Improving the value of ratings by factoring out the generosity of raters

Douglas L. T. Rohde
Department of Brain and Cognitive Science
Massachusetts Institute of Technology

May 29, 2003

Abstract

Gathering human ratings is one of our most common research methods. However, there are potential sources of bias or noise in human ratings that can adversely affect the clarity and consistency of the results, especially when a relatively small amount of data is available. One primary source of disagreement between raters is their varying generosity—some raters tend to give higher ratings than others. This paper evaluates a typical method for dealing with this problem, the use of z-scores, and introduces several new techniques, most notably the Spindle model, for discovering and factoring out reviewer generosity to improve the usefulness of ratings data.

1 Introduction

One of the most common data gathering methods used in the social sciences, and in society at large, is the collection of human ratings. We ask people to rate movies and restaurants, colleges, consumer products, conference papers, job applications, SAT essays, and mutual funds, in addition to experimental items of many sorts. The ultimate goal in collecting these ratings is usually to assign each item a score with which they can be compared. The most common and straight-forward method for producing such scores is simply averaging the ratings received by each item.

However, it is often the case that each item is rated by only a subset of the reviewers. If the reviewers differ in the way they apportion their ratings scale, such that some are more generous than others, simply averaging an item's ratings may result in a biased score. If an item is only reviewed by a small number of reviewers and some of them happen to be sticklers, the item may receive a lower average rating than it deserves. In theory, this bias could be greatly reduced by estimating how generous each of the reviewers is, relative to the other reviewers, and adjusting his or her ratings accordingly.

This paper introduces and evaluates several methods for dealing with the problem of reviewer generosity differences. Principally, it advocates a novel technique, known as the Spindle model, which is designed to discover and factor out reviewer generosity. Unlike other methods, the Spindle model is non-linear and is more appropriate for the predominant situation in which ratings are confined to a finite scale.

2 The Z-Score method

The traditional method for factoring out reviewer generosity is to convert the raw ratings to z-scores. Z-Scores are also useful for comparing ratings based on different scales and they are being used increasingly for the purpose of ranking sports teams and financial products. The model is quite simple. The mean and standard deviation of the ratings are computed for each reviewer. The raw ratings are then converted to z-scores by subtracting the reviewer's mean rating and dividing by the standard deviation. An item's z-scores are then averaged to produce its overall score.

In practice, the use of z-scores can be an improvement over averaged ratings, but this method also has the potential to introduce some new sources of bias. The Z-Score method does not involve any direct or indirect comparison between reviewers. As a result, the adjusted scores assigned by a reviewer can be skewed by a biased sampling of items rated by that reviewer. For example, if the reviewer rated mostly above average items, then an item that is truly of average score would receive a negative z-score from that reviewer.

A reviewer's z-scores can also be affected by the variance of the items he or she rates or the variance of the ratings assigned. If a reviewer tends to have very low variance, assigning most ratings close to the mean, then

[*] In preparation, do not cite.

any outlying items will receive exceptionally low or high z-scores because of the low standard deviation. As a result, this reviewer will have more influence over the mean z-score of such items than would another reviewer who distributes his or her ratings more evenly over the whole scale.

These problems result from differences in mean or variance in the quality of items assigned to different reviewers. Such biased sampling in assigning items may not be a problem in well-designed ratings studies. However, they can result by chance when reviewers rate only a small sample and they can arise systematically if reviewers are self-selecting the items they review, as in most voluntary reviews of movies, books, and other consumer products.

To avoid such problems, we will need to introduce some form of comparison between reviewers. If reviewer A and reviewer B both evaluate several items in common, and reviewer A tends to give them higher ratings, we could reasonably conclude that A is more generous than B. We might then, in the future, wish to discount A's ratings, or elevate B's. Even if A and B did not rate any of the same items in common, we could estimate their relative generosity if they both rated different sets of items in common with reviewer C. Thus, the comparison between reviewers need not be direct.

Our ideal model will learn the generosities of the various reviewers as it learns the adjusted scores of the items. Because the observed ratings result from the interaction of these scores and generosities, in addition to other factors, these two measures are mutually dependent. We will therefore attempt to learn them both simultaneously. The next two sections introduce several multivariate models intended to smooth ratings data by modeling and factoring out reviewer generosity.

3 A linear model

In order to derive scores and generosities on the basis of ratings data, we must begin by specifying a model of how these two factors interact in the assignment of a rating. Let us assume that each of the items under review, i, has an intrinsic value, which we'll call its *score*, and denote s_i. The primary goal of the process is to determine the true scores of the items given the set of ratings and the identities of the reviewers.

The model assumes that a reviewer is able to accurately evaluate the true score of any item, but that the reviewer must translate that score into a rating using the given scale. A more generous reviewer will map the same item to a higher rating than will a less generous reviewer. The following is a very simple linear model of ratings assignment:

$$r_{iv} = s_i + g_v \tag{1}$$

According to this model, the rating, r_{iv}, assigned by reviewer v to item i is predicted to be the sum of two terms. One is the actual quality, or score, of the item, s_i, and the other is the reviewer's generosity, g_v.

Given a set of ratings, the model will attempt to simultaneously find the set of reviewer generosities and the set of item scores that minimize the sum-squared error between the predicted and actual ratings. If $\hat{r}_{iv}$ is the actual rating assigned to item i by reviewer v, the overall ratings error is given by the following formula:

$$E_R = \sum_{i,v} (r_{iv} - \hat{r}_{iv})^2 \tag{2}$$

Although a closed-form solution may be possible, it is simple and effective to adjust the generosities and scores to minimize this error using an iterative process. We begin by assuming that all reviewers have a neutral generosity of 0 and find the locally optimal values for each of the scores, s_i. These are obtained by differentiating Equation 2 with respect to s_i, setting this partial derivative to zero, and solving for s_i. The result is shown in Equation 3:

$$s_i = \frac{\sum_{v \in V_i} \hat{r}_{iv} - g_v}{\mid V_i \mid} \tag{3}$$

where V_i is the set of reviewers who rated item i. Thus, an item's score is estimated to be the average difference between its actual ratings and the generosity of the reviewers, and the initial scores will simply be the average ratings.

The next step of the iteration is to re-estimate the generosities of the reviewers given the new scores of the items. This is done with the corresponding equation:

$$g_v = \frac{\sum_{i \in I_v} \hat{r}_{iv} - s_i}{\mid I_v \mid} \tag{4}$$

where I_v is the set of items reviewed by v. By subtracting the average ratings of the items from a reviewer's ratings, his or her generosity is determined relative to that of the other reviewers.

Equations 3 and 4 are used to iteratively re-estimate the scores and the generosities until the model settles into a stable solution, and the final scores are taken to reflect the actual values of the items. The settling process converges quite rapidly, generally requiring about 6 iterations.

This *Linear model*, as thus described, is actually somewhat under-constrained because there are multiple solutions that result in the same error value. For example, one could add a constant to all of the generosities and subtract that constant from all of the scores without affecting the predicted ratings and thus the error. In practice, there is usually a small drift in the generosities and scores with each iteration. Therefore, we add an additional constraint, which is that the generosities of the reviewers must sum to 0, resulting in a unique solution. This is enforced by subtracting the average generosity from each g_v after every update.

The astute reader will note that the Linear model is quite similar to a two-way ANOVA. However, they differ in that the ANOVA is not designed to handle missing data, while this is a situation with potentially rather sparse data. When computing estimated scores, the iterative model effectively fills in a missing rating, $\hat{r}_{iv}$, using the sum of the current estimate of generosity, g_v and the new estimate of the score, s_i.

3.1 A priori biases

In many circumstances, the performance of statistical models that must rely on a small amount of data can be improved by introducing an *a priori* bias towards a simpler model or towards statistics that are closer to the expected mean. This is especially true in the presence of noise. In this case, we may want to discourage extreme generosities and scores, under the assumption that such values are unlikely to occur in a natural situation. One way to achieve this is to introduce a cost term to the error function that penalizes scores that are far from the expected mean, s_m, and generosities that are far from the mean, g_m, which is 0 in this case. Although there are many ways to do this, the following squared-error function proves quite effective:

$$E = E_R + C\left(\sum_i (s_i - s_m)^2 + \sum_v (g_v - g_m)^2\right) \quad (5)$$

Taking the derivative of the error with respect to s_i, setting it to zero, and solving for s_i results in the new error-minimization function for s_i to replace (3):

$$s_i = \frac{s_m C + \sum_{v \in V_i} \hat{r}_{iv} - g_v}{C + \mid V_i \mid} \quad (6)$$

A corresponding equation is used to find the locally optimal value for g_v. Once ratings have been normalized to a $(0, 1)$ scale, as explained in Section 4.1, a mean score of $s_m = 0.5$ is used. The extreme value penalty, C, can be adjusted depending on the level of bias desired. In practice, a value of $C = 0.25$ is a good default choice for natural tasks, with optimal values for the Linear model ranging from 0 to 0.5. Models with a C value greater than zero will be referred to a *biased* models.

3.2 Problems with the Linear model

The Linear model is quite simple and fast to train, and, as we'll see in Section 5, usually results in a significantly more accurate estimate of an item's true score than either the average rating or the average z-score. However, this model may not be appropriate for most practical situations because it assumes an unbounded rating scale, while the most common procedures for assigning ratings involve bounded, usually discrete, scales, such as the integers from 1 to 5 or 1 to 10. As a result, the Linear model does not behave appropriately at the boundaries.

For example, if an item really is exceptional, one should expect that most reviewers, generous or not, will give it a high rating. On a low-resolution scale, they may all give it the maximum rating. Likewise, reviewers will tend to agree about truly terrible items. Reviewers of differing generosity will tend to disagree the most over the items of moderate quality. The Linear model assumes, to the contrary, that reviewer generosity will have an equal effect on all items and "predicts" that a generous reviewer would give a very good item a rating that is beyond the limit of the scale.

Therefore, we will now turn to an improved model of the review process that is more appropriate for situations with bounded rating scales.

4 A non-linear model

Our goal is to develop a model that accords better with an intuitive understanding of how different reviewers are likely to use the ratings scale. Again we will begin with a model of the review process that relates a generosity and a score to a rating, but in this case the model will be multiplicative, rather than additive.

Let us start by assuming that the true score of an item is unbounded and falls in the range $[0, \infty]$. Scores in this range will be referred to as *unbounded scores* and denoted s'_i. Why use a range from zero to infinity? Let us assume that the items to be reviewed are academic or scientific papers. There is certainly a reasonable lower bound on the quality of a paper. A blank paper, for example, would be worthless. The present report aside, it is hard to imagine that any other paper could be worse than worthless. But how can one place an upper bound on the quality of a paper? Even if a paper is exceptional, it is surely always possible that it could be improved or that a better paper might one day be written. For the time being, let us also assume that a reviewer has a potentially unbounded generosity, g'_v, which also falls in the $[0, \infty]$ range and that the rating assigned by that reviewer to an item is the product of this generosity and the item's unbounded score:

$$r'_{iv} = g'_v s'_i \tag{7}$$

A reviewer with a generosity of 1 assigns accurate ratings, while one with a generosity of 2 thinks all items are twice as good as they really are. But there is a problem with this model, which is that it produces unbounded ratings, while our goal is to model ratings on a fixed interval. Therefore, we will translate the unbounded rating, r'_{iv}, into a *bounded rating*, r_{iv}, in the range $[0, 1]$ through the following transformation:

$$r_{iv} = \frac{r'_{iv}}{1 + r'_{iv}} \qquad r'_{iv} = \frac{r_{iv}}{1 - r_{iv}} \tag{8}$$

The form of equation 8 on the right is also known as the odds ratio. If r_{iv} were a probability, r'_{iv} would be its odds. Note that, if $r'_{iv} = 1$ then $r_{iv} = 0.5$. If the model is to be applied to a case in which the true ratings are not on a $(0, 1)$ scale, the ratings must be normalized accordingly, as discussed in Section 4.1.

Although we began with unbounded scores and generosities, it is not always convenient to work with such values. Therefore, these parameters can also be mapped to bounded versions using similar odds transformations:

$$g_v = \frac{g'_v}{1 + g'_v} \qquad g'_v = \frac{g_v}{1 - g_v} \tag{9}$$

$$s_i = \frac{s'_i}{1 + s'_i} \qquad s'_i = \frac{s_i}{1 - s_i} \tag{10}$$

The bounded value g_v will simply be referred to as a reviewer's *generosity* and s_i as an item's *score*. Using Equations 8–10, we can substitute for r'_{iv}, g'_v, and s'_i in Equation 7 and solve for each of the three variables to produce the following set of equations:

$$r_{vp} = \frac{g_v s_i}{1 - g_v - s_i + 2 g_v s_i} \tag{11}$$

$$g_v = \frac{r_{vp}(1 - s_i)}{r_{vp} + s_i - 2 r_{vp} s_i} \tag{12}$$

$$s_i = \frac{r_{vp}(1 - g_v)}{r_{vp} + g_v - 2 r_{vp} g_v} \tag{13}$$

So, given a rating and the actual score of the item, we can compute the reviewer's generosity and given a rating and the generosity, we can compute the item's score. The unbounded terms, such as s'_i, were a useful tool in deriving this otherwise opaque set of equations. Henceforth, as useful or appropriate, we can work with either these unbounded values or with bounded ratings, generosities, and scores that fall on the $(0, 1)$ interval.

