

November 2015

A New Extension of the Binomial Error Model for Responses to Items of Varying Difficulty in Educational Testing and Attitude Surveys

James A. Wiley
University of California San Francisco

John Levi Martin
University of Chicago

Stephen J. Herschkorn
CUNY College of Staten Island

Jason Bond
Alcohol Research Group

[How does access to this work benefit you? Let us know!](#)

Follow this and additional works at: http://academicworks.cuny.edu/si_pubs

Recommended Citation

Wiley, J. A., Martin, J. L., Herschkorn, S. J. & Bond, J. (2015). A New Extension of the Binomial Error Model for Responses to Items of Varying Difficulty in Educational Testing and Attitude Surveys. *PLoS ONE*, 10(11), e0141981. doi:10.1371/journal.pone.0141981.

This Article is brought to you for free and open access by the College of Staten Island at CUNY Academic Works. It has been accepted for inclusion in Publications and Research by an authorized administrator of CUNY Academic Works. For more information, please contact AcademicWorks@cuny.edu.

RESEARCH ARTICLE

A New Extension of the Binomial Error Model for Responses to Items of Varying Difficulty in Educational Testing and Attitude Surveys

James A. Wiley¹, John Levi Martin^{2*}, Stephen J. Herschkorn³, Jason Bond⁴

1 Department of Family and Community Medicine and Institute for Health Policy Studies, School of Medicine, University of California San Francisco, San Francisco, California, United States of America, **2** Sociology Department, University of Chicago, Chicago, Illinois, United States of America, **3** Department of Mathematics, College of Staten Island, New York City, New York, United States of America, **4** Alcohol Research Group, Emeryville, California, United States of America

* jlmartin@uchicago.edu



Abstract

We put forward a new item response model which is an extension of the binomial error model first introduced by Keats and Lord. Like the binomial error model, the basic latent variable can be interpreted as a probability of responding in a certain way to an arbitrarily specified item. For a set of dichotomous items, this model gives predictions that are similar to other single parameter IRT models (such as the Rasch model) but has certain advantages in more complex cases. The first is that in specifying a flexible two-parameter Beta distribution for the latent variable, it is easy to formulate models for randomized experiments in which there is no reason to believe that either the latent variable or its distribution vary over randomly composed experimental groups. Second, the elementary response function is such that extensions to more complex cases (e.g., polychotomous responses, unfolding scales) are straightforward. Third, the probability metric of the latent trait allows tractable extensions to cover a wide variety of stochastic response processes.

OPEN ACCESS

Citation: Wiley JA, Martin JL, Herschkorn SJ, Bond J (2015) A New Extension of the Binomial Error Model for Responses to Items of Varying Difficulty in Educational Testing and Attitude Surveys. PLoS ONE 10(11): e0141981. doi:10.1371/journal.pone.0141981

Editor: Fabio Rapallo, University of East Piedmont, ITALY

Received: June 8, 2015

Accepted: October 15, 2015

Published: November 6, 2015

Copyright: © 2015 Wiley et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: All relevant data are within the paper.

Funding: The authors have no support or funding to report.

Competing Interests: The authors have declared that no competing interests exist.

Introduction

In this paper we introduce a class of item response models for the analysis of response distributions derived from survey data. The simplest item response function in this class is a generalization of the binomial error model for ability testing of Keats and Lord [1]. Similar item response functions were considered briefly by Lazarsfeld [2] and Coleman [3] but not implemented as models for response data. The distinguishing characteristic of the approach taken here is that the probability of an item response is written as a function of a latent variable that can also be interpreted as a probability. The choice of a probability metric suggests the Beta density as a natural choice to model the distribution of the latent variable. Furthermore, a response function formulated in this way is easy to modify to accommodate variations in the nature and complexity of response tasks.

We first introduce the extended binomial error (EBE) model as a generalization of the binomial error model. We discuss issues of identifiability, estimation and fit, and give two examples, one for a single sample, and another for data from two independent samples, noting the similarities in fit between the extended binomial error and loglinear Rasch models. Finally, we discuss extensions for polychotomous data and for modeling responses produced by non-cumulative response functions.

The resulting class of models have special utility for the investigation of substantively important questions about the nature of response (as opposed to scoring long tests), Further, in contrast to the loglinear Rasch models that have been most influential in sociology due to the work of Duncan [4] in particular, this approach allows us to make strong assertions about the social *distribution* of the latent trait.

The Extended Binomial Error Model

Introducing item difficulty into the binomial error model for dichotomous responses

The binomial error model for test scores was introduced by Keats and Lord [1] [5] as a strong true score theory for dichotomous test items of more or less equal difficulty. This model is based on the assumption that the conditional distribution of the observed score given the true score (conceived as a theoretical “proportion correct” measure) is binomial, with the true score playing the role of the constant probability of a correct answer in M independent trials where M is the number of items. With no additional specifications, Keats and Lord were able to show that the first M moments of the true score distribution can be determined from the moments of the observed score distribution. Furthermore, they showed [6] that a linear regression of true on observed score implies and is implied by a negative hypergeometric distribution of the observed scores. They also proved that a two-parameter Beta distribution for the true scores implies a linear regression of true score on observed score and a negative hypergeometric distribution for the observed scores.

Of course, it is infrequent that we analyze a set of dichotomous items with the same levels of difficulty (see also [7]); hence Keats and Lord indirectly incorporated variations in item difficulties in their compound binomial model, which accordingly lost many of the most attractive features of the binomial error model. Some other modifications of the binomial error assumptions have been proposed [8], but these are not often implemented as IRT models.

