

TWO METHODS FOR THE PRACTICAL ANALYSIS OF RATING DATA

Gunter Maris

Timo Bechger

CITO, NATIONAL INSTITUTE FOR EDUCATIONAL MEASUREMENT

ARNHEM

Citogroep

Arnhem, March 3, 2003

This manuscript has been submitted for publication. No part of this manuscript may be copied or reproduced without permission.

## Abstract

Managing human judgements is complicated by the fact that ratings rarely agree perfectly. This paper presents two methods to handle such disagreement. The first method is to remove inconsistent ratings and it will be shown that this enhances the quality of the data. Further assumptions about the relation between ability and assessability of subjects allows us to recover the information in the inconsistent ratings using a generalized partial credit model (GPCM). The use of the GPCM is illustrated with real data.

## 1. Introduction

Abilities such as speaking or writing are usually measured via human judgement. Managing human judgements is complicated, however, by the fact that ratings rarely agree perfectly and at present there is no generally accepted way to deal with inconsistent ratings. This paper is about practical ways to handle rating data.

It will be assumed that a *true rating* exists. Raters may fail to produce the true rating for a variety of reasons such as fatigue, failure to understand the intensions of the scoring guidelines, distraction due to matters such as poor handwriting, etc. The type of errors may be classified into five different categories:

1. *Rater severity of leniency* is a systematic tendency on the part of the rater to give a score that is somewhat higher or lower than appropriate.
2. *The halo effect* occurs when the judgement is based on an overall impression of the performance of the respondent.
3. *Rater preference for certain subjects*. Raters may assign higher ratings to subjects with certain characteristics.
4. *Central tendency or restriction of range* are similar in that in both cases the rater is not making full use of the scoring range. In the case of central tendency the rater rarely awards scores at the extremes, while in the case of restriction of range, the narrow band of scores awarded may be in any part of the range.
5. *Rater unreliability* refers to random variation in the assessment of judges.

Thus, inconsistencies between ratings contain both systematic (e.g., differences in severity) and stochastic elements (rater unreliability). To wit, it is quite difficult to say what is going on when raters disagree and it is difficult to devise an appropriate data model that includes parameters for individual raters.

There are a number of ways to promote appropriate ratings. We mention three

of them.

1. Halo effects and rater preferences for certain subjects may be avoided if raters are randomly assigned to subject-performance combinations and no background information about subjects is made known to the raters.
2. Raters may be trained to ensure that they fully understand the intentions of the test designers and be aware of the influences that can bias their judgement.
3. Information may be gathered about the consistency and appropriateness of ratings and used to advise raters so that they can improve their judgements.

However praiseworthy such efforts are, the effectiveness of rater training should not be over-estimated. Research shows that strong rater effects can persist even in the presence of stringent rater quality control efforts and attempts to train judges need not be successful (e.g., Raymond, Webb, & Houston, 1991; Stahl & Lunz, 1997). Thus, although one should always try to promote the quality of the ratings, one cannot do away with the need for a procedure to determine the final score given to respondents in such a way that the effect of rater errors is minimized. This is the main focus of the present paper.

There are basically two ways to deal with rating data. The first is to use a model that includes the raters and thus attempt to correct for rater bias (Linacre, 1994; Patz, Junker, & Johnson, 2000; Verhelst & Verstralen, 2001). The second is to use a single rating, assume that this rating is valid, and analyze the data as if the ratings are responses to ordinary test items. Both approaches have their advantages but neither has been very successful in practical applications. The present paper adopts the second approach but combines two independent ratings to form a single rating that is more likely to be the true rating.

It is assumed that respondents perform on a number of assignments and that

each performance is rated independently by two raters that are drawn randomly from a list of raters. Ideally, the number of raters on the list is sufficiently large to ensure that the probability that a single rater assesses two assignments of the same subject is small. Imagine, for instance, subjects taking an examination in order to demonstrate their ability to speak Dutch. Each assignment places the examinees in a different situation which requires them to explain something, to express their opinion, etc. The performance of the examinees may be recorded and sent to the raters that must each pronounce judgment on the examinee's performance. We further assume that each rater produces a vector of negative and positive judgments, which are coded as zero and one, respectively. The two judgements must then be combined to result in a single mark for each examinee. An application of this kind will be discussed in more detail below, and the data will be used as an illustration.