Let's take a look at how these equations govern a reviewer's transformation of an item's score into a rating.

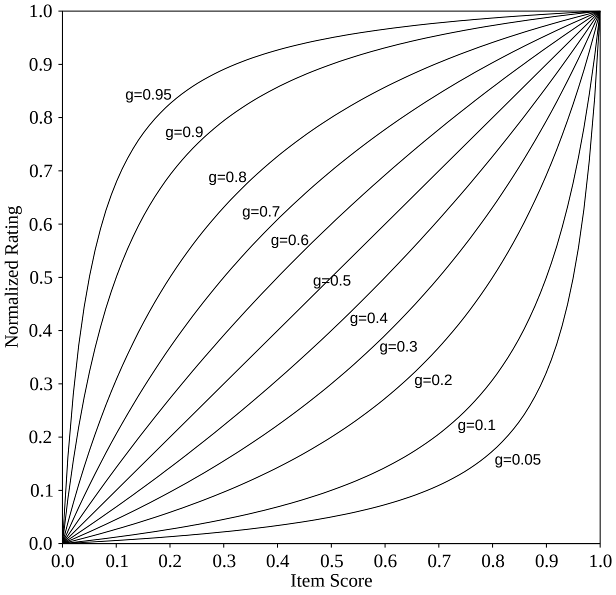

Figure 1: The rating an item will receive as a function of the item's score and the generosity of the reviewer, g, according to the Spindle model.

Figure 1 shows the rating an item will receive as a function of the item's score under several different levels of generosity, g. Note that if the generosity is 0.5, the rating is equal to the score. If the generosity is greater than 0.5, the ratings are inflated. An attractive property of this set of equations is that the curves are symmetric. Each curve is symmetric across the leftward diagonal, $r = 1 - s$, and the curve for $g = x$ is the reflection of the curve for $g = 1 - x$ in the rightward diagonal, $r = s$. These properties of the equation result in the high and low ends of the scale receiving symmetric treatment. This would not be true of many other non-linear models of the relationship between score, generosity, and rating, such as $r_{iv} = s_i^{1/g'_v}$. Incidentally, the name for this model, the *Spindle model*, derives from the fact that Figure 1 seems to resemble the mitotic spindle during cell division.

One important fact to consider is that singularities occur at the endpoints of the rating, score, and generosity scales. If an item has a score of 1, it will be given a rating of 1 by any reviewer, regardless of his or her generosity. In that case, it isn't possible to determine the generosity given the rating and score. In theory, this should not present a problem as no item is expected to be perfect or completely worthless. Similarly, no reviewer should have a generosity of 0 or 1 because he would assign all items the same rating, which is not useful. In practice, this problem can be avoided if no item is allowed to have a rating of exactly 0 or 1, as discussed further in the next section.

4.1 Normalizing the ratings

In order to apply the model to a particular reviewing situation, we must first normalize the range of the ratings given by the reviewers to the $(0, 1)$ scale. Assume, for example, that the items were rated on an integer scale from 1 to 5. The most obvious method of normalizing would be to map a rating of 1 to 0 and a rating of 5 to 1. Not only would this method result in singularities because of the possibility for perfect 0 or 1 ratings, it does not reflect the most reasonable interpretation of the discrete scale. Presumably, each of the five possible integer ratings is the result of rounding a real-valued rating from an underlying continuous scale. Thus a rating of 3 captures all real-valued ratings that would have fallen in the range $[2.5, 3.5)$. If we assume that each of the integer ratings represents an equal portion of the original continuous scale, then a rating of 1 represents values in the range $[0.5, 1.5)$ and the maximum value, 5, captures the range $[4.5, 5.5)$. So the appropriate $(0, 1)$ normalization in this case is to map 0.5 to 0 and 5.5 to 1.

In general, if the minimum interval used in the rating scale is i, the normalized ratings can be computed from the raw ratings, $\hat{r}^*_{iv}$, as follows:

$$\hat{r}_{iv} = \frac{\hat{r}^*_{iv} - (\text{min} - i/2)}{\text{max} - \text{min} + i} \tag{14}$$

Henceforth, we will assume that all ratings, $\hat{r}_{iv}$, have been normalized via this equation.

If the initial ratings were truly on a continuous scale, rather than a discrete scale, the correct setting for i is 0. However, if any items were given the minimum or maximum possible rating on that scale, this will result in problems because of the aforementioned singularity. If this situation arises, one can avoid the problem by setting i to some small epsilon, such as $1e^{-6}$.

4.2 The Logistic model

Like the Linear one, the Spindle model is trained by searching for the set of scores and generosities that minimize the sum-squared error between the predicted ratings and the normalized actual ratings. But the question is, which ratings space should we work in: the bounded space of r_{iv} and $\hat{r}_{iv}$, or the unbounded space of r'_{iv} and $\hat{r}'_{iv}$?

Let us first consider working in the unbounded ratings space. One approach might be to simply minimize the sum-squared error between the terms r'_{iv} and $\hat{r}'_{iv}$. However, in this unbounded space, the ratings at the positive end of the scale will be much larger than those at the low end of the scale and will tend to dominate the error function. Therefore, the error minimization process will primarily focus on the high ratings and may ignore the information in the low ratings.

One solution is to log-transform the predicted and actual ratings and to compute the sum-squared error in log ratings space. This approach has the added benefit of greatly simplifying the mathematics. Equation 7 now becomes:

$$\log r'_{iv} = \log g'_v + \log s'_i \tag{15}$$

This is identical in form to the equation upon which the Linear model is based. The ratings component of the error function we will seek to minimize is:

$$E_R = \sum_{i,v} (\log r'_{iv} - \log \hat{r}'_{iv})^2 \tag{16}$$

Rather than working with s_i or s'_i values, we will perform the error minimization directly with log unbounded scores, $\log s'_i$, and generosities, $\log g'_v$, to simplify the mathematics. We can convert back to bounded values to normalize the generosities and once the training is done. As a result, the error minimization can be performed in much the same way as in the Linear model, using the following update equation for the scores, with a corresponding equation for the generosities:

$$\log s'_i = \frac{\log s'_m C + \sum_{v \in V_i} \log \hat{r}'_{iv} - \log g'_v}{C + \mid V_i \mid} \tag{17}$$

In this case, the extreme value penalty, C, is operating in a different error space and, as it turns out, larger bias values are needed to achieve the same effect.

The astute reader may further note that Equation 15 bears a strong resemblance to a logistic regression. For that reason, this variation of the non-linear model will be referred to as the *Logistic model*. However, this method differs from a true logistic regression in that it is multivariate and is not actually computing a true regression.

4.3 The Spindle model

Although the Logistic model is quite simple mathematically and, unlike the Linear model, conforms to the assumption of a bounded ratings space, it is still possible that the model is overly sensitive to the extreme ends of the domain. Rather than operating in log unbounded ratings space, it would seem more desirable to minimize the error in bounded ratings space, as is essentially done in the Linear model.

Although we will now work with ratings in bounded space, because that is the form in which the error function is to be expressed, we can as a matter of convenience continue to work with either bounded or unbounded scores and generosities. In this case, the error equation is simplest if expressed with unbounded values. By substituting into Equation 2 using Equations 7 and 8, we obtain:

$$E_R = \sum_{i,v} \left(\frac{g'_v s'_i}{1 + g'_v s'_i} - \hat{r}_{iv} \right)^2 \tag{18}$$

We wish to choose the values of s_i and g_v, or s'_i and g'_v, that minimize this error. This is done by iteratively adjusting scores in order to minimize the error given the generosities and then adjusting the generosities to minimize the error given the scores. As before, we can find the bounded score value, s_i, that locally minimizes the error by computing its partial derivative with respect to s_i and setting this equal to zero. The partial derivative is as follows:

$$\frac{\partial E_R}{\partial s_i} = 2 \sum_{v \in V_i} \frac{g'^2_v (s_i - r_{iv} s_i) + g'_v (r_{iv} s_i - r_{iv})}{(1 + g'_v s_i - s_i)^3} \tag{19}$$

Unfortunately, the s_i root of this equation is not easily found in closed form. However, the Newton-Raphson method of successive approximations allows us to iteratively solve for the optimum s_i, provided that we can compute the second partial derivative of the error, which is:

$$\frac{\partial^2 E_R}{\partial s_i^2} = 4 \sum_{v \in V_i} \frac{g'^3_v (r_{iv} s_i - s_i) + g'_v (r_{iv} s_i - r_{iv})}{(1 + g'_v s_i - s_i)^4} + \frac{g'^2_v (r_{iv} + s_i - 2 r_{iv} s_i + \frac{1}{2})}{(1 + g'_v s_i - s_i)^4} \tag{20}$$

Like the Linear model, the Spindle model can benefit from the introduction of an *a priori* bias against extreme generosities and scores. Adding the same cost function to the error equation to penalize extreme values, we obtain:

$$E = E_R + C \left(\sum_i \left(s_i - \frac{1}{2} \right)^2 + \sum_v \left(g_v - \frac{1}{2} \right)^2 \right) \tag{21}$$

$$\frac{\partial E}{\partial s_i} = \frac{\partial E_R}{\partial s_i} + 2C \left(s_i - \frac{1}{2} \right) \tag{22}$$

$$\frac{\partial^2 E}{\partial s_i^2} = \frac{\partial^2 E_R}{\partial s_i^2} + 2C \tag{23}$$

The Newton-Raphson method begins with an estimate of the root to be computed, s_i in this case, and then iteratively refines the estimate according to the following update equation:

$$s_{i(t+1)} \leftarrow s_{i(t)} - \frac{\frac{\partial E}{\partial s_{i(t)}}}{\frac{\partial^2 E}{\partial s^2_{i(t)}}} \tag{24}$$

If a function is well-behaved and the initial estimate is close to correct, the Newton-Raphson method converges quite rapidly and consistently. In the first iteration of the Spindle algorithm, the generosities are all initialized to 0.5. In this case, the optimal score for an item is actually the average of its ratings and there is no need to run the Newton-Raphson iteration to optimize s_i. Subsequently, the starting point for the iteration is simply the previous value of s_i. In this case, the method converges on a new value, to a criterion of $1e^{-6}$, in about 4 iterations on average.[1]

Because the model is symmetric with respect to the scores and generosities, the equations for approximating the optimal unbounded generosities, g_v, are equivalent to Equations 19–24 after swapping s_i and g_v, g'_v and s'_v, V_i and I_v, and so on.

To recap, the Spindle method operates as follows. First, the raw ratings are normalized to the range (0,1) as described in Section 4.1. The initial bounded generosities are all set to 0.5, which is equivalent to an unbounded generosity of 1, and the initial scores are set to the averaged ratings. Then the generosities are re-estimated using the Newton-Raphson method to minimize the error. This involves corresponding versions of Equations 19–24. As with the Linear method, the generosities are additively adjusted to maintain an average bounded generosity of 0.5. Next the scores are recomputed using the same process, without adjusting to a mean of 0.5. The steps of re-computing the generosities and the scores repeat until the entire model settles to the point that no score changes by more than $1e^{-4}$. This usually requires 6–7 iterations.

5 Evaluation

We have introduced five methods for estimating overall item score from a collection of ratings. The Average model is the most common approach, and simply involves averaging the ratings for each item. The Z-Score method relies on the mean and variance of the ratings given by each reviewer, with no comparison between reviewers. Finally, the Linear, Logistic, and Spindle models simultaneously estimate both item scores and reviewer generosity by means of inter-reviewer comparisons. These final three methods are parametric in that they permit a scalable bias against extreme scores and generosities.

[1] In rare circumstances on very large tasks, the Newton-Raphson method fails to converge. In this case, the Spindle program searches for a better initial value until convergence is achieved or it is forced to give up.

In this section, we evaluate the performance of these models on an artificial task and on three tasks involving actual human ratings. At issue is how effectively the models can estimate the true scores of the items given just a limited sampling of ratings data.

5.1 Evaluation measures

There is no single best metric for evaluating the models' performance, because different aspects of their behavior will be more or less relevant for various applications. Therefore, four main evaluation measures will be used here. In each case, we will assume that there is a correct set of item scores against which the models' estimates can be tested. With artificial data, there can actually be a set of scores that are known to be correct. In a natural task, the "correct" scores will be determined by the average ratings of each item across all reviewers, not just the limited set of data that will be available to the models.

The first metric is the root mean squared difference between the correct scores and the models' estimated scores across the items. This will be known as the *RMS error*. This measure is the most strict in that it requires the model to reproduce the scores exactly, not just their relative magnitudes. However, this measure is not appropriate for the Z-Score method, which does not produce scores in the $(0, 1)$ range as the other methods do. Therefore, a variation on the RMS error is to first perform a linear transformation of the models' scores that minimizes the error. This variant will be referred to as *best-fit RMS error*.

On the other hand, in many cases we are not actually interested in accurately reconstructing the true scores of the items. Rather, we are interested only in properly ordering them. Therefore, we could instead ask how closely the estimated ranks of the items match their true ranks. One measure of this is the percentage of item pairs that have been ranked in the wrong order on the basis of the estimated scores.[2] This will be known as *rank error*. A perfect ordering should result in a rank error of 0%, while a random ordering is expected to have 50% rank error.