We propose to modify the binomial error model by direct incorporation of item-difficulty parameters and by assuming a bounded proportion-like latent trait. The model assumptions allow a compact expression for the probability of any observed response pattern in terms of item-difficulty parameters and the parameters of the distribution of the latent trait, assumed to be two-parameter Beta. This expression leads in turn to an algorithm for simultaneous ML estimation of both sets of parameters. In contrast to a Rasch model, therefore, our estimation of item parameters is not separable from the estimation of person parameters; the advantage of specifying a flexible family of distributions like the Beta comes in the capacity to easily analyze factorial experiments in which either the trait, or item hardnesses, or both may be affected by treatments, as we show below.

A Generalization

Let $\mathbf{x} = [x_1, x_2, \dots, x_M]$ be a response vector for a set of M dichotomous items, coded so that $x_i = 1$ implies completion of a task related to some underlying ability or acceptance of a statement consistent with some hypothesized underlying attitude dimension and $x_i = 0$ otherwise. (The 1/

0 coding of responses is arbitrary but useful in writing down the probabilities associated with complete response patterns.) Consider a latent variable y , where $0 < y < 1$, and let the difficulty of the i^{th} item be k_i ($k_i > 0$ for all i). We propose the simple response function

$$\Pr[x_i = 1|y] = y^{k_i} \tag{1}$$

The conditional probability of any set of M responses under conditional independence is

$$\Pr[\mathbf{x}|y] = \prod_{i=1}^n (y^{k_i})^{x_i} (1 - y^{k_i})^{(1-x_i)} \tag{2}$$

and the unconditional probability is

$$\Pr[\mathbf{x}] = \int \Pr[\mathbf{x}|y]\varphi(y)dy \tag{3}$$

where $\varphi(y)$ is the density of y .

There are a number of advantages to constructing the model using the form of Eq 1, as the latent trait may be interpreted as the probability of answering standardized ($k_i = 1$) item in a positive direction. However, a complication arises for the interpretation of standard errors of the k parameters given that they must be strictly non-negative. As is the case for other models with necessarily non-negative parameters (such as models for variances), the standard errors cannot be interpreted symmetrically; indeed, in cases of very poor fit, standard errors may imply that the true population value might quite plausibly be negative. When the model fits well, these issues are minor, but it makes sense to carry out statistical tests on the equivalence of k parameters not by comparison of standard errors, but by the comparison of chi-squares of nested models constraining or not constraining parameters to be equal. If the use of confidence intervals is required, one can rewrite the item response function as a double-exponential whose argument is the difference between an unbounded latent variable ξ and an unbounded item difficulty parameter χ_i , where $\xi = -\log[-\log(y)]$ and $\chi_i = \log(k_i)$, $\xi, -\infty < \xi, \chi_i < +\infty$. Thus we can write

$$\Pr[x_i = 1|\xi] = \exp[-(\exp(-(\xi - \chi_i)))] \tag{4}$$

We note, however, that in contrast to a Rasch model, Eq 4 does not imply equiprobability when $\xi = \chi_i$; instead, $\Pr[x_i = 1|\xi] = .5$ when $(\xi - \chi_i) = -.3665$. Given that Eq 1 has the more intuitive relation to a probability statement, we prefer this for the development of a family of models for stochastic response.

The choice of a probability metric suggests the Beta density as a natural choice to model the distribution of the latent variable, as it, like a probability, is defined for the interval [0,1]. Thus we propose:

$$\varphi(y) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} y^{(a-1)}(1-y)^{(b-1)} \tag{5}$$

where $a, b > 0$, $0 < y < 1$, and Γ is the complete gamma function. (We note that such a specification was suggested but not implemented by Lazarsfeld [2].) Note that $E[y] = a/(a+b)$; $V[y] = ab/[(a+b)^2(a+b+1)]$. For $k_i = k$ for all i , this leads to the binomial error model with a Beta density for the latent trait, sometimes called the beta-binomial model [7].

It can be shown that given this density, for any $k > 0$,

$$E[y^k] = \frac{\Gamma(a+b)\Gamma(a+k)}{\Gamma(a)\Gamma(a+b+k)} \tag{6}$$

Thus, for example, the unconditional probability of the unit response vector $\mathbf{x} = [1, 1, 1, \dots, 1]$ is simply

$$\Pr[1, 1, \dots, 1] = E[\Pr[1, 1, \dots, 1|y]] = E \left[y^{\sum_{i=1}^M k_i} \right] = \frac{\Gamma(a+b)\Gamma(a + \sum_{i=1}^M k_i)}{\Gamma(a)\Gamma(a+b + \sum_{i=1}^M k_i)} \tag{7}$$

More generally, any $\Pr[\mathbf{x}]$ may be written as a function of gamma functions whose arguments are the sums of the distribution parameters a and b and/or item difficulties k_i .

Representation of the Unconditional Response Probabilities

Let $\Pr[\mathbf{x}]$ represent the unconditional probability of a response. For any \mathbf{x} , let $I(\mathbf{x}) = \{i: x_i = 0\}$ and $C(\mathbf{x}) = \{i: x_i = 1\}$. Let A be any subset of $I(\mathbf{x})$, including the null set \emptyset and let $|A|$ be the number of elements in A . Given these definitions, we show in Appendix A that

$$\Pr[\mathbf{x}] = \Pr[x_1, x_2, \dots, x_M] = \sum_{A \subset I(\mathbf{x})} (-1)^{|A|} E \left[\prod_{i \in AUC(\mathbf{x})} \Pr[x_i = 1|y] \right] \tag{8}$$

with

$$E \left[\prod_{i \in AUC(\mathbf{x})} \Pr[x_i = 1|y] \right] = \frac{\Gamma(a+b)\Gamma(a + \sum_{i \in AUC(\mathbf{x})} k_i)}{\Gamma(a)\Gamma(a+b + \sum_{i \in AUC(\mathbf{x})} k_i)} \tag{9}$$

for the EBE model. These equations are useful for constructing joint item response functions in programming routines for ML estimation.