## 2. Filtering

In this section it is explained how additional ratings may be used to “filter” the data in such a way that the filtered ratings are more likely to equal the true rating.

Consider the performance of a subject on assignment  $i$ . Let  $X_{ir}$  denote the rating by the  $r$ -th judge,  $Y_i$  the true rating, and  $\theta$  the ability of the subject. It is assumed that

$$P(Y_i = 1|\theta) = \frac{\exp(2a_i [\theta - \delta_i])}{1 + \exp(2a_i [\theta - \delta_i])} \quad (1)$$

which is known as the *two-parameter logistic model (2PL)*. The parameter  $\delta_i$  represents the difficulty of the assignment, and  $a_i$  is a discrimination parameter. It is seen that the relation between the subjects ability and the true rating of his performance is stochastic. The probability of a positive assessment increases with increasing ability but even subjects that are very able may sometimes fail to perform well.

It is assumed that the subject's performance on each assignment is assessed

by two raters. Ideally, both raters produce the true rating of the performance in which case they agree and  $X_{i1} = X_{i2} = Y_i$ . It is further assumed that raters are randomly assigned to subjects and assignments. That is, the performance of subject  $p$  on assignment  $i$  is rated by two judges that are drawn at random from a list of raters. This implies that

$$P(X_{i1} = x_{i1}, X_{i2} = x_{i2}, Y_i = y_i | \theta) = P(X_{i1} = x_{i2}, X_{i2} = x_{i1}, Y_i = y_i | \theta)$$

which is called *partial exchangeability (PE) of raters*. Intuitively, PE follows because the probability that  $X_{i1}$  refers to the judgement of any rater is equal to the probability that  $X_{i1}$  refers to the rating of any other rater. Note that it was not assumed that all raters behave in the same way.

Under PE, it can be shown that

$$P(X_{i1} = Y_i | X_{i1} = X_{i2}, \theta) \geq P(X_{i1} = Y_i | \theta),$$

where the probabilities are equal in the case of perfect agreement or when the ratings were produced by tossing coins. Thus, the probability that consistent ratings correspond to the true quality is at least as large as the probability that the ratings of either rater correspond to the true rating. This result justifies that inconsistent assessments are ignored.

The proof is as follows: First, note that if two raters disagree, one of the ratings equals  $Y_i$ . Under PE, it is equally likely that this is the first rating or the second rating and it follows that

$$P(X_{i1} = Y_i | X_{i1} \neq X_{i2}, \theta) = \frac{1}{2}.$$

This implies that

$$\begin{aligned}
P(X_{i1} = Y_i|\theta) &= P(X_{i1} = Y_i|X_{i1} \neq X_{i2}, \theta)P(X_{i1} \neq X_{i2}|\theta) \\
&+ P(X_{i1} = Y_i|X_{i1} = X_{i2}, \theta)P(X_{i1} = X_{i2}|\theta) \\
&= \frac{1}{2}[1 - P(X_{i1} = X_{i2}|\theta)] + P(X_{i1} = Y_i|X_{i1} = X_{i2}, \theta)P(X_{i1} = X_{i2}|\theta) \\
&= \frac{1}{2} + \left[ P(X_{i1} = Y_i|X_{i1} = X_{i2}, \theta) - \frac{1}{2} \right] P(X_{i1} = X_{i2}|\theta).
\end{aligned}$$

It follows that

$$\begin{aligned}
P(X_{i1} = Y_i|\theta) - \frac{1}{2} &= \left[ P(X_{i1} = Y_i|X_{i1} = X_{i2}, \theta) - \frac{1}{2} \right] P(X_{i1} = X_{i2}|\theta) \\
&\leq P(X_{i1} = Y_i|X_{i1} = X_{i2}, \theta) - \frac{1}{2}.
\end{aligned}$$