One additional measure of interest is how consistently the models assign scores to an item across many small samplings of the full set of ratings. If the rating of an item varies significantly from trial to trial depending on the identity of the reviewers, then the model has not effectively accounted for reviewer bias, and the scores resulting from a single experiment may not accurately reflect the scores obtained in a replication of that experiment. The measure of *inconsistency* that we will use is the ratio of the average standard deviation of the estimated score of each item across the trials to the standard deviation of the mean item scores across the trials. This can be thought of as the ratio of the variance within items to the variance between items. A low score indicates that the model is consistent in its scoring despite changes in reviewer, and other factors that vary from trial to trial.

5.2 Artificial ratings

We will begin with a test involving artificial data, the advantage of which is that we have control over the true scores of the items as well as the generosity of the reviewers and other factors. However, generating artificial data requires the selection of a function for generating the simulated ratings and the nature of this function could bias the results in favor of one model over another.

These simulated experiments involve 50 items whose true scores range in equal increments from 0.02 to 0.98 and a pool of 9 reviewers whose generosities ranged from 0.1 to 0.9 in equal increments. In a given trial, three reviewers were assigned to each item at random and ratings were generated on the basis of the item's score and the reviewers' generosities. In the first four conditions, simulated ratings were generated using the Spindle equation given in (11). In two of the four conditions (R$-$), the ratings were real-valued in the range $(0.5, 10.5)$, while in the other two conditions the ratings were rounded (R+) to whole numbers to reflect the discretization involved in most rating scales. Two of the four trials used no noise (N$-$), with the real-valued rating directly generated by Equation 11, while the other two added noise (N+) randomly generated in the range $[-1, 1]$ to each rating. In the R+ N+ conditions, the noise was introduced prior to rounding. One hundred trials were performed for each model in each condition and the results were averaged.

Figure 2 shows the RMS error for the five models and the eight ratings conditions. The Linear, Logistic, and Spindle models appear twice in each condition, first with no extreme-value penalty and then with a moderate C value. The error for the Z-Score method is not depicted because this method does not produce scores in the appropriate range for this measure. Across all of the conditions, the Average model, shown in the white bars, produces rather poor results. We will begin by considering the four conditions on the left, marked Spindle. The unbiased Linear model, with $C = 0$, has an error rate that is just over half that of the Average model. The biased Linear model, with $C = 0.2$, is somewhat worse than the unbiased model on all conditions. Thus, the bias does not seem to help the Linear model on this artificial task.

In the first condition, in which the ratings are directly generated by the Spindle equations with no noise or rounding, both the unbiased Logistic and Spindle mod-

[2] In computing ranks, items with the same score receive a rank that is the average of the ranks they would have received if it were not an exact tie.

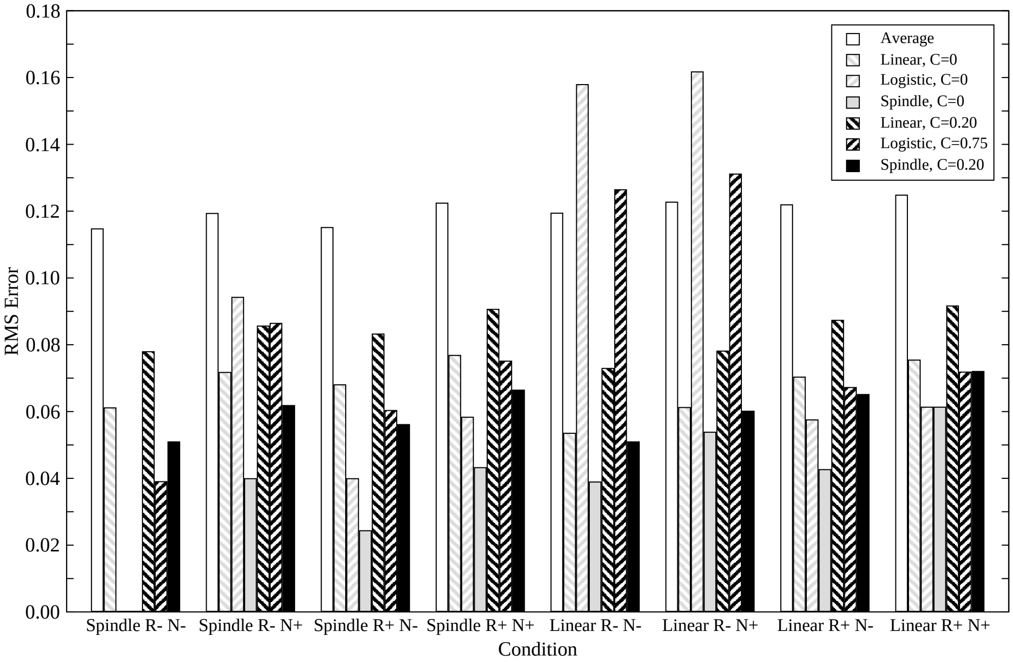

Figure 2: The root mean squared difference between the correct scores and the models' scores across the eight conditions in the Artificial task.

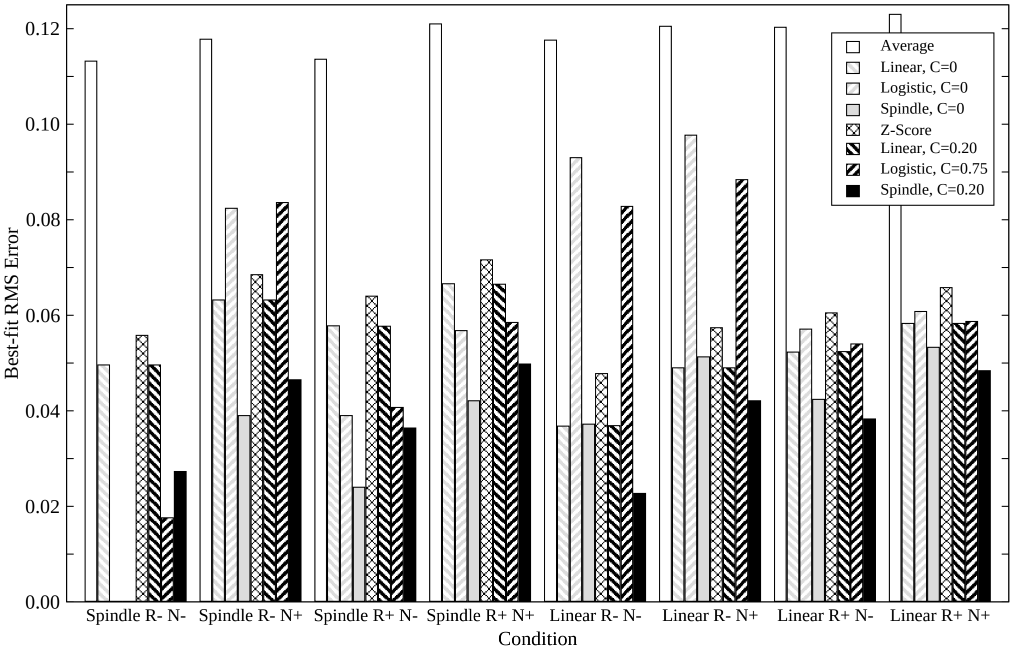

Figure 3: The root mean squared difference between the correct scores and the best linear fit of the models' scores across the eight conditions in the Artificial task.

els have zero error. This is to be expected and is less a validation of the approach than it is an indication that the implementations indeed achieve the correct minimal solution. With noise added, the Logistic model is somewhat worse than the Linear one without rounding, but somewhat better with rounding. However, in all four conditions the unbiased Spindle model achieves the lowest error. Noise and rounding do degrade the performance of the unbiased Spindle model, but they have less of an effect than on the Logistic model.

If the best-fit transformation of the scores is used in computing the error, as shown in Figure 3, the Z-Score method is comparable to the others, although the Average method continues to be much worse. By this measure, the Z-Score method is similar to but consistently worse than the unbiased Linear model. Again, the unbiased Spindle model has the lowest error, followed by the biased Spindle model and the Logistic models.

Of course, these results are biased by the fact that the Spindle model itself was used in generating the artificial ratings, so perhaps it is not surprising that it and the Logistic model perform better than the others. Therefore, in the next four conditions, the artificial ratings were generated using the Linear model. That is, the rating assigned by a reviewer was the sum of the item's score and the reviewer's generosity, bounded by the $(0.5, 10.5)$ interval. In this case, the generosities ranged from -0.4 to 0.4. As shown in the right half of Figure 2, this change has little effect on the Average and Linear models, but the unbiased Logistic model performs much worse on the non-rounded conditions. This may be because these conditions involve some ratings that are close to the bounds of the scale and which therefore dominate this model's error function.

Interestingly, the same is not true of the Spindle model. According to the RMS error, the unbiased Spindle model is the best, while the biased Spindle model actually has the lowest best-fit RMS error. Thus, even if the artificial data has been generated by the Linear model's equations, the unbiased Spindle model still outperforms the Linear model. The reason is that the data has been truncated to fall in the range $(0.5, 10.5)$, in order to fit our overriding assumption that the ratings should use a bounded scale. This truncation violates the assumptions of the Linear model, impairing its performance. This reflects the fact that the Linear model is not appropriate when working with ratings on a bounded scale, unless few of the ratings fall near the boundaries.

Figure 4 shows the rank error percentage on the artificial tasks. Again, the Average model is considerably worse than the other. The Z-Score model performs reasonably well, but is worse than both the Linear and Spindle models. Again the Spindle model achieves the lowest error across all eight conditions, although the biased Logistic model does quite well on the Spindle R$-$ N$-$ condition. In the first four conditions, the unbiased Spindle model outperforms the biased model, although the biased model is somewhat better on the non-rounded Linear conditions. Thus, the biased model looks much better when measured using rank error. The reason is that the bias term tends to produce a warping of the estimated scores and generosities, with the very low and high values drawn towards the middle. While this monotonic transformation affects the RMS error, it has little effect on the rank error because it is generally order-preserving.

Finally, Figure 5 depicts the inconsistency of the item scores across the 100 trials. These results largely parallel the rank error findings. As we might expect, the Average model is quite inconsistent, as is the Logistic model, which has trouble with the non-rounded conditions based on the Linear equations or having noise. The Z-Score model is significantly better than the Average model, but is consistently worse than the Linear and Spindle models. The unbiased Spindle model is better than the unbiased Linear model in all but one condition, while the biased Linear model is virtually identical to the unbiased in all cases. The biased Spindle model is generally worse than the unbiased Spindle model, except for the two Linear conditions with no rounding.

In summary, all of the other models result in considerably better performance on this artificial task, across a range of measures, than is achieved by simply averaging ratings. The Z-Score method was useful, but was generally outperformed by the multivariate techniques. The Logistic model was quite effective in five of the conditions, but had difficulty when real-valued ratings were used that potentially fell close to the bounds of the rating scale. With few exceptions, the Spindle model, either with or without the extreme value penalty, achieved the lowest error and the most consistent scores.

5.3 Attractiveness ratings

The performance of these models is one thing in an artificial context, where ratings are generated according to a formula. But how will they perform in a natural context in which reviewers may actually differ not just in their generosity but in the ranking they would assign to the items? The possibility of non-monotonic differences in the opinions of reviewers, discussed further in Section 6.1, is not explicitly accounted for in these models. Thus we might ask how well they can cope with this additional source of, what is for our purposes, noise.

The first task used to address this question involves rating physical attractiveness. 42 male participants viewed the photos of 60 women and rated their attractiveness on an integer scale from 1 to 7. This task was chosen because

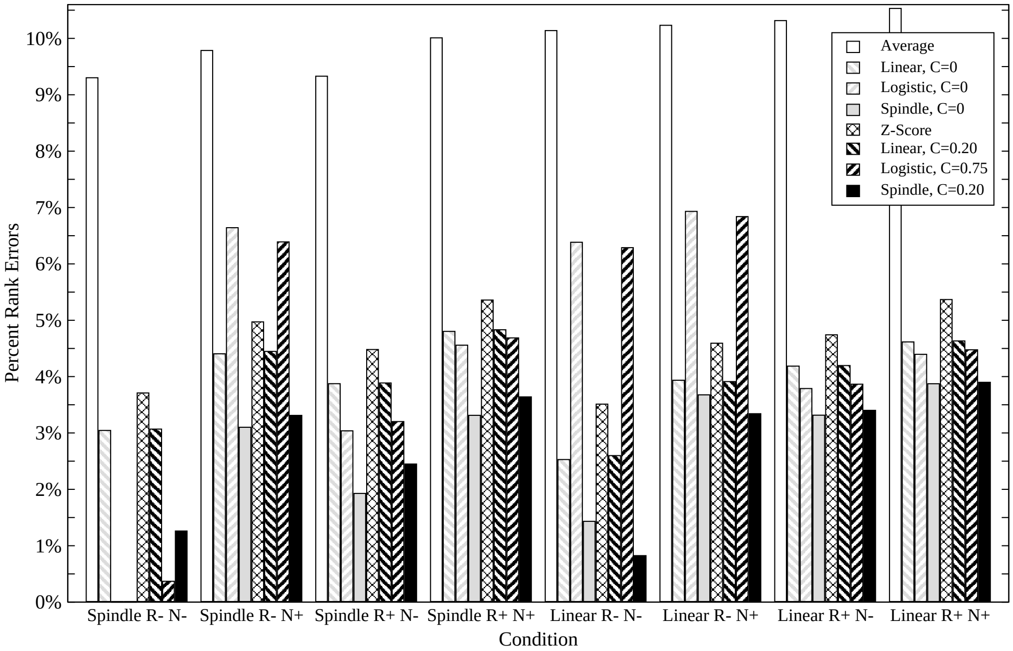

Figure 4: The percentage of misordered item pairs in the Artificial task.