Initial estimates for the model parameters may be obtained from bivariate cross-classifications. Given the resulting table from any such cross-classification of the i^{th} and j^{th} items, and fixing one of the two relevant item parameters (say $k_i = 1$), the other three parameters k_j , a and b can be estimated from the three degrees of freedom in the 2×2 table. Repeating this $M-1$ times for all cross classifications of the i^{th} item with all other items produces a complete vector of starting values. (We give the details in [Appendix B](#).) We then use the Polak and Ribière [9] version of the Fletcher and Reeves conjugate gradient method [10] to find maximum likelihood estimates; we also use the Nelder and Mead [11] simplex methods to check that there are no preferable solutions in the area of the start values. In no case did we find that our maximum was local only.

Scores and Score Variances from the Posterior Distribution of y given \mathbf{x}

The posterior density of y given \mathbf{x} is defined as

$$h[y|\mathbf{x}] = \frac{\Pr[\mathbf{x}|y]\varphi(y)}{\int_0^1 \Pr[\mathbf{x}|y]\varphi(y) dy} = \frac{\Pr[\mathbf{x}|y]\varphi(y)}{\Pr[\mathbf{x}]} \tag{10}$$

Accordingly, the expectation of the trait score associated with a manifest response vector is just

$$E[y|\mathbf{x}] = \int_0^1 h[y|\mathbf{x}]y dy = \frac{\int_0^1 \Pr[\mathbf{x}|y]y\varphi(y) dy}{\Pr[\mathbf{x}]} . \tag{11}$$

We shall regard $E[y|\mathbf{x}]$ as an estimate of the person score corresponding to the response vector \mathbf{x} . The variance of the posterior distribution serves as a measure of the precision of $E[y|\mathbf{x}]$ as a representation of the person score and can be calculated from the relation $V[y|\mathbf{x}] = E[y^2|\mathbf{x}] - [E[y|\mathbf{x}]]^2$. As discussed in Appendix A, there is a form for the score of any \mathbf{x} that is similar to Eqs 8 and 9.

Local Identifiability

Eq 8 is locally identifiable if the Jacobian matrix of the transformation from model parameters to response probabilities has full column rank for a given vector of parameter values [12]. Our numerical explorations suggest that this model, without the imposition of any additional constraints, is locally identifiable for plausible regions of the parameter space. Nevertheless, the imposition of normalizations of the form $k_i = 1$ for some i , providing an item-specific metric of the latent variable, produces more stable full-information maximum likelihood estimates than the unconstrained model. This is the result of the following interesting circumstance: the power of a Beta-distributed random variable is a random variable that is nearly—but not quite—distributed as Beta variable. The choice of i to fix is arbitrary in the following senses: a) the fit of the model to data is virtually independent of this choice and b) ML estimates of item parameters for all normalizations are related in a simple way. This last point deserves a brief elaboration.

Suppose for some set of data the model is normalized with respect to the h^{th} item and that ML estimates are written as k_i^h , and a^h, b^h with $k_i^h = 1$ if $i = h$. Let k_i^h and a^h, b^h be the ML estimates for normalization with respect to item h so that $k_i^h = 1$ if $i = h$. To a close approximation, $k_i^j = k_i^h / k_j^h$. We also note that the special case in which the b parameter is set *a priori* to 1 is underidentified and requires an arbitrary constraint on one item parameter or, equivalently, on the a distribution parameter. With this constraint, there are simple closed form solutions for the ratios of item parameters to the single distributional parameter. For this case, Eq 6 simplifies to

$$\Pr[x_i = 1] = E[y^{k_i}] = \frac{a}{a + k_i} \tag{12}$$

and hence

$$\frac{k_i}{a} = \frac{[1 - \Pr[x_i = 1]]}{\Pr[x_i = 1]} . \tag{13}$$

Thus the ratio of any item hardness to the single distribution parameter can be recreated simply as a ratio of failures to successes at the marginals.

Illustrative Analyses of Dichotomous Responses

Issues of Fit

We go on to apply this model to sets of dichotomous items, making comparisons to results obtained with the Rasch model [13] for the simplest cases. The overall fit of the model can be assessed using the likelihood ratio chi-square, which can also be used to test nested models we

shall introduce below. But for comparison of non-nested models, we use the model selection criterion of Raftery’s BIC [14], which is equal to $L^2 - df \ln(N)$, where L^2 is the likelihood ratio chi-square, df represents the degrees of freedom, and N the number of persons in the sample. (The saturated model has a BIC of 0, any model with a negative BIC is preferred to the saturated model, and the model with the lowest BIC is preferred.) As recently emphasized by Weakliem [15], BIC is not without its drawbacks, and more rigorous implementations of Bayesian logic are now tractable [16]. However, the chief practical drawback of BIC seems to be its overly conservative nature, as it prefers more parsimonious models, and given the temptation to over-fit data sets, we think that BIC serves well as a general criterion for model comparison in which many models are fit to the same data set.

The Single Group Case

We begin by re-analyzing the classic Army data presented by Stouffer [17] and analyzed using a Rasch model by Duncan [4] and Kelderman [18]. The data are to be found in Table 1. We present the model that results when we constrain $k_4 = 1$; the likelihood chi-square is 17.42, with 10 degrees of freedom ($p = .066$). Raftery’s BIC for this model is -51.66 . By contrast, the Rasch model implemented by Duncan had a likelihood ratio chi-square of 10.93 at 8 degrees of freedom, ($p = .206$) leading to a BIC of -44.33 . In sum, the loglinear Rasch model has a closer fit but uses up more degrees of freedom; while deviations from the extended binomial error are marginally significant according classical criteria, the more parsimonious model is preferred according to the Bayesian criterion.

