responses: A latent appraisal model for information pooling. *Journal of Mathematical Psychology*, In Press.

[6] W. H. Batchelder and A. K. Romney. *The statistical analysis of a general Condorcet model for dichotomous choice situations*, pages 103–112. Information Pooling and Group Decision Making: Proceedings of the Second University of California, Irvine Conference on Political Economy. JAI Press, Greenwich, CT, 1986.

[7] W. H. Batchelder and A. K. Romney. *New results in test theory without an answer key*, pages 229–248. Mathematical Psychology in Progress. Springer-Verlag, Heidelberg, Germany, 1989.

[8] W. H. Batchelder, A. Strashney, and A. K. Romney. *Cultural Consensus Theory: Aggregating Continuous Responses in a Finite Interval*, pages 98–107. Social Computing, Behavioral Modeling , and Prediction 2010. Springer-Verlag, New York, NY, 2010.

[9] K. A. Bollen and K. H. Barb. Pearson's r and coarsely categorized measures. *American Sociological Review*, 46(2):232–239, Apr. 1981.

[10] J. S. Breese, D. Heckerman, and C. Kadie. Empirical analysis of predictive algorithms for collaborative filtering. In *Proceedings of Fourteenth Conference on Uncertainty in Artificial Intelligence*, pages 43–52, Madison, WI, 1998. Morgan Kaufmann.

[11] J. Carifio and R. Perla. Resolving the 50-year debate around using and misusing Likert scales. *Medical Education*, 42(12):1150–1152, 2008.

[12] S. L. France and W. H. Batchelder. Maximum likelihood item easiness models for test theory without an answer key. *Educational and Psychological Measurement*, pages 1–30, In Press.

[13] S. L. France and W. H. Batchelder. Unsupervised consensus analysis for on-line review and questionnaire data. *Information Sciences*, pages 1–25, In Press.

[14] S. L. France, M. Vaghefi, and W. H. Batchelder. FlexCCT: Software for Continuous CCT. In S. K. D'Mello, R. A. Calvo, and A. Olney, editors, *Proceedings of the 6th International Conference on Educational Data Mining*, pages 394–395, Worcester, MA, July 6-9 2013. International Educational Data Mining Society.

[15] J. Gaito. Measurement scales and statistics: Resurgence of an old misconception. *Psychological Bulletin*, 87(3):564–567, 1980. ID: 1980-22405-001; M3: doi:10.1037/0033-2909.87.3.564.

[16] M. Greenacre and J. Blasius. *Multiple Correspondence Analysis and Related Methods*. Chapman & Hall, Boca Raton, 1 edition, 2006.

[17] J. Herlocker, J. A. Konstan, and J. Riedl. An empirical analysis of design choices in neighborhood-based collaborative filtering algorithms. *Information Retrieval*, 5(4):287–310, 2002. 10.1023/A:1020443909834.

[18] F. Hewlett. The Hewlett Foundation: Automated essay scoring, 2012.

[19] R. Kennedy, C. Riquier, and B. Sharp. Practical applications of correspondence analysis to categorical data in market research. *Journal of Targeting, Measurement, and Analysis for Marketing*, 5(1):56–70, 1996.

[20] Z. Oravecz, K. Faust, and W. H. Batchelder. An extended cultural consensus theory model to account for cognitive processes in decision making in social surveys. *Sociological Methodology*, In Press.

[21] A. K. Romney, S. C. Weller, and W. H. Batchelder. Culture as consensus: A theory of culture and informant accuracy. *American Anthropologist*,

88(2):313–338, 1986.

[22] X. Su and T. M. Khoshgoftaar. A survey of collaborative filtering techniques. *Adv.in Artif.Intell.*, 2009:4:2–4:2, January 2009.

APPENDIX

A. FIXED POINT ESTIMATION

The fixed point estimation procedure calculates each vector of variables ($\mathbf{z}$, $\mathbf{d}$, $\boldsymbol{\beta}$, and $\mathbf{b}_A$) sequentially using the first order conditions for the variables. The procedure runs until convergence, i.e., any change in the variables is less than a certain threshold (default = 10^{-6}). The log-likelihood equation is given in (19) and the weighting function is given in (20). The fixed point equations are given for each parameter set. The estimation equations are given for the multiplicative item easiness variant of the consensus model. The estimation equations for $\mathbf{z}$, $\mathbf{d}$, and $\mathbf{b}_A$ can be utilized for the basic model by replacing each β_k with 1.

$$L\left(\mathbf{d},\boldsymbol{\beta},\mathbf{z},\mathbf{b}_A|\mathbf{X}\right) = \prod_{k=1}^{m}\prod_{i=1}^{n}\sqrt{\frac{d_i\beta_k}{2\pi}}e^{\left(\frac{-d_i\beta_k\left(x_{ik}-w_k b_{Ai}-z_k\right)^2}{2}\right)} \tag{19}$$

$$w_k = \begin{cases} \frac{U-z_k}{U-L},\ if\ b_i > 0 \\ \frac{z_k-L}{U-L},\ if\ b_i < 0 \end{cases} \tag{20}$$

The estimation for d_i for all i is given below.