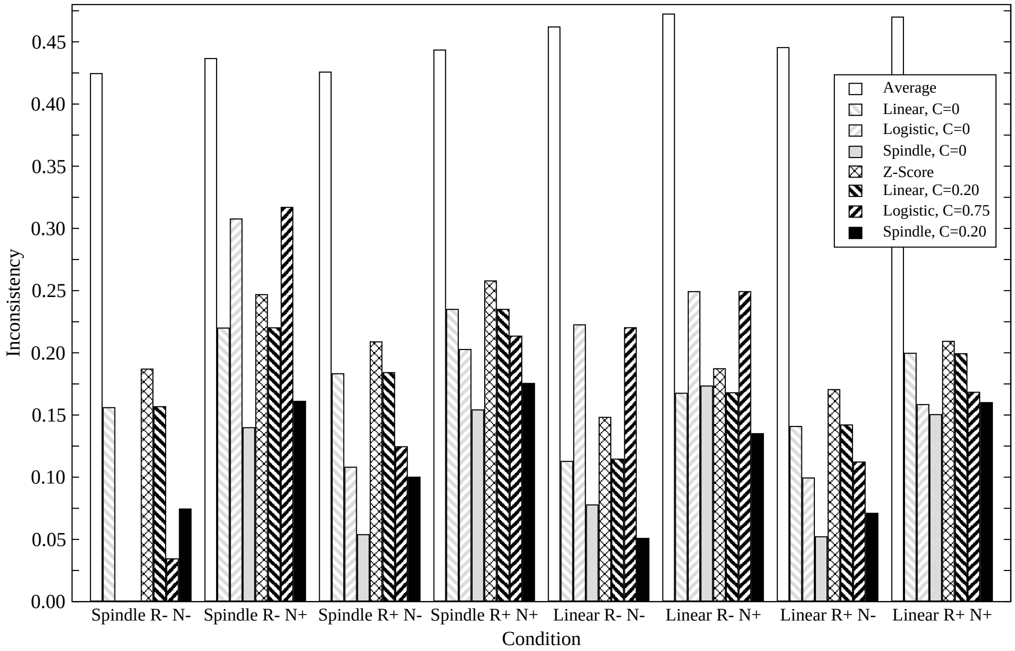

Figure 5: The inconsistency of the item scores over the 100 trials in the Artificial task.

it is one in which raters can be expected to basically agree with one another, although they may have distinct preferences and their opinions are likely to differ over the true ranking of the pictures.

In working with the artificial data in the previous experiment, we had the advantage of knowing the true scores of the items and were thus able to use these scores as the basis against which the models' estimates were measured. However, in a real-world survey such as this, we don't know the true scores of the items. Therefore, the correct scores were taken to be the ratings averaged over all 42 raters. Because each rater has rated all of the items in this controlled experiment, any discrepancy in the generosity of the raters should affect all of the items. In fact, the Linear model and the Average model will both produce the same set of ratings in this circumstance. On the other hand, the Spindle, Logistic, and Z-Score models would produce a slightly different set of scores if given the complete set of ratings. Therefore, these models will be at somewhat of a disadvantage in these tests because we will not be measuring against their preferred baselines.

We are going to test the ability of the models to estimate the true scores, derived from the average over all of the data, using only a subset of the ratings. In each trial of the experiment, 16 of the 42 reviewers were selected at random. Each of the items was then rated by a fixed-size subset of these 16 reviewers, subject to the constraint that the reviewers each provide roughly the same number of ratings. The number of reviewers per item was varied from 2 to 10. For each of these levels, 100 trials were conducted and the results of the models averaged across the trials. With just 2 ratings per item, each of the 16 reviewers contributed 7 or 8 ratings, providing a reasonable basis for estimating his generosity.

Figure 6 shows the RMS score error as a function of the number of ratings per item. As we might expect, the performance of all of the models improves with more reviewers per item. The Z-Score model, again, cannot be evaluated by this measure. The unbiased Linear model performs slightly better than the Average model with more than 2 reviewers per item. The unbiased Logistic and Spindle models are somewhat worse, as is the biased Linear model. However, the two models that consistently achieve the lowest RMS error are the biased Logistic and Spindle models, with the Spindle performing slightly better. The advantage of these methods over the others is most evident with fewer reviews per item.

Figure 7 displays the best-fit RMS error. In this case, there is much less difference between the models, although all of them perform better than the Average model with more than 2 items per reviewers. The biased Linear and Spindle models are consistently quite good, with the biased Logistic, Z-Score, and the unbiased models typically just a bit worse.

Figure 8 shows the rank error on the same task. This measure patterns very much like the best-fit RMS error. This also happens to be true on the other tasks to follow. With more than 2 reviews per item, all of the other models are an improvement over the Average. The biased Linear and Spindle models are consistently the best or nearly so, with the biased Logistic model performing well with less data and the unbiased and Z-Score models performing well with more data. Finally, Figure 9 shows the inconsistency of the item scores across the 100 trials. With the exception of the 2 reviews per item condition, the Average model is less consistent than the others. The unbiased Spindle model is the least consistent with fewer available reviews, while the Z-Score model is the best with more data. The biased models differ very little from one another.

In summary, the biased Spindle model performs quite well according to all four measures on the Faces task relative to the other methods. The other biased models are nearly as good, although the biased Linear has higher RMS error and the biased Logistic performs slightly worse as the available data increases. This latter fact may be because more data results in more extreme scores, despite the bias against this, and these extreme scores dominate the error minimization. The unbiased models and the Z-Score method are generally better than using the average rating, except when there are very few ratings per reviewer, but are more inconsistent than the biased methods and are much worse with small data sets.

It is clear, however, that the gap between the Average model and the others is not as pronounced on this task as it was on the artificial data. One obvious reason may be that the human ratings are simply noisier, increasing the theoretically optimal error rate that even an ideal model could achieve. This noise could be truly random noise in the raters' responses, or it could reflect the fact the raters actually have somewhat different standards of attractiveness. Thus, although we will continue to use 0 as a baseline for the graphs, the optimal performance in the face of noise and factors other than generosity may be much closer to the error levels actually achieved by the models.

Another reason for the smaller advantage of the more advanced methods is that the correct scores in the Faces task are not as evenly distributed as those in the Artificial task. The white bars in Figure 10 depict the histogram of correct scores on the Faces task. The scores are distributed around 0.425 in a roughly Gaussian manner, with only a few very low scores and no very high scores. Some of the raters reported that they were reserving their highest ratings in case an exceptionally attractive photo appeared, which apparently did not happen. When there is less variance in the true scores, ranking the items becomes more

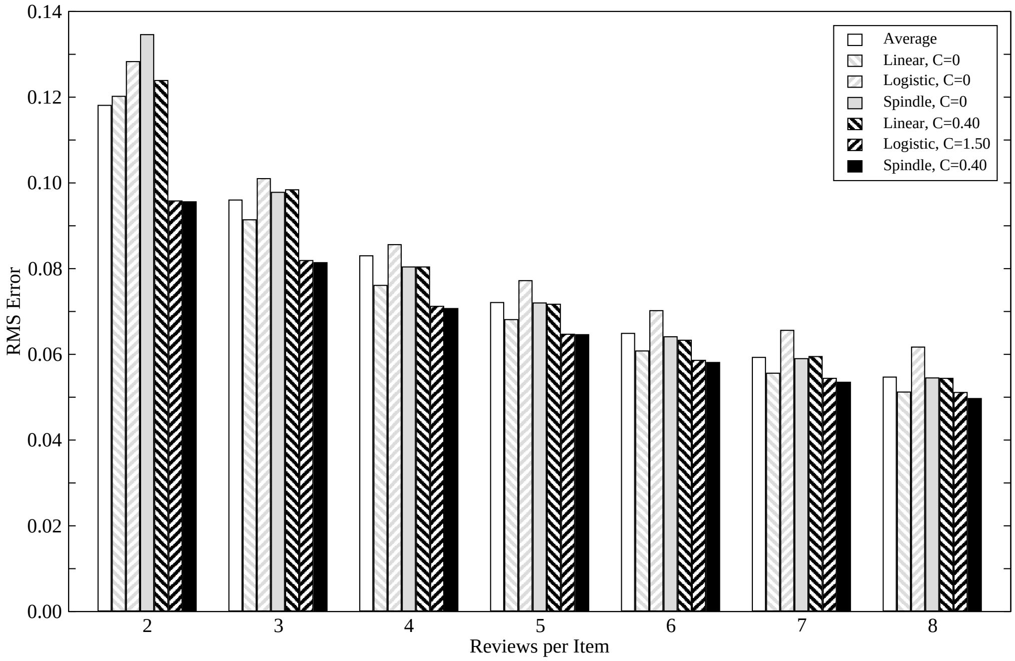

Figure 6: The root mean squared difference between the “correct” scores and the models’ scores as a function of the number of ratings per item on the Faces task.

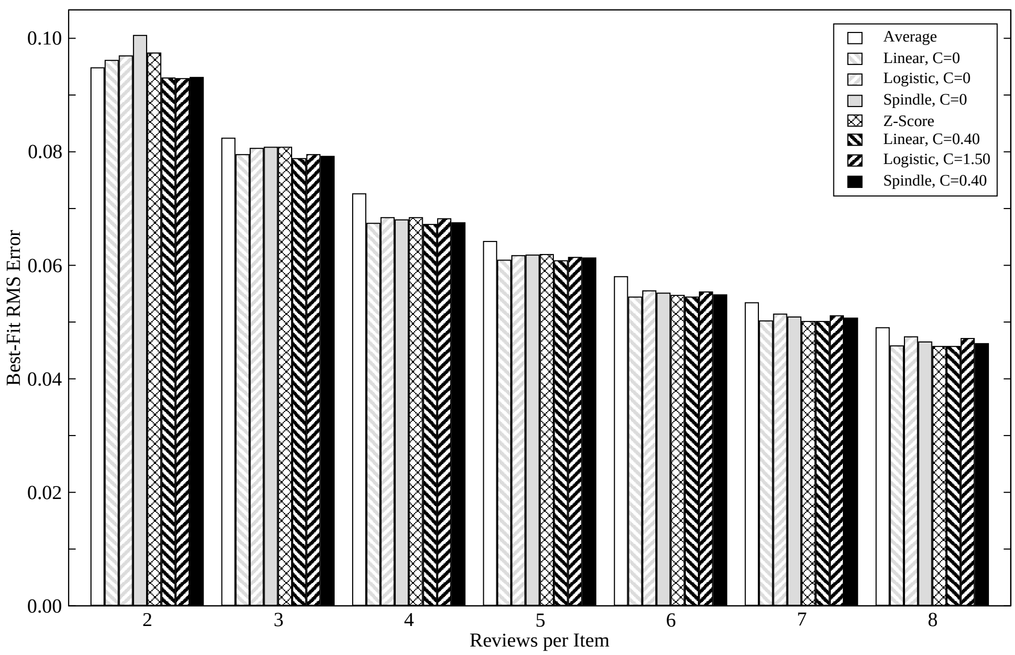

Figure 7: The root mean squared difference between the “correct” scores and the best linear fit of the models’ scores as a function of the number of ratings per item on the Faces task.

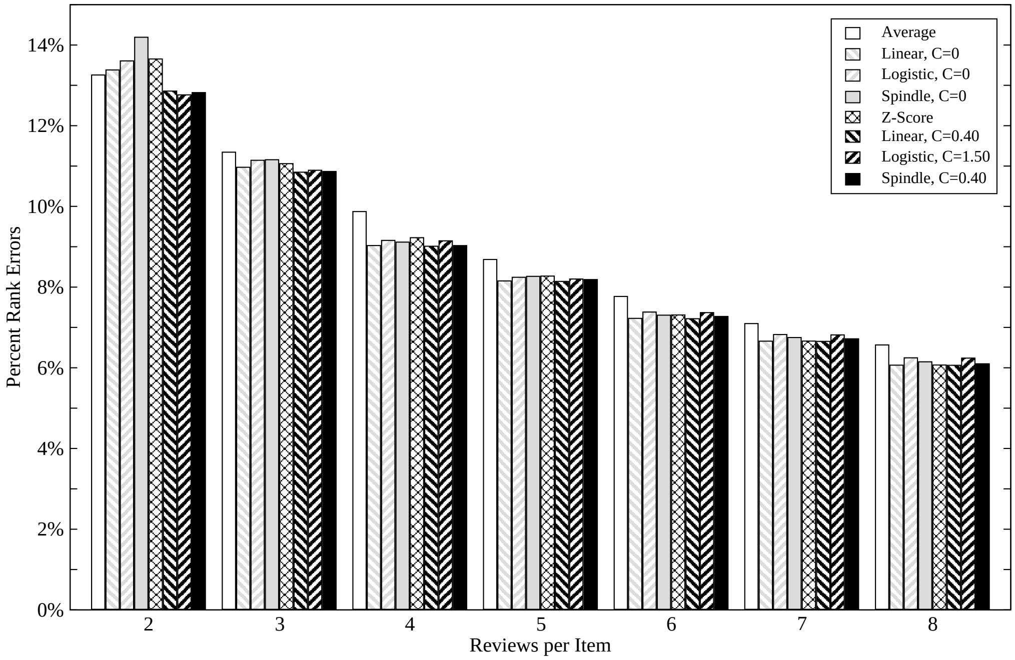

Figure 8: The percentage of misordered item pairs as a function of the number of ratings per item on the Faces task.