Most importantly, the two models agree closely as to the positions of the different items. We find a correlation $> .99$ between the natural log of the extended binomial error model’s item parameters and those of the Rasch model. (Here we use results from a conditional maximum likelihood fit. Duncan’s item parameters are quite different, which we assume to be a mistake, since the order of parameter hardnesses differs not only from our Rasch analyses, but from the marginal distribution of the items. Kelderman did not report item parameters.) Inspection of the estimated scores (Table 2) from the extended binomial error (Eq 11) model shows that

Table 1. Stouffer’s Army Data.

Item A	Item B	Item C	Item D	Frequency
0	0	0	0	229
0	0	0	1	199
0	0	1	0	52
0	0	1	1	96
0	1	0	0	25
0	1	0	1	60
0	1	1	0	16
0	1	1	1	69
1	0	0	0	16
1	0	0	1	45
1	0	1	0	8
1	0	1	1	55
1	1	0	0	10
1	1	0	1	42
1	1	1	0	3
1	1	1	1	75

doi:10.1371/journal.pone.0141981.t001

Table 2. Results of Fit to Army Data.

Parameter	Value	Standard Error
k_1	4.0879	.3124
k_2	3.4005	.2534
k_3	2.5674	.1875
k_4	1.0000*	-----
a	3.0223	.3350
b	1.7173	.1623

* = fixed; Log-likelihood chi-square = 17.422; df = 10; p = .066.

doi:10.1371/journal.pone.0141981.t002

while there are clear score differences between patterns with the same Rasch raw score (the number of “positive” responses), the ranking of respondents with respect to posterior scores is roughly the same as the ranking with respect to raw scores. The differences between posterior expected value scores over response patterns are small compared to the posterior standard deviations for each pattern; this is to be expected given that the number of items is small.

In this case and most others we have investigated the log linear version of the Rasch and extended binomial error models lead to similar conclusions, though latter tends to be more parsimonious and the former to fit somewhat better, as it has $M-2$ more free parameters. Nevertheless, the two need not agree. Numerical investigations demonstrate the existence of cases in which response distributions generated by the extended binomial error model parameters fit the Rasch model well (as judged by goodness-of-fit measures) but produced estimates of log-linear trait distribution parameters that violate the moment inequalities conditions of Cressie and Holland [19].

In sum, the EBE behaves similarly to the Rasch model; given an implementation involving greater distributional flexibility (the loglinear version) the Rasch model will naturally fit somewhat better at the cost of more parameters. There are, however, data that are fit by one model and not by the other. Most importantly, the probability-like metric of the latent trait allows for extensions that may be of great interest when we wish to examine the mechanisms of response processes (as opposed to scoring long tests). We go on to examine several such extensions, starting with models for independent groups under experimental conditions.

The Multiple Group Case

The multiple group extension of the extended binomial error model permits investigation of issues related to item bias and, under certain conditions, study of group differences in the distribution of the latent trait. Let $g = 1, \dots, G$ index groups which may be formed by partitioning a single random sample, sampling from diverse populations, or by random assignment of different item formats. To allow for differences in item and distribution parameters over these groups, we may rewrite Eqs 4 and 5 as follows:

$$\Pr[x_{ig} = 1 | y, k_{ig}] = y^{k_{ig}} \tag{14}$$

$$\varphi_g(y) = \frac{\Gamma(a_g + b_g)}{\Gamma(a_g)\Gamma(b_g)} y^{(a_g-1)}(1 - y)^{(b_g-1)} \tag{15}$$

To illustrate, we take data from a national telephone survey of Italian adults 18–69 years old that was conducted in April and May, 1994 [20]. The survey was part of a study of regional and ethnic prejudice in Italy and included a series of items dealing with stereotypical beliefs about

Africans and East Europeans living in Italy. Each respondent was assigned at random to a set of questions targeted at one of three immigrant groups: a) North Africans, such as Moroccans, Tunisians, or Algerians (probability = .25); b) Africans from regions of Central Africa, such as Senegal and Somalia (probability = .25); and c) Eastern Europeans, such as Poles, Albanians, or Slavs (probability = .5). For each target group, the respondent was presented with a statement incorporating a positive or negative adjective pertaining to the group, e.g.: “do you agree that most of them are complainers? (they try to make others feel sorry for them)”. The interviewer then asked respondents “Do you agree strongly, agree somewhat, disagree somewhat, or disagree strongly with this description?”

For our example, we selected responses to items incorporating four negative adjectives which roughly translate into “selfish”, “slackers”, “violent”, and “complainers”, combined the North and Central African target groups into a single target group (based on similarity of the marginal distributions), and dichotomized the responses into “Agree” (coded 1) and “Disagree” (coded 0). The response distribution for $N = 2001$ respondents in the 1994 survey is shown in [Table 3](#).

As an exercise, we can use these data to determine whether the responses are consistent with the hypothesis of an underlying trait of prejudice or hostility to minorities, and if so, if Africans and Eastern Europeans are perceived identically. If there are differences in the perception of Africans and Eastern Europeans, we can determine whether there is still a single trait of “prejudice” with perhaps different thresholds for attributing negative characteristics to Africans and Eastern Europeans, or whether there seem to be different latent traits involved when it comes to judging members of the two target groups.

[Table 4](#) presents fit statistics for selected models for these data. The first model corresponds to Eqs [14](#) and [15](#), only adding the constraint $k_{3g} = 1$ for $g = 1, 2$ (the reason the third item is chosen will become clear below). This model fits quite well, generating a likelihood ratio chi-square of 15.93 with 20 degrees of freedom ($p = .721$). Models 2 and 3 impose substantively important constraints. Model 2 sets $k_{ig} = k_i$ for all i ; it is a test of identical item meanings (semantic invariance) that allows the choice of target group to evoke different traits (e.g., degree of hostility to Africans or degree of hostility to Eastern Europeans). This sort of model might be used to examine item bias across two non-experimental groups; if this model failed to fit one could attempt to see if there were particular items that had different hardnesses across groups.