This implies, finally, that

$$P(X_{i1} = Y_i|\theta) \leq P(X_{i1} = Y_i|X_{i1} = X_{i2}, \theta)$$

which had to be proven. If the raters were behaving as if they were throwing coins,

$$P(X_{i1} = Y_i|\theta) = P(X_{i1} = Y_i|X_{i1} = X_{i2}, \theta) = \frac{1}{2}.$$

Schematically, the data can be depicted as in the following table:

$X_{i1}$	$X_{i2}$	$X_i^*$	$C_i$
1	1	1	1
1	0	<i>missing</i>	0
0	1	<i>missing</i>	0
0	0	0	1

where  $X_i^*$  represents the rating when raters are consistent, and  $C_i$  indicates whether the raters were consistent. In practice, we equate the consistent rating with the true

quality and consider the 2PL (1) for the consistent data, that is,

$$P(X_i^* = 1|C_i = 1, \theta) = \frac{\exp(2a_i[\theta - \delta_i])}{1 + \exp(2a_i[\theta - \delta_i])}. \quad (2)$$

The consistent judgements may then be used to estimate the subject's ability using software that allows for incomplete data. The larger the number of consistent ratings for a subject, the better his or her ability can be determined. This procedure entails a loss of information to the extent that the knowledge of whether ratings were consistent or not contains information about ability. The relation between  $C_i$  and ability is the subject of the ensuing section. It will be shown that  $C_i$  and  $X_i^*$  can be analyzed simultaneously at the cost of an additional assumption about the relation between  $C_i$  and  $\theta$ . Any information that remains in the identity of the rater is left unused because it would require a sound understanding of the behavior of individual raters. In general, however, we cannot presume to have such understanding.

### 3. Consistency and Ability

Whether two raters agree depends upon the raters themselves, the assignments and the ability of the respondent. It is reasonable to assume that raters are likely to agree when a performance is extremely good or extremely bad and inconsistencies will occur mainly when raters assess average performances. These considerations lead us to model the probability  $P(C_i = 1|\theta)$  as a “single-dipped” function of  $\theta$ ; namely as,

$$P(C_i = 1|\theta) = \frac{1 + \exp(2a_i[\theta - \delta_i])}{1 + \exp(a_i[\theta - \lambda_i - \delta_i]) + \exp(2a_i[\theta - \delta_i])} \quad (3)$$

which equals the *collapsed partial credit model* proposed by Verhelst and Verstralen (1993), albeit in a different parameterization. An illustration is given in Figure (1). The probability  $P(C_i = 1|\theta)$  may be interpreted as a measure of the *assessability* of subjects with ability  $\theta$ . It achieves its lowest value at  $\theta = \delta_i$ , and  $\delta_i$  represents the

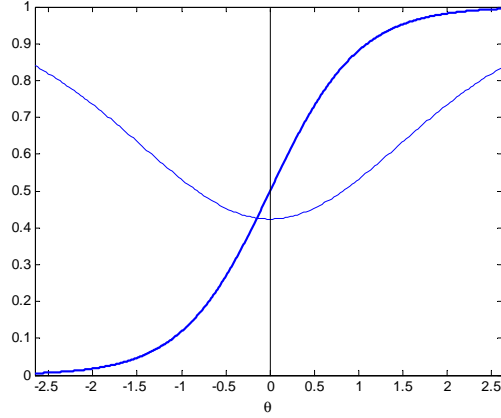


FIGURE 1.

The thick line is the plot of  $P(X_i^* = 1|C_i = 1, \theta)$ . The thin line is a plot of  $P(C_i = 1|\theta)$ . Indicated in the figure is the position of the difficulty of the item,  $\zeta_i$ .

difficulty of the assignment.