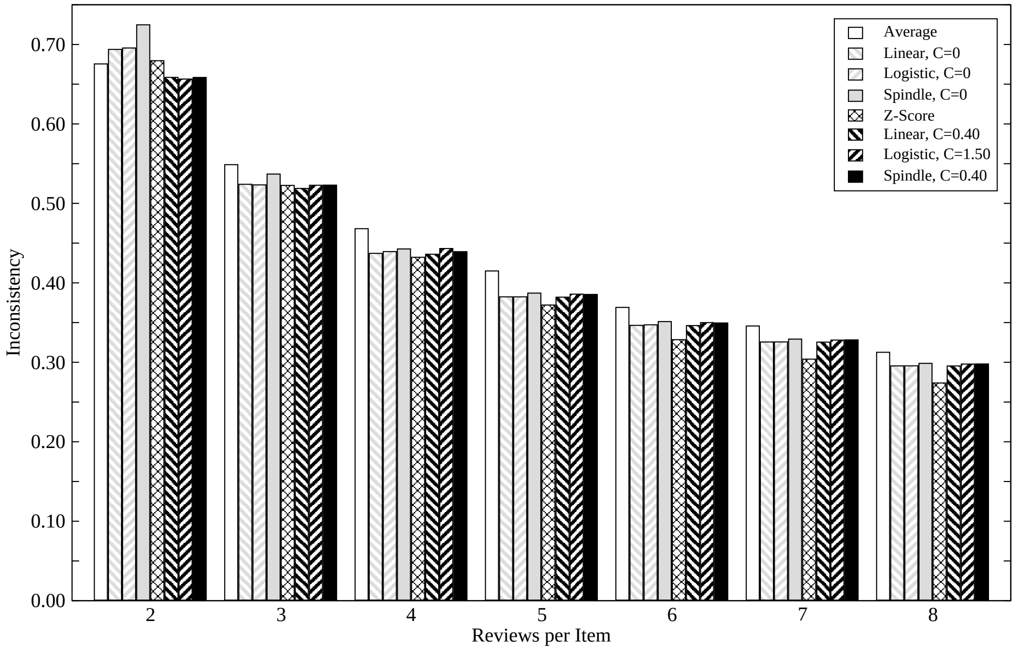

Figure 9: The inconsistency of the item scores over the 100 trials as a function of the number of ratings per item on the Faces task.

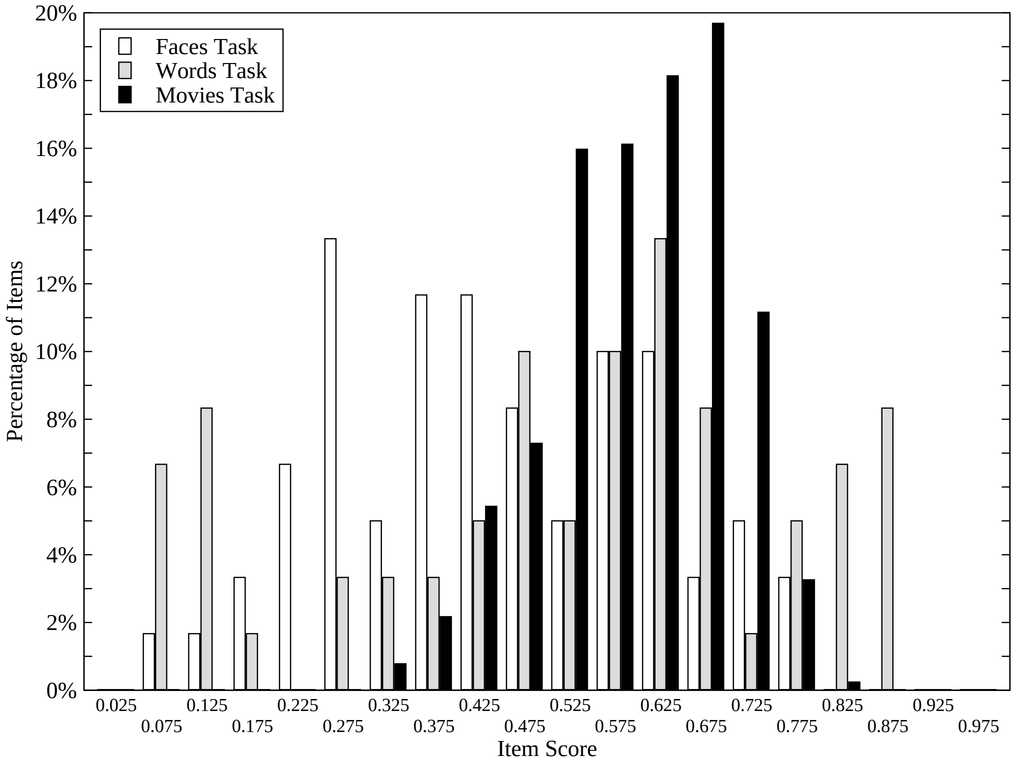

Figure 10: Histograms of the "correct" scores (the average rating for each item based on all of the data) for the three natural tasks used here.

difficult. Furthermore, when there are few items with extremely low or extremely high scores, the potential advantage of the Spindle model over the linear methods diminishes, because it is the extreme items on which their assumptions are most violated.

5.4 Word similarity ratings

This next task was intended to involve more items that would receive very high or very low average ratings. In this case, 32 participants were asked to rate the semantic similarity of 60 word pairs, again on an integer scale from 1 to 7. Some of the pairs were very close synonyms, such as *cry–weep*, while others were intended to be entirely unrelated, such as *donkey–gave*. Most of the pairs fell somewhere in the middle and were either members of the same superordinate category, such as *table–sofa*, or were associated nouns and verbs. In order to help anchor their judgments, raters were given three practice examples of very similar, very dissimilar, and intermediate pairs.

The "correct" scores were again determined by the average score of each item over all 32 raters. As shown in Figure 10, the efforts to produce more evenly distributed scores were partially successful. The grey bars in the histogram form a broader distribution than that for the Faces task. However, although there are more items with high and low scores, there are still no extremely high scores. Many raters continued to reserve their highest score for truly marvelous synonyms.

As before, the models were tested by choosing 16 raters at random and retaining between 2 and 10 ratings per item from this pool of raters. Because the scores on the Words task are more broadly distributed, the Linear and Spindle models perform better with extreme value penalties that are only half as large. The models' RMS score error is shown in Figure 11. The smoothed Spindle and Logistic models again achieve the lowest RMS error, with the Spindle model improving relative to the Logistic model with more data. The unbiased Linear model also performs quite well.

The rank error results on the Words task, which patterns almost exactly like the best-fit RMS error, is shown in Figure 12. There is little difference between the multivariate models, although the unbiased Spindle model is not as good with just 2 or 3 ratings per item. However, with more than 2 ratings per item, all of these models have lower rank error than the Average or Z-Score methods. The Z-Score performs particularly poorly on both the RMS and rank error measures with less data. In terms of inconsistency, shown in Figure 13, the Logistic models

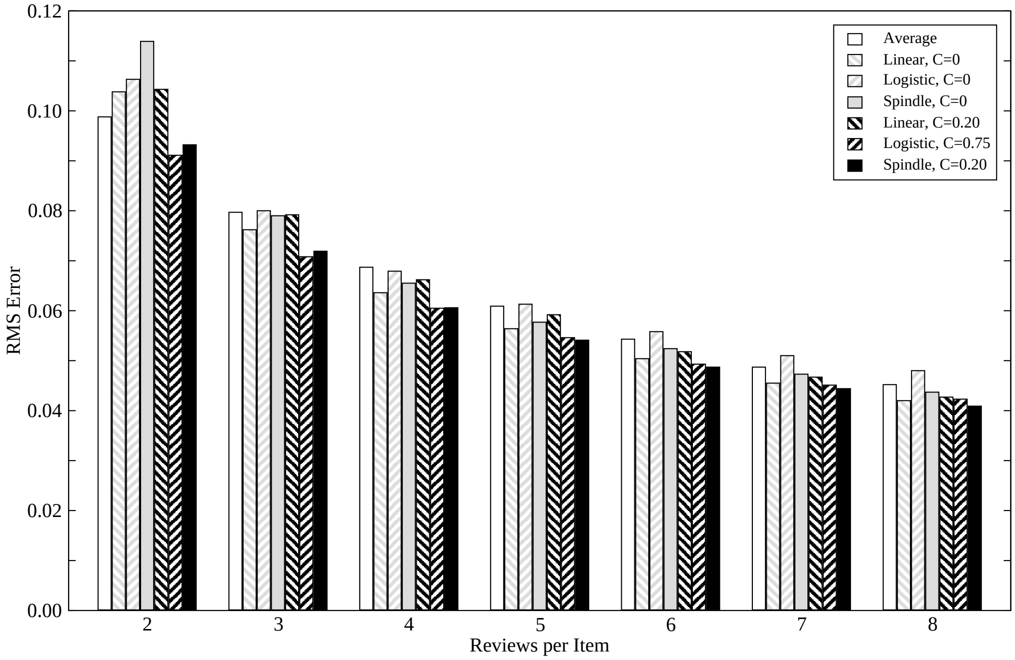

Figure 11: The root mean squared difference between the "correct" scores and the models' scores as a function of the number of ratings per item on the Words task.

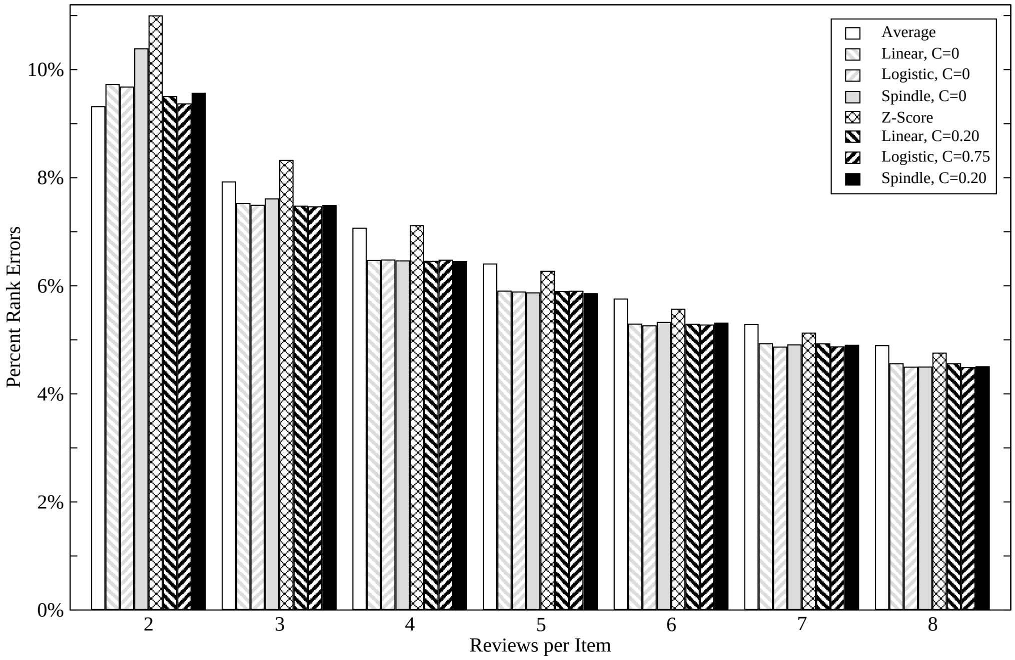

Figure 12: The percentage of misordered item pairs as a function of the number of ratings per item on the Words task.

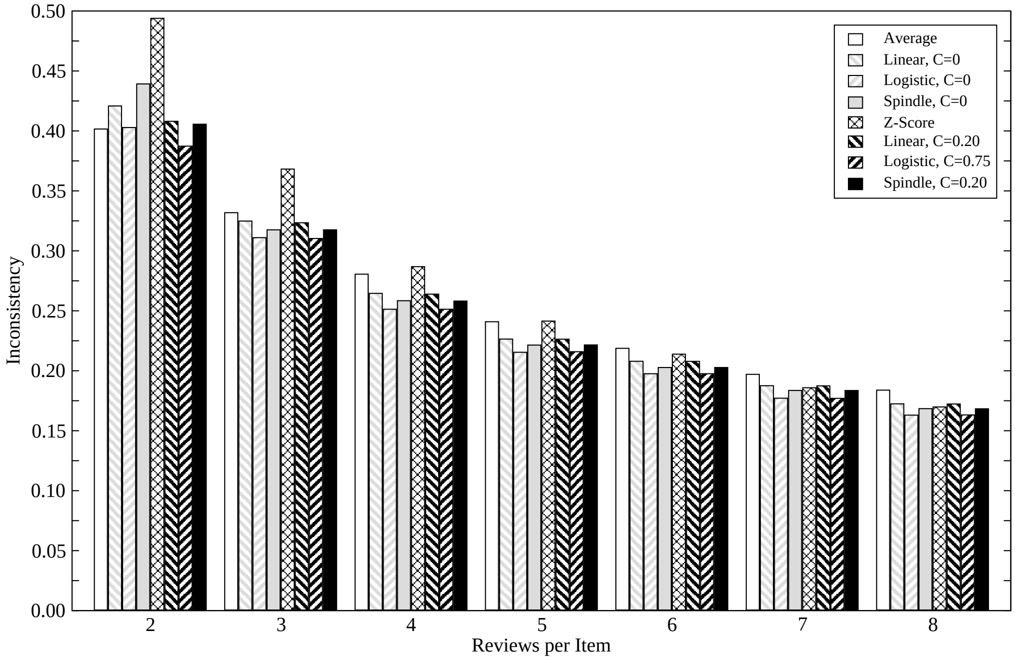

Figure 13: The inconsistency of the item scores over the 100 trials as a function of the number of ratings per item on the Words task.

are the best, followed by the biased Spindle, and then the unbiased Spindle and Linear models. As with the rank error, the Z-Score model is the worst with small amounts of data and the Average model is the worst with more data.