Given that our groups are the results of experimental treatments, however, we may instead begin by assuming identity of the distribution of the latent trait. Model 3 sets $a_g = a$ and $b_g = b$; given the normalization $k_{3g} = 1$ for $g = 1, 2$ this is equivalent to saying that there is only one trait of overall prejudice, but the items (except for the third) can have different hardnesses depending on the target group.

Given model 1, model 2 must clearly be rejected as the difference in chi-square is significant (10.31 at 3 df, $p = .016$). Given model 1, however, loss of fit due to the constraints associated with model 3 is insignificant (chi-square of .62 at 2 df, $p = .733$). (The results here are not independent of the choice of item that is fixed; setting $k_{3g} = 1$ for $g = 1, 2$ resulted in the lowest chi-square for model 3 and was hence also used for models 1 and 2.) Further, model 4 demonstrates that the location of item 2 can also be equated for the two groups (the chi-square difference of 2.01 between models 3 and 4 is insignificant at 1 df with $p = .156$), although models 5 and 6 demonstrate that this is not true of items 1 and 4. Inspection of the item parameters demonstrates that the item locations are more spread out on the underlying continuum when the target group is Africans than when the target is Europeans. Such a result is consistent with the interpretation that there is a single trait of out-group hostility among the respondents, but that Italians have a more differentiated stereotype of Africans than they do of Eastern Europeans.

Table 3. 1994 Italian Survey: Stereotype Data (Sniderman, et al., 1995).

Item 1	Item 2	Item 3	Item 4	Immigrants are African	Immigrants are East European
0	0	0	0	342	361
0	0	0	1	164	129
0	0	1	0	35	32
0	0	1	1	44	47
0	1	0	0	53	41
0	1	0	1	60	51
0	1	1	0	18	20
0	1	1	1	51	38
1	0	0	0	35	45
1	0	0	1	39	47
1	0	1	0	9	20
1	0	1	1	28	27
1	1	0	0	11	10
1	1	0	1	39	36
1	1	1	0	13	17
1	1	1	1	66	73
Totals				1007	994

doi:10.1371/journal.pone.0141981.t003

Other Extensions

Generalization To Ordered Polychotomies

The generalization of the extended binomial error model to ordered polychotomies relies on a standard threshold parameterization for ordered categories that was developed in an IRT context by Edwards and Thurstone [21] and also for regression analysis of ordered categories (see, for example, [22–24]). Let $j = 1, \dots, J$ represent the *a priori* order of the response categories for the i^{th} item where $j = 1$ is “low” and $j = J$ is “high”. We denote the hardness of each of these response categories as $k_{i,j}$, where $k_{i,1} < k_{i,2} < \dots < k_{i,J-1}$. By analogy with Eq 1 for dichotomous responses, we write the probability of that a response fall into category j or higher as

$$\Pr[x_i \geq j|y] = y^{k_{i,j}}. \tag{16}$$

For a fixed y value, the probabilities thus defined diminish as j increases. Given this definition, the probabilities of the intermediate response categories are calculated as differences between the probabilities associated with adjacent dichotomizations. This then implies that the model

Table 4. Results of Fitting Models to Italian Data.

Model	Parameters Constrained To Be Identical Across Groups	Likelihood Ratio Chi-Sq	df	Probability
1	$k3^*$	15.93	20	.721
2	$k1, k2, k3^*, k4$	26.24	23	.290
3	$k3^*, a, b$	16.55	22	.788
4	$k2, k3^*, a, b$	18.56	23	.726
5	$k2, k3^*, k4, a, b$	21.49	24	.610
6	$k1, k2, k3^*, k4, a, b$	27.61	25	.326

* = fixed parameter

doi:10.1371/journal.pone.0141981.t004

for ordered categories can be written as follows:

$$\Pr[x_i = j|y] = y^{k_{ij}} - y^{k_{i,j+1}} \tag{17}$$

where we set the second term to zero for $j = J$.

The model given as Eq 17 has a number of welcome features. First, the probability function for any category j is single-peaked; if we denote its maximum y^*_j , then $(j < k) \Leftrightarrow (y^*_j < y^*_k)$; $y^*_1 = 0$ and $y^*_J = 1$. Finally, because this is a threshold model, it satisfies the two principles Jansen and Roskam [25] called the *joining assumption* (the probability of an aggregated response category is the sum of the probabilities of its constituent, usually adjacent, response categories) and ξ -*equivalence* (the latent traits embedded in aggregated and disaggregated versions of the item response functions are identical or are related by an admissible transformation). In contrast, the most methodologically tractable generalizations to the polychotomous case in the Rasch model (the rating scale and partial credit models and their relatives [26]) do not satisfy these criteria which means that a model that fits the polychotomous data may not fit dichotomous or otherwise collapsed data [27–30]. (The graded response model [31] does satisfy the collapsing conditions but is somewhat more complex.)

An Illustration of the Model for Ordered Polychotomies

To illustrate the application of the polychotomous model to social data, we use a small example drawn from the 2000 U.S. National Alcohol Survey, a cross-sectional national probability household survey on alcohol use and problems [32]. Table 5 shows the response distribution pertaining to a cross-classification of two 4-category items dealing with reasons for abstention from alcohol in a sample of 547 women aged 18–29. They were asked “how important to you are each of the following reasons for abstaining from alcohol beverages or being careful about how much you drink.” The two reasons represented in Table 6 are “drinking is bad for your

Table 5. National Alcohol Survey Data.