$$\begin{aligned}\frac{\partial f}{\partial d_i} &= \sum_{k=1}^{m}\left[\sqrt{\frac{2\pi}{d_i\beta_k}}\cdot\frac{1}{2}\sqrt{\frac{2\pi}{d_i\beta_k}}\frac{\beta_k}{2\pi} - \frac{w_k\beta_k\left(x_{ik}-w_k b_{Ai}-z_k\right)^2}{2}\right] \\ &= \sum_{k=1}^{m}\left[\frac{1}{2d_i} - \frac{\beta_k\left(x_{ik}-w_k b_{Ai}-z_k\right)^2}{2}\right]\end{aligned} \tag{21}$$

Setting $\frac{\partial f}{\partial d_i} = 0$ gives (22).

$$\sum_{k=1}^{m}\frac{1}{2d_i} = \sum_{k=1}^{m}\frac{\beta_k\left(x_{ik}-w_k b_{Ai}-z_k\right)^2}{2} \tag{22}$$

Rearranging for d_i gives (23).

$$d_i = \frac{m}{\sum_{k=1}^{m}\beta_k\left(x_{ik}-w_k b_{Ai}-z_k\right)^2} \tag{23}$$

The estimation for b_{Ai} for all i is given below.

$$\frac{\partial f}{\partial b_{Ai}} = \sum_{k=1}^{m}\left[\frac{2\cdot -d_i\beta_k\left(x_{ik}-w_k b_{Ai}-z_k\right)\cdot -w_k}{2}\right] \tag{24}$$

Setting $\frac{\partial f}{\partial b_{Ai}} = 0$ gives (25).

$$\sum_{k=1}^{m} d_i\beta_k w_k\left(x_{ik}-w_k b_{Ai}-z_k\right) = 0 \tag{25}$$

Rearranging for b_{Ai} and canceling d_i terms gives (26).

$$b_{Ai} = \frac{\sum_{k=1}^{m}\beta_k w_k\left(z_k - x_{ik}\right)}{\sum_{k=1}^{m}\beta_k w_k^2} \tag{26}$$

The estimation for β_k for all k is given below.

$$\begin{aligned}\frac{\partial f}{\partial \beta_k} &= \sum_{i=1}^{n}\left[\sqrt{\frac{2\pi}{d_i\beta_k}}\cdot\frac{1}{2}\sqrt{\frac{2\pi}{d_i\beta_k}}\frac{d_i}{2\pi} - \frac{w_k d_i\left(x_{ik}-b_{Ai}-z_k\right)^2}{2}\right] \\ &= \sum_{i=1}^{n}\left[\frac{1}{2\beta_k} - \frac{d_i\left(x_{ik}-w_k b_{Ai}-z_k\right)^2}{2}\right]\end{aligned} \tag{27}$$

Setting $\frac{\partial f}{\partial \beta_k} = 0$ and rearranging gives (28).

$$\sum_{i=1}^{n}\frac{1}{2\beta_k} = \sum_{i=1}^{n}\frac{d_i\left(x_{ik}-w_k b_{Ai}-z_k\right)^2}{2} \tag{28}$$

Rearranging for β_k gives (29).

$$\beta_k = \frac{n}{\sum_{i=1}^{n} d_i\left(x_{ik}-w_k b_{Ai}-z_k\right)^2} \tag{29}$$

The estimation for z_k for all k is given below. The value of w_k is dependent on whether $b_{Ai} > 0$, $b_{Ai} < 0$, or $b_{Ai} = 0$. If $b_{Ai} > 0$:

$$\frac{\partial f}{\partial z_k} = \sum_{i=1}^{n}\left[\frac{-d_i\beta_k\left(x_{ik}-w_k b_{Ai}-z_k\right)\cdot 2\cdot}{2}\left(-1+\frac{b_{Ai}}{U-L}\right)\right] \tag{30}$$

If $b_{Ai} < 0$:

$$\frac{\partial f}{\partial z_k} = \sum_{i=1}^{n}\left[\frac{-d_i\beta_k\left(x_{ik}-w_k b_{Ai}-z_k\right)\cdot 2\cdot}{2}\left(-1-\frac{b_{Ai}}{U-L}\right)\right] \tag{31}$$

The two cases can be combined along with the trivial $b_{Ai} = 0$ case to give (32).

$$\frac{\partial f}{\partial z_k} = \sum_{i=1}^{n}\left[-d_i\beta_k\left(x_{ik}-w_k b_{Ai}-z_k\right)\left(-1+\frac{|b_{Ai}|}{U-L}\right)\right] \tag{32}$$

Setting $\frac{\partial f}{\partial z_k} = 0$ and rearranging gives (33).

$$\begin{aligned}&\sum_{i=1}^{n} z_k d_i\beta_k\left(-1+\frac{|b_{Ai}|}{U-L}\right) \\ =&\sum_{i=1}^{n} d_i\beta_k\left(x_{ik}-w_k b_{Ai}\right)\left(-1+\frac{|b_{Ai}|}{U-L}\right)\end{aligned} \tag{33}$$

Rearranging for z_k gives (34).

$$z_k = \frac{\sum_{i=1}^{n} d_i\left(x_{ik}-w_k b_{Ai}\right)\left(-1+\frac{|b_{Ai}|}{U-L}\right)}{\sum_{i=1}^{n} d_i\left(-1+\frac{|b_{Ai}|}{U-L}\right)} \tag{34}$$