Despite the greater variance of scores, the overall profile of results on the Words task is very similar to that on the Faces task, although all of the models performed a bit better in this case. The biased Spindle and Logistic models are again the most effective, or nearly so, in all measures, followed by the biased Linear model.

5.5 Movie ratings

The Faces and Words tasks were useful tests because they were expected to conform reasonably well with the assumption, which is inherent in the models, that the raters will use similar standards in judging the items and that the ratings between any two reviewers will tend to differ monotonically as a function of the generosity of the reviewers. Indeed, the average correlation between the ratings of paired reviewers was 0.569 on the Faces task and 0.770 on the Words task, where a correlation closer to 1 indicates greater agreement between reviewers. However, in many real-world tasks, reviewers cannot be expected to base their judgments on similar standards. For example, in rating movies, some reviewers may prefer dramas, while others may prefer action or comedy. Will these methods continue to be effective in this situation?

This final task uses movie ratings from the MovieLens database, gathered on the web by the GroupLens Research Project at the University of Minnesota. The complete data set contains 100,000 ratings of 1,682 movies by 943 users, and is thus considerably larger than the data sets used in the other tasks. Because some of the movies received only a few ratings and not all users contributed the same number of ratings, the data set was reduced to permit the accurate evaluation of the models. The 500 most prolific reviewers were retained as well as the 776 movies which were reviewed by at least 40 of them, in order to provide a reasonably good estimate of the “true scores” of the movies. Following this trimming, 72,200 ratings remain. Again the models were tested on their ability to estimate the true scores of the movies given a small subset of the ratings and the results in each condition were averaged over 100 trials.

Several factors contribute to the difficulty of this task. First of all, not all reviewers rated all of the movies, so the baseline scores may be less accurate. Furthermore, as shown in the histogram in Figure 10, the average scores of the movies fall in a fairly tight Gaussian cluster. Therefore, small errors in the estimated scores may result in large errors in rank order. More importantly, however, the reviewers do not agree with one another. In contrast with the inter-reviewer correlations of 0.569 and 0.770 on the other tasks, the average correlation of the movie reviewer ratings was only 0.151.[3] Thus, it does in fact seem that the movie ratings are either extremely noisy or, more likely, that the reviewers are in fact relying on different standards and have strong preferences for different types of movies.

Finally, to make the task even more difficult, all 500 reviewers were used when selecting the 2–10 ratings per movie. As a result, there were fewer ratings per reviewer on this task than on the Faces and Words task, in which a reduced reviewer pool of 16 was used for each trial. With 2 ratings per movie, there were just 2 or 3 ratings per reviewer on each trial.

Despite these difficulties, the biased Spindle and Logistic models continue to perform reasonably well. Figure 14 shows the RMS score error on the Movies task. In this case, the Average model is not bad either, and is actually better than the unbiased models and the biased Linear model. Only the biased Logistic and Spindle models outperform the Average, significantly so with less data. Presumably the unbiased models are so much worse on this task because there are fewer ratings available per reviewer with which to estimate generosity. In this case, the *a priori* bias has a greater influence on the estimates and is quite helpful for the Spindle and Logistic models.

The rank error, which again patterns like the best-fit RMS error, is shown in Figure 15. The differences between the models are slight, with the exception that the Z-Score method is noticeably worse than the others. Figure 16 displays the inconsistency of the score estimates across trials. In general, the inconsistencies are much higher on this task because the inter-item score variability is so low, as is the inter-rater agreement. With only two or three reviews per item, and thus five or fewer ratings per reviewer, the unbiased models are quite inconsistent. Given more data, all of the models are about equally consistent with the exception of the Z-Score method.

As mentioned before, the Movie task results that were just presented involved 500 active reviewers in each trial. Because there were 776 movies in the trimmed database, if 3 ratings were chosen per item on a given trial, that translates to 4–5 ratings per reviewer. This is not a very large sample with which to estimate the generosity of a reviewer, particularly because the reviewer's opinions were so poorly correlated on this task. As a result, the Linear and Spindle models may have been severely hindered. By using a smaller set of active reviewers for each trial, and still choosing 3 ratings per item, we can model a situation in which more ratings are available per reviewer. To test the effect of the number of ratings per reviewer on the RMS score error, the number of active reviewers was varied from 500 down to 100. With only 100 active reviewers, there are 23–24 ratings per reviewer, which should be quite sufficient for estimating generosity.

Figure 17 shows the RMS error for trials with 3 ratings per item as a function of the number of active reviewers from which those ratings were drawn. As we might expect, the performance of the unbiased models, particularly the unbiased Spindle model, improves significantly with more ratings per reviewer. The performance of the biased Linear model improves, but less so, and it is matched by the unbiased Spindle model with 100 active reviewers. However, the performance of the biased Spindle and Logistic models, like that of the Average model, remains essentially stable. That is, the bias term is particularly useful in improving the estimates of reviewer generosity when there are only a small number of ratings available per reviewer, allowing the Logistic and Spindle models to perform nearly as well with much less data per reviewer.

5.6 Significance testing

The analyses conducted thus far have focused on the tasks of assigning overall scores or rankings to the items. However, another common goal of ratings studies is to test for significant differences between two or more sets of items. For example, a researcher might be interested in whether subjects rate one set of words used in a memory experiment as more imageable than another set of words. In situations like these, generosity differences between subjects will be a source of noise that may result in the detection of spurious differences if the assignment of items is not well balanced, or may reduce the power of the significance tests even if the items are well balanced.

Significance tests can be performed using these models if the model is used to produce a set of adjusted ratings, rather than a set of item scores. This ratings adjustment is intended to remove the effect of reviewer generosity from each of the ratings. With the Z-Score method, this is done by using the individual rating z-scores, rather than the raw ratings. For the Linear model, this is done by solving for s_i in Equation 1, which amount to subtracting the reviewer's generosity from the rating. For the Logistic and Spindle models, the adjusted rating is given by Equation 13.

The first test of the models' ability to detect a significant difference between conditions was based on the Artificial task, with ratings generated using the Spindle equation with noise and rounding. The items were partitioned into two equal groups. The true scores of the items in the first group were drawn at random from the range $[0.23, 0.73]$. The scores of the items in the second group were 0.04 higher than the scores of the corresponding items in the first group. Therefore, there was in fact a small but consistent statistical difference between the

[3] In computing the average correlation of the ratings between pairs of reviewers, only reviewer pairs who rated at least 5 movies in common were used.

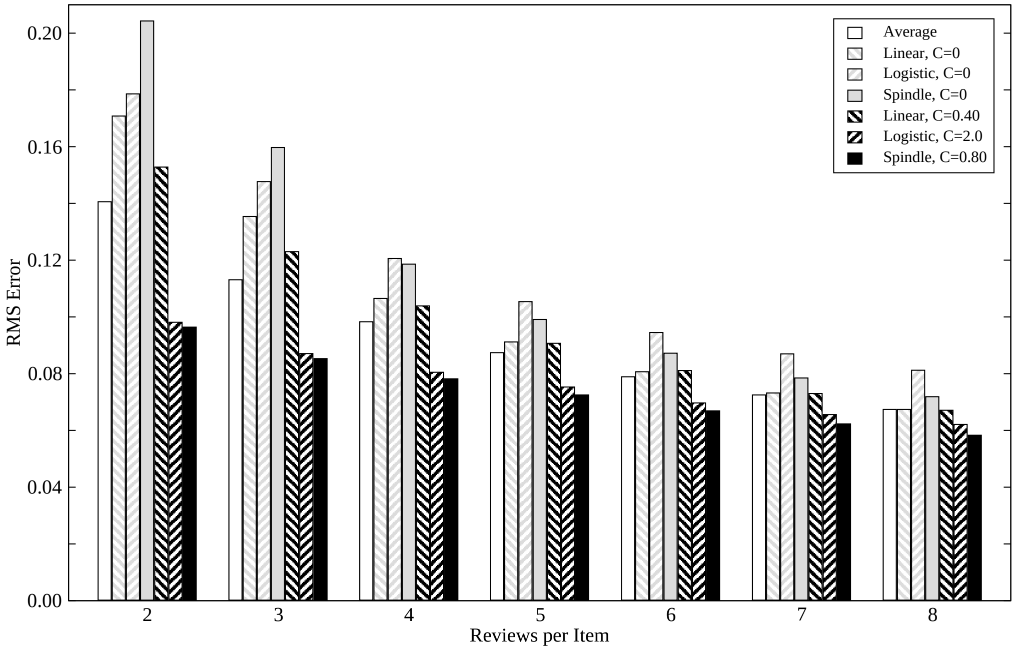

Figure 14: The root mean squared difference between the “correct” scores and the models’ scores as a function of the number of ratings per item on the Movies task.

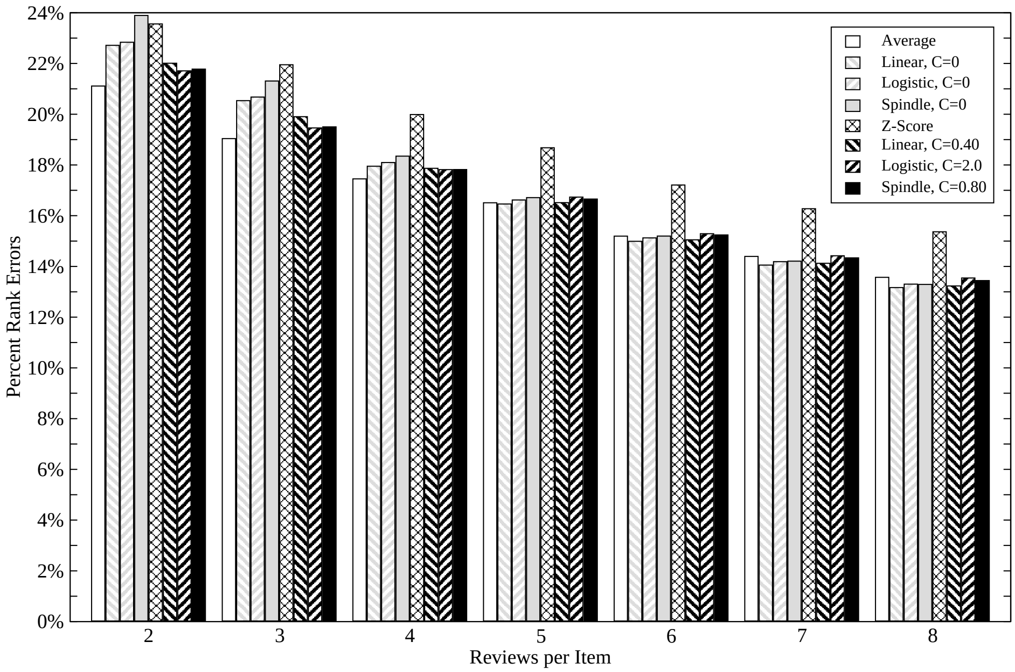

Figure 15: The percentage of misordered item pairs as a function of the number of ratings per item on the Movies task.

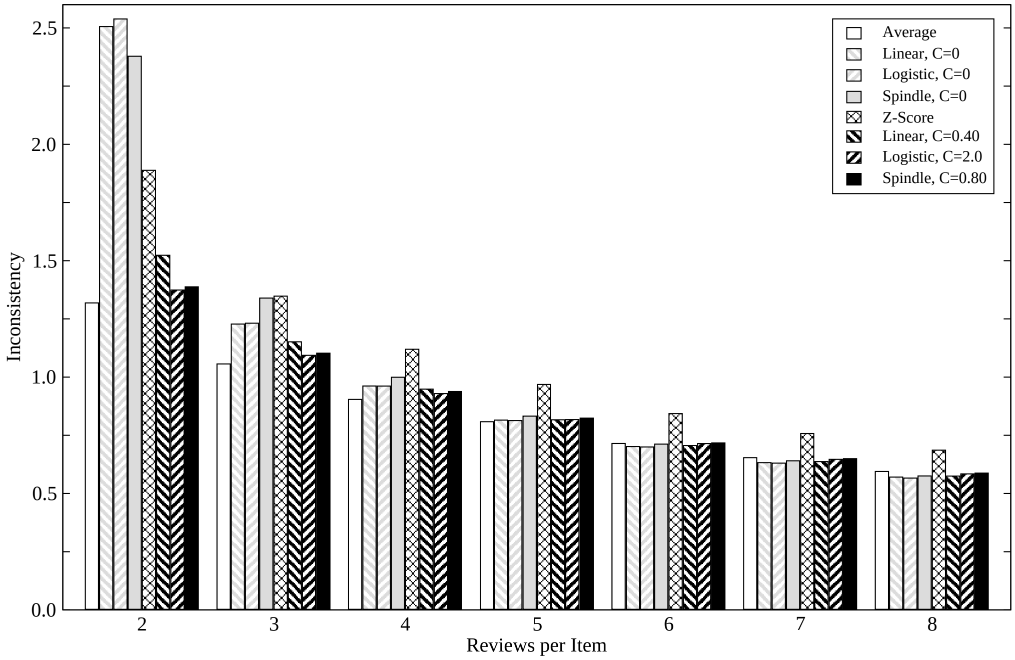

Figure 16: The inconsistency of the item scores over the 100 trials as a function of the number of ratings per item on the Movies task.