Item 1	Item 2	Frequency
1	1	9
1	2	3
1	3	7
1	4	1
2	1	4
2	2	3
2	3	7
2	4	6
3	1	15
3	2	19
3	3	65
3	4	51
4	1	15
4	2	14
4	3	80
4	4	248
Totals		547

Response categories: 1 ‘Not a reason’, 2 ‘Not an Important Reason’, 3 ‘Somewhat Important Reason’, 4 ‘Very Important.’ Item 1: “Drinking is bad for your health”; item 2: “Drinking can get you sick.”

doi:10.1371/journal.pone.0141981.t005

Table 6. Results of Fitting Models to Data in Table 5.

Parameter	Estimate	Standard Error
$k_{1,1}$	1.000*	–
$k_{1,2}$	0.132	0.021
$k_{1,3}$	0.062	0.014
$k_{2,1}$	1.565	0.177
$k_{2,2}$	0.301	0.042
$k_{2,3}$	0.141	0.025
a	1.112	0.194
b	0.598	0.097

*fixed; Log-likelihood ratio chi-squared value = 12.78; df = 8; p>0.173

doi:10.1371/journal.pone.0141981.t006

health” and “drinking can get you sick” and the response categories for both reasons are “not a reason at all,” “not an important reason,” “somewhat important” and “very important.” In this illustration, the category response functions are written so that high values are associated with responses indicating the reasons stated are important for decisions to abstain or to be careful about drinking. Thus $y^{k_{i,1}}$ is the conditional probability, given y , that a respondent will regard the “drinking is bad for your health” as “an important reason” for abstaining or being careful about drinking.

Table 6 shows the parameter estimates and standard errors for fitting the extended binomial error model for ordered polychotomies to these data. As judged by the likelihood ratio chi-squared value (12.78, 8df), the fit of the model to these data is acceptable. The beta distribution generated by $a = 1.112$ and $b = 0.598$ is skewed toward high values of y implying that the majority of respondents consider both reasons for abstinence to be important. A comparison of the category threshold parameters indicates that “bad for your health” is considered a more compelling reason for abstinence than “getting sick” among the young adult women in the national sample.

Single-Peaked Response Functions

This threshold formulation can also be adapted to model response processes in which subjects are more likely to give a positive response to some item if that item’s location is near their own on the latent trait. Following the method of Andrich and Luo [33–36], each item is turned into a trichotomy, with two failure regions (the item is too far above the subject for her to answer in a positive direction and the item is too far below her), with the intermediate leading to a positive response. However, there is an alternative approach that is sufficiently flexible to handle a variety of empirical cases.

Here we approximate the response function in question for some set of such dichotomous items as follows:

$$\Pr[x_i = 1 | y, k_i] = \alpha_i y^{k_i} (1 - y^{k_i}) \tag{18}$$

which is a unimodal function achieving its maximum when $y^{k_i} = .5$, which is equivalent to $\xi = \chi_i$ in an unbounded metric if we again define $\chi_i = \log(k_i)$, and $\xi = -\log[-2\log(y)]$. The leading α_i may be considered akin to a discriminating ability parameter for cumulative models, or it may be considered an overall normalization factor if constrained $\alpha_i = \alpha$ for all i ; it may also be fixed in advance, such that (for example) the predicted probability of a positive response is 1 when ξ

$= \chi_i$. Again, there is a parsimonious representation of the unconditional probability of any vector \mathbf{x} for any set of parameter values $\mathbf{h} = (a, b, \alpha_1, \dots, \alpha_M, k_1, \dots, k_M)$ (see [Appendix A](#)).

Conclusion

In sociology the loglinear Rasch model has become the most widely known and used IRT model in sociology [37–41]. There are two reasons for this popularity. First, Duncan [4] argued that its indifference to the distribution of the latent trait made it a better scientific instrument than the covariance-based methods that dominate social modeling. Second, the Rasch model’s parameters turned out to be estimable with a very simple loglinear approach [18, 42]. This approach treats the distribution of the latent variable as a set of nuisance parameters—the total score preserves all useful information in grading respondents, and the item hardnesses can be estimated without further investigation of the distribution of the latent trait.

But preserving a metric representation of the trait aids the investigation of the response process. While other IRT models, including the Rasch model, can be adapted in this way, the model presented here allows for a wide-range of substantively important extensions to be modeled rather simply, due to the combination of a) a latent trait that can be interpreted as a probability, b) a simple family of response functions that can model dichotomous, ordered polychotomous, and unfolding-type responses, and c) a flexible two-parameter Beta distribution for the latent trait. Such a family of models can be particularly useful in the analysis of experiments that involve changes in wording, response formats, and item order.

Appendix A

A compact representation of the predicted response probabilities for any value of the parameters can be derived from the exclusion-inclusion theorem (see, e.g., [43], 72f). Let the vector $\mathbf{h} = (a, b, k_1, \dots, k_M)$ represent the model parameters and $\Pr[\mathbf{x}|\mathbf{h}]$ the unconditional probability of a response for a given \mathbf{h} . For any \mathbf{x} , let $I(\mathbf{x}) = \{i: x_i = 0\}$ and $C(\mathbf{x}) = \{i: x_i = 1\}$. Let A be any subset of $I(\mathbf{x})$, including the null set \emptyset and let $|A|$ be the number of elements in A . Now since the responses to each item are assumed to be independent conditional on the value of the latent trait, we can write $\Pr[\mathbf{x}|\mathbf{h}]$ in product form, where the expectation is taken with respect to the distribution of the latent trait:

$$\begin{aligned} \Pr[\mathbf{x}|\mathbf{h}] &= \Pr[x_1, x_2, \dots, x_N] = E[\Pr[x_1, x_2, \dots, x_N|y]] = E\left[\prod_{i=1}^N \Pr[x_i|y]\right] \\ &= E\left[\prod_{i \in C(\mathbf{x})} \Pr[x_i = 1|y] \prod_{i \in I(\mathbf{x})} (1 - \Pr[x_i = 1|y])\right] \\ &= E\left[\prod_{i \in C(\mathbf{x})} \Pr[x_i = 1|y] \sum_{A \subset I(\mathbf{x})} (-1)^{|A|} \prod_{i \in A} \Pr[x_i = 1|y]\right] \\ &= E\left[\sum_{A \subset I(\mathbf{x})} (-1)^{|A|} \prod_{i \in A \cup C(\mathbf{x})} \Pr[x_i = 1|y]\right] = \sum_{A \subset I(\mathbf{x})} (-1)^{|A|} E\left[\prod_{i \in A \cup C(\mathbf{x})} \Pr[x_i = 1|y]\right] \end{aligned} \tag{A.1}$$