We now proceed by combining  $C_i$  and  $X_i^*$  in a *generalized partial credit model* (GPCM). For each subject, the ratings have as possible outcomes: negative consistent ( $X_i^* = 0$  and  $C_i = 1$ ), inconsistent ( $C_i = 0$ ), and positive consistent ( $X_i^* = 1$  and  $C_i = 1$ ). The probabilities corresponding to each outcome are:

$$\begin{aligned}
 P(X_i^* = 0, C_i = 1|\theta) &= P(X_i^* = 0|C_i = 1, \theta)P(C_i = 1|\theta) \\
 &= \frac{1}{1 + \exp(2a_i[\theta - \delta_i])} \frac{1 + \exp(2a_i[\theta - \delta_i])}{1 + \exp(a_i[\theta - \lambda_i - \delta_i]) + \exp(2a_i[\theta - \delta_i])} \\
 &= \frac{1}{1 + \exp(a_i[\theta - \lambda_i - \delta_i]) + \exp(2a_i[\theta - \delta_i])} \\
 P(C_i = 0|\theta) &= \frac{\exp(a_i[\theta - \lambda_i - \delta_i])}{1 + \exp(a_i[\theta - \lambda_i - \delta_i]) + \exp(2a_i[\theta - \delta_i])} \\
 P(X_i^* = 1, C_i = 1|\theta) &= \frac{\exp(2a_i[\theta - \delta_i])}{1 + \exp(a_i[\theta - \lambda_i - \delta_i]) + \exp(2a_i[\theta - \delta_i])}
 \end{aligned}$$

These are precisely the probabilities under a GPCM for an item with three categories where the categories correspond to different values of  $X_{i1} + X_{i2}$ . The

relation between  $X_{i1} + X_{i2}$ ,  $C_i$ , and  $X_i^*$  is summarized in the following table:

$X_{i1}$	$X_{i2}$	$X_i^*$	$C_i$	$X_{i1} + X_{i2}$
1	1	1	1	2
1	0	<i>missing</i>	0	1
0	1	<i>missing</i>	0	1
0	0	0	1	0

It is seen, for instance, that  $X_{i1} + X_{i2} = 0$  when the ratings are both negative. Thus, we may proceed by analyzing  $X_{i1} + X_{i2}$  with the GPCM using any of the software packages for this model.

Care must be taken in the interpretation of the parameters of the GPCM. The parameter  $\lambda_i$  determines the *assessability* of the assignment and its value should be as high as possible; when  $\lambda_i \rightarrow \infty$ , the probability of a consistent response becomes unity while it becomes zero if  $\lambda_i \rightarrow -\infty$ . If  $\lambda_i \leq 0$ , it can be shown that  $[\delta_i + \lambda_i, \delta_i - \lambda_i]$  is the *inconsistency interval* where the probability of an inconsistent response is greater than the probability of either of the consistent responses (see Figure 2). When  $\lambda_i > 0$  this doesn't occur and the inconsistency interval can be considered empty. The discrimination parameter  $a_i$  determines the steepness of the curve. Further illustration is provided by Figure (3) which shows how the value of the different parameters affect assessability as a function of ability. It is seen in Figure (3) that changing the difficulty parameter,  $\delta_i$ , simply shifts the curves over the  $\theta$ -axis. Finally, note that if the ratings are perfectly consistent  $Y_i$  is observed and the model is given by Equation (1). On the other hand, if the ratings are tosses of a coin,  $\lambda_i \rightarrow \infty$ , and  $P(C_i = 1|\theta) \rightarrow 0.5$  if  $a_i = \log(2)/(-\lambda_i)$ .

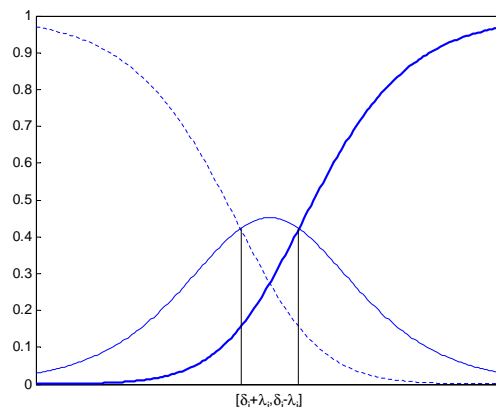


FIGURE 2.

Plot of the probability of an inconsistent response, the probability of a positive consistent response, and the probability of a negative consistent response. The location of the inconsistency interval is indicated.