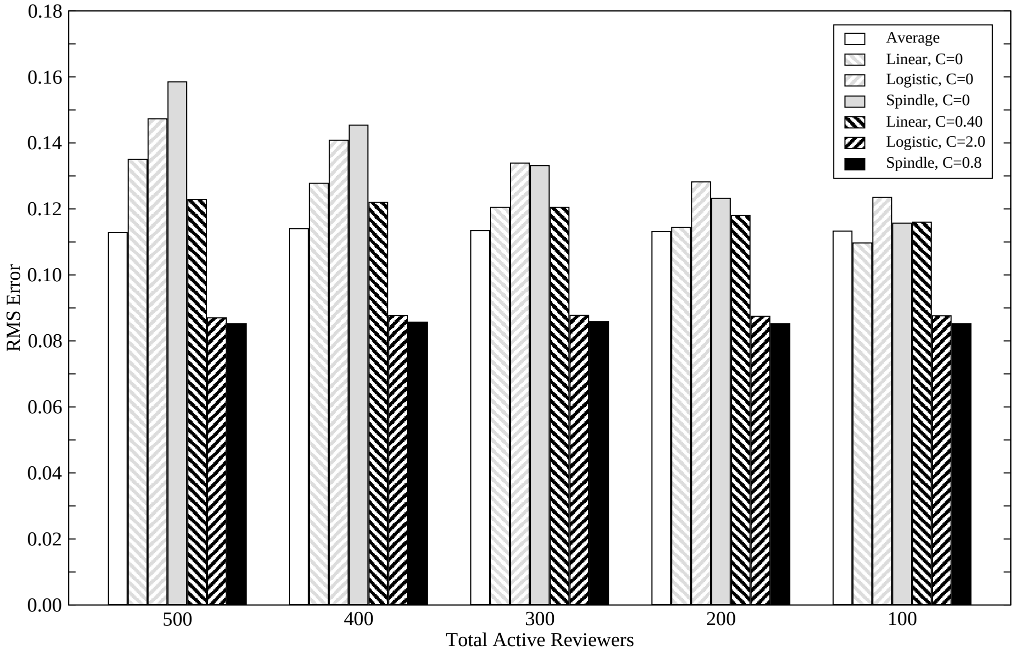

Figure 17: The root mean squared difference between the “correct” scores and the models’ scores as a function of the total number of active reviewers per trial, with 3 active reviewers assigned to each item.

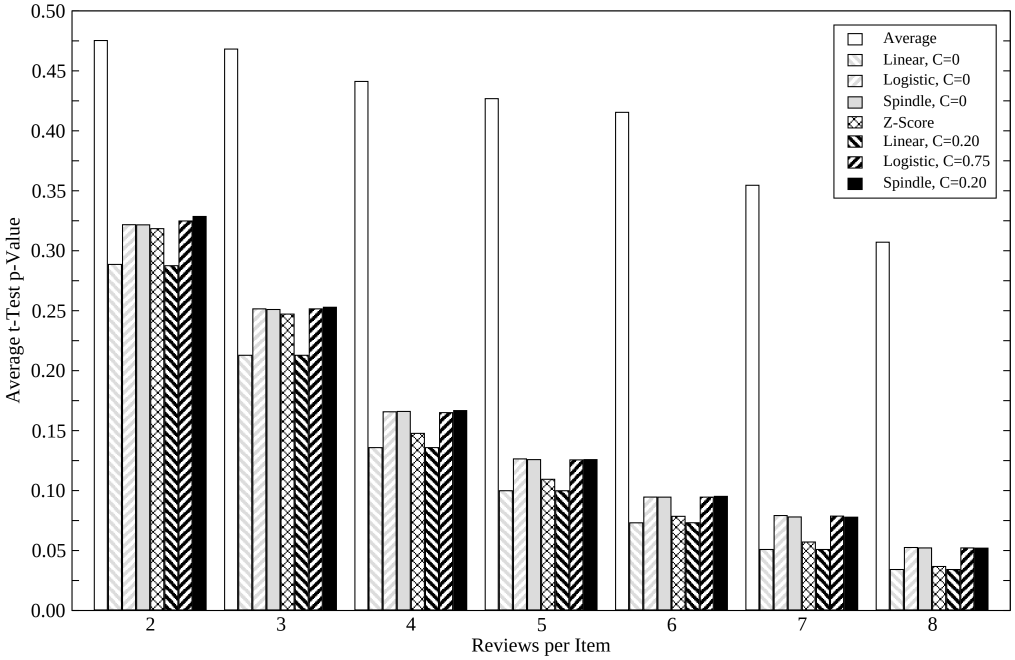

Figure 18: Average t-test p-value for two groups of items on the Artificial task.

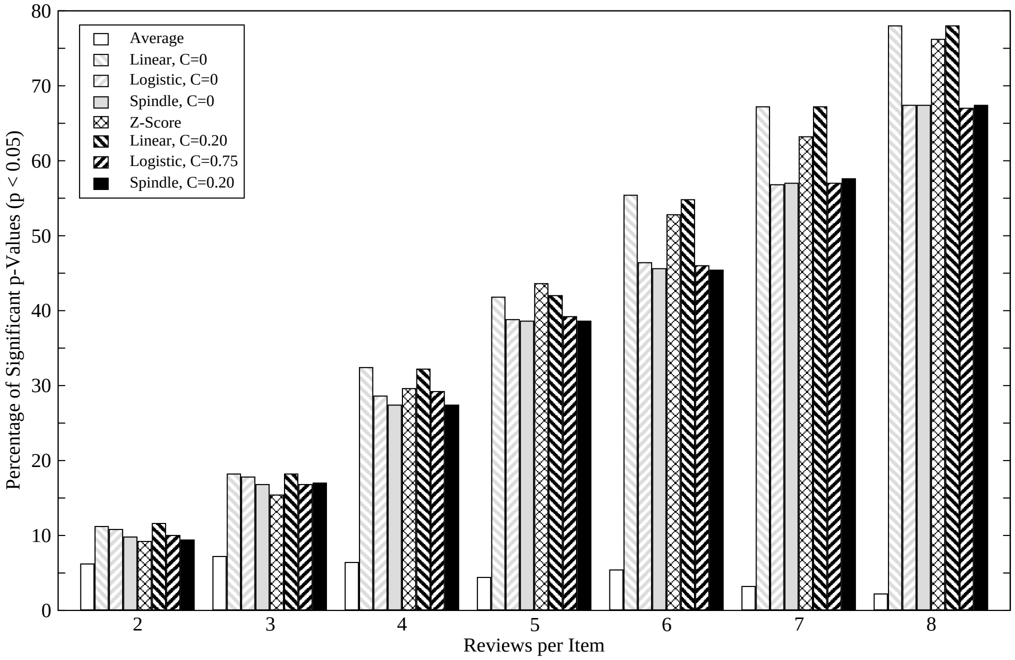

Figure 19: Percentage of trials in which the p-value was significant on the Artificial task.

groups in all trials.

The models were trained on a subset of the resulting ratings, 500 times per condition, with a varying number of ratings per item. In each trial, the models converted these ratings to adjusted ratings, as described above, and a t-test was performed to determine the p-value, indicating the probability that a single distribution would have generated two data sets whose means differ by as much as those in the sampled data. A lower p-value indicates a greater likelihood of a true difference between the groups. Figure 18 shows the average p-value for each model in each condition. Clearly the Average model, which is not doing any adjustment to the raw ratings, results in a much higher mean p-value than the other models. With eight reviews per item, the mean p-values of the other models are all well below the standard criterion of 0.05. There is relatively less difference between these models, although the Linear models consistently perform well.

Another measure of the models' ability to detect a significant difference is the percentage of the 500 trials on which the p-value fell below the 0.05 criterion, as shown in Figure 19. With the exception of the Average model, the percentage of detected differences all increase with additional data, with the Linear and Z-Score methods performing the best. Interestingly, the percentage of significant p-values for the Average model actually decreases with additional ratings. Therefore, the difference between the scores of the items in the two conditions, which is known to exist in this artificial task, cannot easily be detected unless the reviewer generosity is factored out.

Again, it is important to verify these findings in a more natural task. Therefore, in order to create two conditions for the Faces task, the images were partitioned on the basis of hair color into 23 blondes and 37 brunettes. The difference between these groups is small but significant in the full data set.[4] Figure 20 shows the average p-values as a function of the number of ratings per item. In this case, the Z-Score method does not perform as well as the others, especially with less data, and the Average model results in the least-significant p-values. There is little difference between the multivariate methods, although the Logistic model seems a bit better than the Linear or Spindle. These results are reflected in the percentage of significant p-values, shown in Figure 21. Similar findings are obtained on the Words task by comparing word pairs that look similar (having edit distance 3 or less) with pairs that do not, which tend to receive lower semantic similarity ratings. In that case, the Z-Score method was considerably worse than the Average model with fewer than 5 reviews per item.

6 Discussion

This paper has introduced several methods for improving the signal to noise ratio in ratings surveys by modeling and factoring out reviewer generosity. These methods can significantly improve the ability to recover from a limited body of data scores that are representative of the average ratings in a complete data set. Although, the models are most useful in cases where the data is sparse and two items may be rated by different subsets of the reviewers, they can improve the power of statistical tests on fully-balanced data sets by eliminating the noise introduced by generosity differences between reviewers.

One requirement for the success of these methods is that a reasonable number of ratings should be provided by each reviewer in order to accurately estimate his or her generosity. From our tests, it seems that as few as five ratings per reviewer may be sufficient if the reviewers are in reasonable agreement with one another in terms of the rank order of their ratings. More ratings will be needed if reviewers are likely to be employing different standards, as on the Movies task, or if the ratio of total items to reviewers is large. Using a stronger bias against extreme generosities and scores helps the models to behave appropriately when less data is available.

Aside from the most common approach of simply averaging the ratings for each item, the Z-Score method has been widely used to account for generosity differences. However, because it does not rely on any form of comparison between reviewers on common items, this method only works well if the items assigned to reviewers are well-distributed across the range of scores. As a result, the Z-Score method often results in poor performance unless there are a relatively large number of ratings per reviewer. Another drawback of this approach is that it produces z-scores, which occupy an unbounded scale, which may have to be transformed if bounded scores are needed. The Z-Score method performed better than the Average model on the artificial task, but consistently worse than the Linear and Spindle models. On the Faces task it performed quite well, but was worse on the Words task, especially with small amounts of data, and was particularly bad on the Movies task, even with large amounts of data. At the task of detecting significant group differences, the Z-Score method worked quite well on the artificial data, but was little better than the Average model on the Faces task, and worse on the Words task. This method is not consistently better than the Average model and, with more available data, can actually be significantly worse.

The Linear model improves upon the Z-Score method by obtaining estimates of reviewer generosity through indirect comparisons of their ratings. However, it rests upon the assumption of an unbounded rating scale. This sim-

[4] Yes, the blondes were considered more attractive, but only slightly.

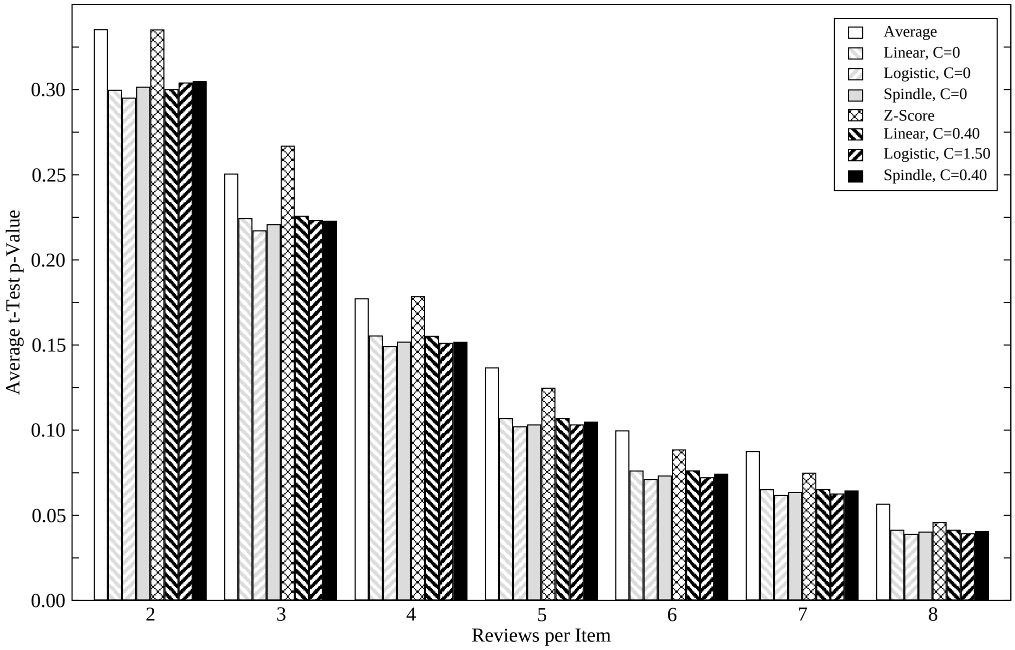

Figure 20: Average t-test p-value for two groups of items on the Faces task.