This representation is valid for all latent trait models for dichotomous responses that assume conditional independence. For the extended binomial error model, we make the following

substitution:

$$E \left[\prod_{i \in AUC(\mathbf{x})} \Pr[x_i = 1|y] \right] = \frac{\Gamma(a + b)\Gamma(a + \sum_{i \in AUC(\mathbf{x})} k_i)}{\Gamma(a)\Gamma(a + b + \sum_{i \in AUC(\mathbf{x})} k_i)} \tag{A.2}$$

The response pattern score and score variance can be expressed similarly. Written in terms of the model parameters, the score $E[y|\mathbf{x}]$ is given as:

$$E[y|\mathbf{x}] = \frac{\sum_{A \subset I(\mathbf{x})} (-1)^{|A|} \frac{\Gamma(a + b)\Gamma(a + 1 + \sum_{i \in AUC(\mathbf{x})} k_i)}{\Gamma(a)\Gamma(a + b + 1 + \sum_{i \in AUC(\mathbf{x})} k_i)}}{\sum_{A \subset I(\mathbf{x})} (-1)^{|A|} \frac{\Gamma(a + b)\Gamma(a + \sum_{i \in AUC(\mathbf{x})} k_i)}{\Gamma(a)\Gamma(a + b + \sum_{i \in AUC(\mathbf{x})} k_i)}} \tag{A.3}$$

Finally, there is a parsimonious representation of the unconditional probability of any vector \mathbf{x} for any set of parameter values in the unfolding-type model of Eq 19. With $I(\mathbf{x})$ and $C(\mathbf{x})$ defined as above, and A any subset of $I(\mathbf{x})$, let B be any subset of the union of this A with $C(\mathbf{x})$, and let $|B|$ be the number of elements in B . Then it can be shown that

$$P[\mathbf{x}|\mathbf{h}] = \sum_{A \subset I(\mathbf{x})} (-1)^{|A|} \left\{ \left(\prod_{i \in AUC(\mathbf{x})} \alpha_i \right) \sum_{B \subset \{AUC(\mathbf{x})\}} (-1)^{|B|} E \left[y^{\left(\sum_{i \in AUC(\mathbf{x})} k_i + \sum_{i \in B} k_i \right)} \right] \right\} \tag{A.4}$$

For the expectation involving y we make the following substitution

$$E \left[y^{\left(\sum_{i \in AUC(\mathbf{x})} k_i + \sum_{i \in B} k_i \right)} \right] = \frac{\Gamma(a + b)\Gamma(a + \sum_{i \in AUC(\mathbf{x})} k_i + \sum_{i \in B} k_i)}{\Gamma(a)\Gamma(a + b + \sum_{i \in AUC(\mathbf{x})} k_i + \sum_{i \in B} k_i)} \tag{A.5}$$

and algorithms related to those discussed above can be used to obtain simultaneous estimates of score and distribution parameters.

Appendix B

The cross-classification of responses pertaining to any two items i and j generates a 2 by 2 table of population proportions. Consider three summary measures that are sufficient to reconstruct the marginal and interior probabilities of this table: $\Pr[x_i = 1, x_j = 1]$, $\Pr[x_i = 1]$, and $\Pr[x_j = 1]$. (Note that here, $\Pr[\]$ indicates *observed* probabilities.) Under the model represented by Eqs 1

and 5, these have following structure:

$$\Pr[x_i = 1, x_j = 1] = E[y^{k_i+k_j}] = \frac{\Gamma(a+b)\Gamma(a+k_i+k_j)}{\Gamma(a)\Gamma(a+b+k_i+k_j)} \tag{B.1}$$

$$\Pr[x_i = 1] = E[y^{k_i}] = \frac{\Gamma(a+b)\Gamma(a+k_i)}{\Gamma(a)\Gamma(a+b+k_i)} \tag{B.2}$$

$$\Pr[x_j = 1] = E[y^{k_j}] = \frac{\Gamma(a+b)\Gamma(a+k_j)}{\Gamma(a)\Gamma(a+b+k_j)} \tag{B.3}$$

These equations cannot be solved uniquely to obtain a, b, k_i , and k_j . However, if we set $k_i = 1$ (in effect, normalizing the model by setting $y = \Pr[x_i = 1|y]$) and use $\Gamma(s+1) = s\Gamma(s)$, we get

$$\Pr[x_i = 1, x_j = 1] = \frac{\Gamma(a+b)(a+k_j)\Gamma(a+k_j)}{\Gamma(a)(a+b+k_j)\Gamma(a+b+k_j)} \tag{B.4}$$

$$\Pr[x_i = 1] = \frac{\Gamma(a+b)a\Gamma(a)}{\Gamma(a)(a+b)\Gamma(a+b)} = \frac{a}{(a+b)} \tag{B.5}$$

It follows that

$$\frac{\Pr[x_i = 1, x_j = 1]}{\Pr[x_j = 1]} = \frac{a+k_j}{a+b+k_j} \tag{B.6}$$

and with manipulation we get

$$\frac{k_j}{a+b} = S_j \tag{B.7}$$

where

$$S_j = \frac{\Pr[x_i = 1, x_j = 1] - \Pr[x_i = 1]\Pr[x_j = 1]}{\Pr[x_j = 1] - \Pr[x_i = 1, x_j = 1]} \tag{B.8}$$

for all i and j . Thus, for each item other than the i^{th} , its cross-classification with x_i gives us its ratio to $a + b$.