#### 4. An Application

The examination of Dutch as a second language is intended to measure the ability to use the Dutch language in everyday situations. The examination is taken at two levels, with level one intended for those who wish to work, and level two intended for those who wish to enter higher education. The ability to speak, write, listen, and read are examined separately. For our present purpose we focus on the ability to speak, which is graded by human raters.

The examinations consist of a small number of assignments; write a letter, talk about a day at school, etc. Each performance is graded on a varying number of subscales or *aspects* such as content, sentence construction, etc. by field experts. Since a rater assesses a single performance on multiple aspects there may be dependencies between the ratings that are due to the rater. We therefore conducted analyses on separate aspects.

Each examinee carries out each assignment and his/her performance is graded independently by only two raters. Currently, raters are assigned to examinees in

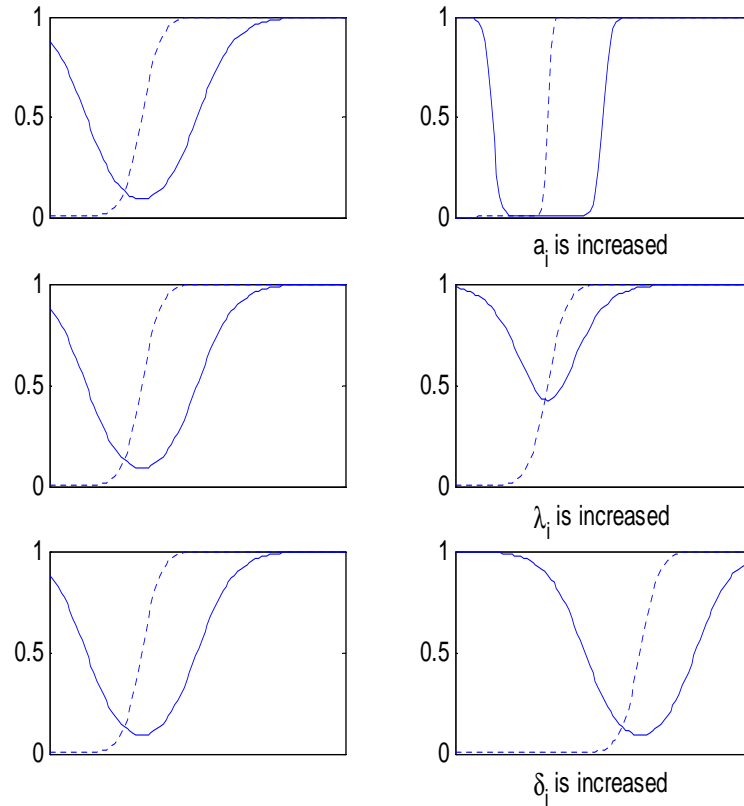


FIGURE 3.

The dashed line is the plot of  $P(X_i^* = 1 | C_i = 1, \theta)$ . The solid line is a plot of  $P(C_i = 1 | \theta)$ . The plots show the effect on the curve if the values of individual parameters are changed.

a systematic way according to availability. Furthermore, each rater assesses the performance of a candidate on each assignment in the examination so that the assumption that raters are assigned randomly to assignments is clearly violated. The reason for this is that the examinees oral performance is recorded on tape at this time and it is unfeasible in practice to record performance on different assignments on different tapes. Our present purpose is to analyze the data as an illustration. We treat the data as if the assignment of raters was random and we have conducted separate analyses for the aspects content and choice of words. These aspects were

chosen because we had enough data.

We have collected the data from several examinations that were administered from 1992 on to 2002. Only the assessments that were dichotomous have been selected. The resulting datasets are incomplete and very large; at level 1  $n \approx 24,000$ , at level 2  $n \approx 38,000$ . Although the data were incomplete there was enough overlap between examinations to allow conditional maximum likelihood estimation. Unfortunately, the overlap was too small for level 2 which forced us to focus on level 1. Half of the data (*the calibration data*) was used to calibrate the model and the remaining half (*the test data*) was used to validate the model.