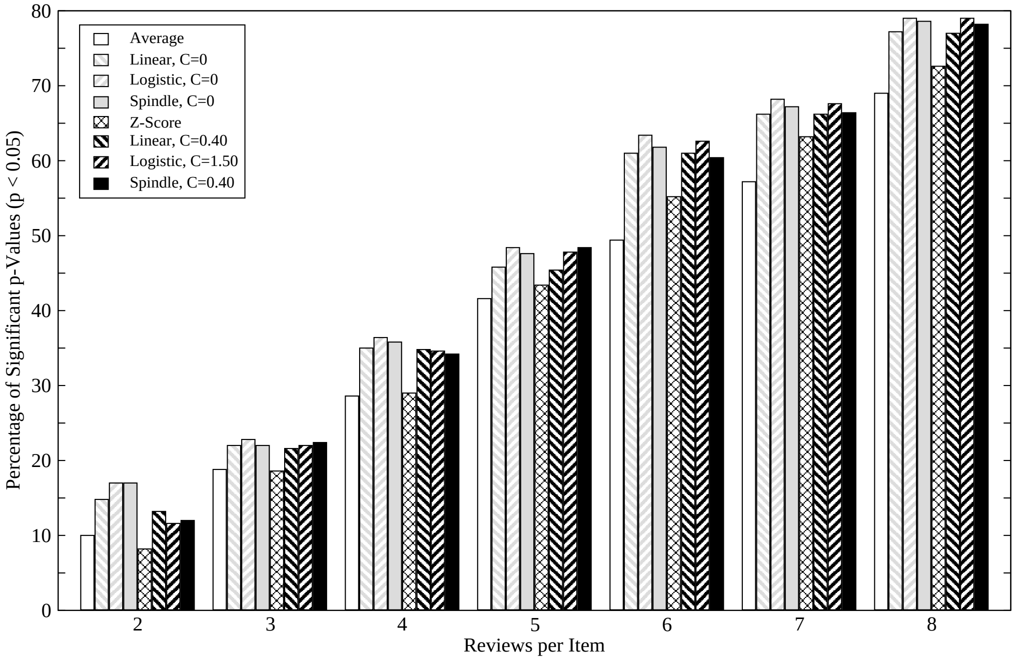

Figure 21: Percentage of trials in which the p-value was significant on the Faces task.

plifies the equations for training the model, but hinders its performance somewhat. On the artificial task, the Linear model was better than the Z-Score method, but not as good as the Spindle model. On the natural tasks, the unbiased Linear model performed better than the other unbiased models, according to RMS error, but the biased Linear model was significantly worse than the Logistic and Spindle models. The models all performed about equally based on rank error and inconsistency. In the task of detecting significant group differences, the Linear model, with or without bias, performed very well on the artificial task and quite well on the Faces task.

It is interesting that, although its assumptions are not appropriate for a bounded ratings scale, the Linear model actually performs reasonably well in many cases. This may be because the Linear model, when applied to a task with a bounded scale, has a natural bias against extreme generosities and scores. Consider a case in which the current generosity and score are quite high so the Linear model would predict a rating that is above the maximum of the scale. Because the actual rating will necessarily be lower than this, the model will re-estimate the generosity and/or the score to a lower value. A similar compression will occur with ratings at the bottom end of the scale. As a result, the scores and generosities will be less extreme. This is similar in influence to the extreme value cost, although its effect does not diminish with additional observations.

The Logistic and Spindle methods are based on the same non-linear model, but minimize different error functions. Under most circumstances, the two models have very similar performance and the biased Logistic model was even more consistent on the Words task and resulted in more easily detected group differences on the Faces task. However, there are certain conditions in which the Logistic model is considerably worse. It performed quite poorly in three of the four non-rounded artificial conditions. On the natural tasks, the unbiased Logistic model had high RMS error. The biased Logistic was as good as the Spindle model with less data, but slightly worse with more data.

On the artificial data, the unbiased Spindle model had the lowest RMS error and either the biased or unbiased Spindle model achieved the lowest rank, best-fit RMS error, and inconsistency. On the natural tasks, the unbiased Spindle model did not perform well with a small number of reviews per item. However, the biased Spindle model was quite good, with the lowest or close to the lowest RMS and rank error in all cases. Thus, although the biased Spindle model is not always the best model for all tasks, it is consistently quite good and can safely be used in many different situations.

The advantage of the Linear, Logistic, and Spindle models over the Average method was less apparent on the real data than it was on the artificial task. Presumably, this is because the real data includes sources of variance other than generosity or random noise, such as the fact that the reviewers might employ different standards in rating the items. On the Faces task, the reviewers' responses were moderately correlated and there was a moderate range of average scores across the items. On the Words task, the reviewers were even more in agreement with one another and there was a broader range of average scores. Finally, the reviewers did not agree strongly with one another on the Movies task and there was little variance in average review. Despite these differences, the results were fairly similar across the three domains.

Computational complexity is not a significant limitation of these models. Although they involve an iterative minimization process, the Linear, Logistic, and Spindle models all converge quite quickly and consistently, so the running time is not much worse than linear in the number of ratings. Because it uses a Newton-Raphson search, the Spindle model is somewhat slower than the Linear and Logistic methods, but this is essentially a constant factor and the difference is unlikely to be important for typical users. Another consideration is the ease of implementation. The Linear and Logistic models use simpler equations and are easier to code. Therefore, although they are not as reliable, the other models may be preferable to the Spindle model when it is expected that few items will have extremely high or low ratings.

Under most circumstances, the Linear, Logistic, and Spindle models are improved by the use of a reasonable extreme value penalty, C. An appropriate choice of penalty can reduce the sensitivity of the estimated scores to the potentially arbitrary assignment of reviewers to items and can substantially improve the match between the estimated scores and the true scores, in cases where those true scores are known. For the purpose of evaluating the effect of this parameter, it has been adjusted to appropriate levels for the various tasks. It is, of course, bad practice to adjust parameters on the basis of performance on the testing set, as we have done. In a real application, this would not be possible. However, our results on the tasks reported here suggest an independent method of selecting an appropriate value. One could plot a histogram of the average scores of the items, or the scores of the items according to an unbiased Spindle model, as in Figure 10. If the items have a broad distribution of scores, as in the Words task, a lower penalty, such as 0.25 or less, is appropriate. If the scores fall in a narrow distribution, as in the Movies task, a penalty of 0.5 to 1.0 may be the best choice. The Spindle and Logistic models are less sensitive than the Linear model to this bias and larger values can be used with little change in performance.

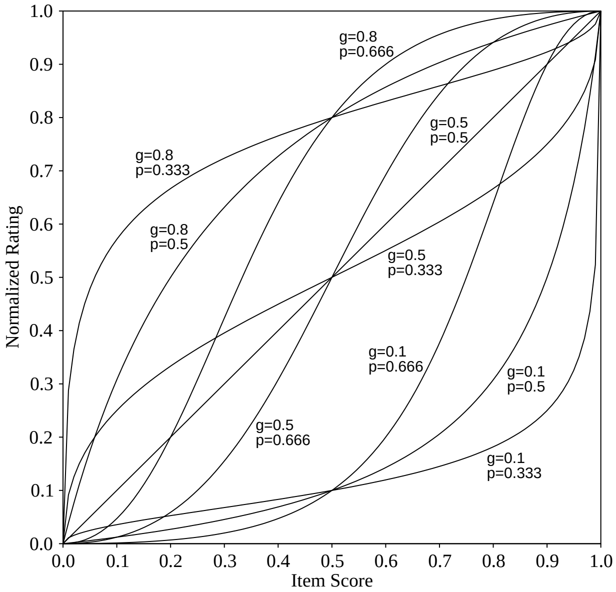

Figure 22: The rating an item will receive as a function of the item's score and the generosity, g, and polarity, p, of the reviewer according to Equation 25.

A working demonstration of the Spindle method is available at the following web address:

http://tedlab.mit.edu/∼dr/ReviewSmoother/

6.1 Related ideas

Both the Linear and Spindle models use a single parameter, generosity, to characterize the behavior of individual reviewers. In practice, there are other ways that reviewers may differ in how they apportion the ratings scale. Some reviewers, for example, are conservative and tend to use only the middle of the scale, while others prefer to use the extremes. These reviewers may not differ at all in their generosity. Rather, they differ in what we might call *polarity*. Reviewers with higher polarity are more likely to assign very high or low ratings. This idea can be captured quantitatively with a simple addition to the Spindle equation:

$$r'_{iv} = g'_v {s'_i}^{p'_v} \tag{25}$$

Here, p'_v is again a value between zero and infinity, which is related by the odds equation to the bounded value, p_v or *polarity*, which falls in the range $(0, 1)$. Figure 22 shows the predicted ratings for a variety of g_v and p_v values. With neutral generosity, a p_v above 0.5 results in an S-shaped curve. As a result, the ratings tend to be more extreme. A polarity below 0.5 results in a symmetric compression towards moderate ratings. With various choices of g_v and p_v, these equations can produce quite a range of monotonic functions and therefore may be capable of more accurately modeling the actual behavior of human reviewers.

Aside from the fact that all of the equations become more complex, the primary drawback of adding polarity to the model is that it doubles the number of mutually-dependent parameters that must be estimated. In order to accurately estimate these parameters, more data is needed. In the absence of sufficient data, the performance of the model may suffer. However, the Spindle model is most appropriate for and will most often be used with datasets having a limited number of ratings per reviewer. Therefore, it is not clear if the addition of polarity will actually improve the usefulness of the model.

Both the current models and the proposed polarity model assume that items can be characterized by a single underlying score and that reviewers' ratings merely reflect a monotonic transformation of this score. Although this is undoubtedly not accurate, it is an appropriate approach if one is primarily interesting in achieving a one-dimensional ranking of the items. However, there are cases in which ratings are based on a more complex underlying structure and our goal may be to uncover that structure. For example, the ratings on the Movies data were not highly correlated from one reviewer to the next. Presumably this is because movies differ on a wide range of dimensions. There are dramas, comedies, action movies, foreign language films, and they will differ along numerous dimensions, or factors, such as *humor*, *violence*, and *plot complexity*. Likewise, reviewers will differ in their appreciation of these factors, so their pairwise ratings may not be expected to vary monotonically.

In such a situation, we may wish to extract the hidden factors that best explain the variance in reviewer ratings. In doing so, the model will estimate not only the weighting of each factor for each item, but the appreciation of each reviewer for each of the factors. The following is a linear ratings model that reflects these ideas:

$$r_{iv} = \sum_f a_{vf} w_{if} \tag{26}$$

Where a_{vf} is the appreciation of reviewer v for factor f and w_{if} is the weighting of factor f on item i. The one free parameter in this model is the number of hidden factors. This model is closely related to established techniques of factor analysis (Harman, 1967; Neal & Dayan, 1997) and independent component analysis (Comon, 1994; Everson & Roberts, 1999). Such algorithms have been applied to ratings data under the guise of *collaborative filtering*, which involves predicting a reviewer's preferences given some of his or her ratings along with those of other reviewers (Canny, 2002).

Incidentally, this model is naturally implemented, along with non-linear sigmoidal filters to bound the ratings and factor weightings, in a standard feed-forward neural network with a single hidden layer. Each input unit represents an item, each output unit represents a reviewer, and each hidden unit represents a hidden factor. The factor weighting, w_{if}, is reflected in the output of the hidden unit f when input unit i is activated, and the reviewer appreciation, a_{vf}, takes the form of the weight from hidden unit f to output unit v. The activations of the output units reflect the ratings that the reviewers are expected to assign to the active item. Thus, a multilayer neural network can profitably be viewed as a device that performs hidden factor analysis.

Acknowledgements

This research was supported by NIMH NRSA 1-F32-MH65105-01. Correspondence regarding this article may be sent to Douglas Rohde (dr@tedlab.mit.edu), MIT, NE20-437E, 3 Cambridge Center, Cambridge, MA 02139.

References

Canny, J. (2002). Collaborative filtering with privacy via factor analysis. In *Acm conference on research and development in information retrieval.*

Comon. (1994). Independent component analysis—a new concept? *Signal Processing*, *36*, 287–314.

Everson, R., & Roberts, S. (1999). Independent Component Analysis: A flexible nonlinearity and decorrelating manifold approach. *Neural Computation*, *11*, 1957–1983.

Harman, H. H. (1967). *Modern factor analysis, 2nd edition.* Chicago, IL: University of Chicago Press.

Neal, R. M., & Dayan, P. (1997). Factor analysis using delta-rule wake-sleep learning. *Neural Computation*, *9*, 1781–1803.