The model that was used to calibrate the data was the PCM which equals the GPCM with equal discrimination parameters. The parameters were estimated using the method of conditional maximum likelihood with the OPLM software package. The calibration data were used to estimate the parameters and remove unfitting items. The test data are used to test the model. Estimation and testing are done with separate data as a safeguard against capitalization on chance. Note that OPLM uses a different parameterization of the GPCM than the one used here. When  $\beta_1$  and  $\beta_2$  denote the parameters from OPLM, the assessability and difficulty parameters are

$$\delta_i = \frac{1}{2} [\beta_1 + \beta_2],$$

and

$$\lambda_i = \beta_1 - \delta_i,$$

respectively.

Model fit was assessed with the  $R_{1c}$  statistic that is provided by OPLM. For “content”, the  $R_{1c}$  was 568.1 with 473 degrees of freedom which had an exceedance

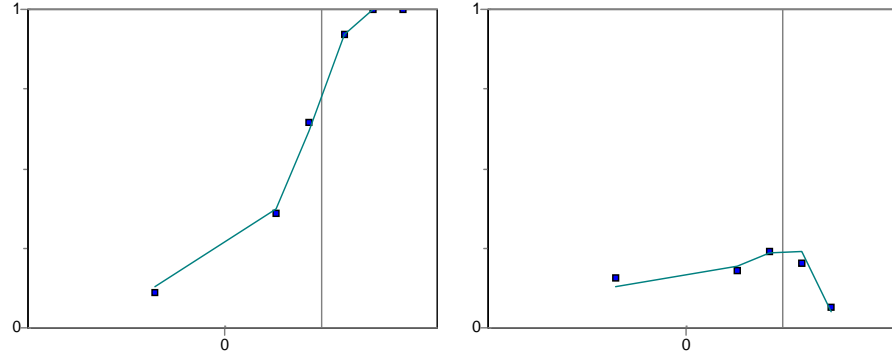


FIGURE 4.

Plots of  $P(X_i^* = 1|C_i = 1, S = s)$  (left panel) and  $P(C_i = 0|\theta)$  (right panel).

probability of 0.002 under the model. For “word choice”, the  $R_{1c}$  was 554.5 with 462 degrees of freedom which had an exceedance probability of 0.002 under the model. The number of observations per assignment ranged from 266 to 3249.

To assess the fit of the model to individual assignments we used plots of observed and expected probabilities as a function of test score. Specifically, we produced separate plots for  $P(X_i^* = 1|C_i = 1, S = s)$  and  $P(C_i = 0|S = s)$ , where  $s$  denotes the sum score. Scores were grouped and expected probabilities were calculated for the mean in each group. Thus, in Figure 4 the line connects the probabilities expected under the PCM while the boxes are the observed probabilities. We have indicated the location of the cut-off score that is used in the examination. It is especially here that assessability should be high.

In general, the fit of the PCM was quite reasonable especially in view of the fact that the data were not collected under the assumption of random assignment of raters. Figure 4 shows a plot for one of the few less fitting assignments.

There were quite a few assignments that were difficult to assess in the vicinity of the cut-off score. An illustration is given in Figure (4). It is interesting to note that, in general, word choice was seen to be less assessable than content.

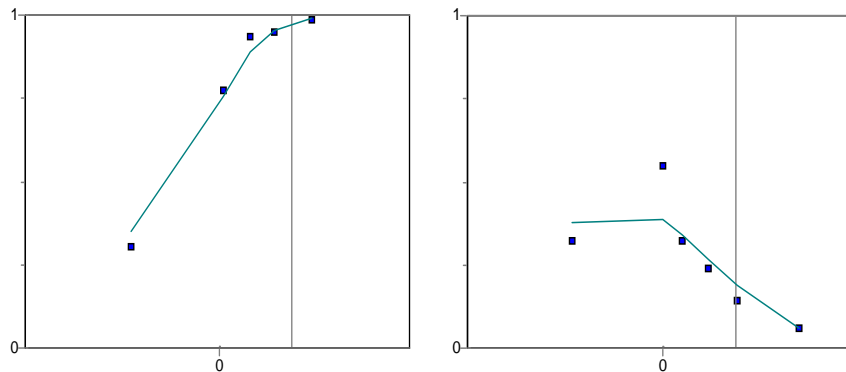


FIGURE 5.

Plots of  $P(X_i^* = 1 | C_i = 1, S = s)$  (left panel) and  $P(C_i = 0 | \theta)$  (right panel).

## 5. Discussion

We have started our account by showing that consistent ratings are more likely to reflect the true rating so that the quality of the data is enhanced if inconsistent ratings are ignored. The number of consistent judgements will differ across subjects since some are more assessable than others which means that a greater effort is needed to measure the ability of subjects that are less assessable. We have found, however, that information provided by inconsistent ratings may still be made to use if the ratings are summed and the GPCM is used as a model for the summed ratings. When the GPCM fits the data it provides a useful tool for test construction and data analysis. To assess the fit of the GPCM we have used plots of predicted and observed probabilities that were tailored for the present purpose. These plots are affected by sampling error and it is a topic for future research to add confidence bounds around the predicted probabilities.

For those items whose assessability can not be modeled with the GPCM one would like to take refuge to the filtering procedure and use only the consistent ratings. Unfortunately, standard software fails when there are many items with this defect and many subjects. It is a topic for future research to construct software that

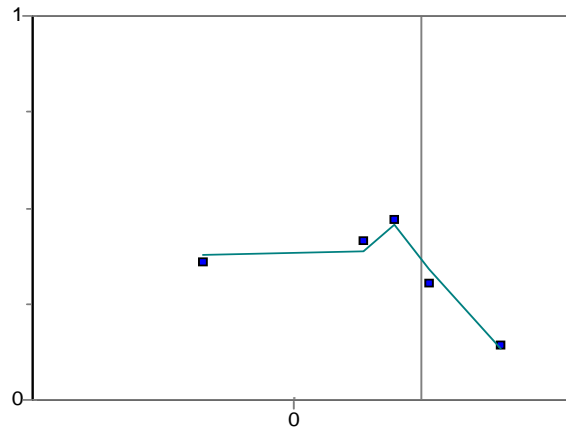


FIGURE 6.

Plot of  $P(C_i = 0 | \theta)$  for an item where the probability of disagreement was close to a half and fairly constant.

allows for large incomplete datasets.

It is important to note that the quality of the ratings remains a concern. In the application we have seen that there are many assignments that were difficult to assess in the vicinity of the cut-off score. Furthermore, when there are more consistent ratings, the amount of data increases when filtering is used, and the fit of the GPCM is expected to improve.

Finally, we note that the present procedure is subject to the condition that raters are assigned randomly to both subjects, and assignments which may complicate the logistic handling of the testing. We nevertheless think that it is important. Not just because it justifies the assumption that raters are PE but also because it helps to diminish rater biases.

## References

- Linacre, J. M. (1994). *Many-facetted Rasch measurement* (2nd ed.). Chicago: Mesa Press.
- Patz, R. J., Junker, B. W., & Johnson, M. S. (2000). *The hierarchical rater model for rated test items and its application to large scale educational assessment data* (Tech. Rep. No. 712). Department of Statistics, Carnegie Mellon University.
- Raymond, M. R., Webb, L., & Houston, W. (1991). Correcting performance rating errors in oral examinations. *Evaluation and the Health professions, 14*, 100-122.
- Stahl, J. A., & Lunz, M. (1997). Judge performance reports: Media and message. In G. J. Engelhard & M. Wilson (Eds.), *Objective measurement: Theory into practice*. Norwood NJ: Ablex Press.
- Verhelst, N., & Verstralen, H. H. F. M. (1993). A stochastic unfolding model derived from the partial credit model. *Kwantitatieve Methoden, 43*, 73-92.
- Verhelst, N. D., & Verstralen, H. H. F. M. (2001). IRT models with multiple raters. In A. Boomsma, M. A. J. Van Duijn, & T. A. B. Snijders (Eds.), *Essays on item response theory* (p. 89-106). New York: Springer Verlag.