To construct an initial estimate of $a + b$, which we will denote Θ for brevity, we take advantage of the fact that from Eq B.3,

$$\Pr[x_j = 1] = \frac{\Gamma(\Theta)\Gamma(\Theta(\Pr[x_i = 1] + S_j))}{\Gamma(\Theta(\Pr[x_i = 1]))\Gamma(\Theta(S_j + 1))} \tag{B.9}$$

and hence

$$\Pr[x_j = 1] - \frac{\Gamma(\Theta)\Gamma(\Theta(\Pr[x_i = 1] + S_j))}{\Gamma(\Theta(\Pr[x_i = 1]))\Gamma(\Theta(S_j + 1))} = 0 \tag{B.10}$$

For admissible values of $\Pr[x_i = 1]$, $\Pr[x_j = 1]$, and S_j , Eq B.10 has a single positive root Θ .

When the model describes the data perfectly, the $M-1$ equations corresponding to each j ($j \neq i$) should yield the same root. For actual data, we average our derived values of Θ . We then substitute this value in Eq B.7 to get initial values of k_j , $j \neq i$. Our starting estimate of the parameter a can be recovered from the product of $(a + b)$ and $\Pr[x_i = 1]$ (given that, by construction, $k_i = 1$), and that of b from $b = \Theta - a$.

Author Contributions

Analyzed the data: JAW JLM JB. Contributed reagents/materials/analysis tools: JAW. Wrote the paper: JAW JLM. Mathematical solutions: SJH. Identifiability checks: SJH. Algorithm: SJH JLM JAW. Programming: JLM JB.

References

1. Keats JA, Lord FM. A theoretical distribution for mental test scores. *Psychometrika* 1962; 27:59–62.
2. Lazarsfeld PF. Latent structure analysis and test theory. In: Lazarsfeld PF Henry NW, editors. *Readings in mathematical and social science*. Cambridge, Mass.: MIT Press; 1966. p. 78–88.
3. Coleman JS. *Introduction to mathematical sociology*. Glencoe, Ill.: Free Press; 1964.
4. Duncan OD. Rasch measurement in survey research: Further examples and discussion. In: Turner CF and Martin E, editors. *Surveying subjective phenomena, Vol. II*. New York: Russell Sage Foundation; 1984. p. 367–404.
5. Keats JA. Some generalizations of a theoretical distribution of mental test scores. *Psychometrika* 1964; 29:215–231.
6. Lord RM, Novick MR. *Statistical theories of mental test scores*. Reading, Mass.: Addison-Wesley; 1968.
7. Huynh H. Error rates in competency testing when test retaking is permitted. *J of Ed Stats* 1990; 15: 39–52.
8. Wilcox RR. A review of the beta-binomial model and its extensions. *J of Ed Stats* 1981; 6:3–32.
9. Polak E, Ribière G. Note sur la convergence de directions conjuguée. *Rev. Francaise Informat Recherche Operationelle* 1969; 16: 35–43.
10. Fletcher R, Reeves CM. Function minimization by conjugate gradients. *Comput J.*, 1964; 7: 149–154.
11. Nelder JA, Mead R. A simplex method for function minimization. *Comput J* 1965; 7: 308–313.
12. Wald A. Note on the identification of economic relations. In: Koopmans TC, editor. *Statistical inference in dynamic economic models*. New York: Wiley; 1950. p. 305–310.
13. Rasch G. *Probabilistic models for some intelligence and attainment tests*. Copenhagen: The Danish institute of educational research; 1960.
14. Raftery A. A note on Bayes factors for log-linear contingency table models with vague prior information. *JRSS Ser. B* 1985; 48: 249–250.
15. Weakliem DL. A critique of the Bayesian information criterion for model selection. *Sociological Methods and Research* 1999; 27:359–397.
16. Raftery A. Bayes factors and BIC. *Sociological Methods and Research* 1999; 27:411–427.
17. Stouffer SA, Guttman L, Suchman EA, Lazarsfeld PF, Star SA, Clausen JA. *Measurement and prediction*. Princeton: Princeton University Press; 1950.
18. Kelderman H. Loglinear Rasch model tests. *Psychometrika* 1984; 49: 223–245.
19. Cressie N, Holland PW. Characterizing the manifest probabilities of latent trait models.” *Psychometrika* 1983; 48:129–141.
20. Sniderman P, Piazza T, Peri P, Schizzerotto A. *Codebook for the 1994 survey on regional and ethnic prejudice in Italy*. Berkeley: Survey Research Center, University of California; 1994.
21. Edwards AL, Thurstone LL. An internal consistency check for scale values determined by the method of successive integers.” *Psychometrika* 1952; 17:169–180.
22. Bock RD. *Multivariate statistical methods in behavioral research*. New York: McGraw-Hill; 1975.
23. Cox DR *Analysis of binary data*. London: Chapman and Hall Ltd; 1970.
24. Maddala GS. *Limited dependent and qualitative variables in econometrics*. Econometric Society Monograph 3. Cambridge: Cambridge University Press; 1983.

25. Jansen PGW, Roskam EE. Latent trait models and dichotomization of graded responses. *Psychometrika* 1986; 51:69–91.
26. Thissen D, Steinberg L. A taxonomy of item response models. *Psychometrika* 1986; 51:567–577.
27. Rasch G. An individualistic approach to item analysis. In: Lazarsfeld PF Henry NW, editors. *Readings in mathematical and social science*. Cambridge, Mass.: MIT Press; 1966. p. 89–107.
28. Andrich D. Models for measurement, precision, and the nondichotomization of graded responses. *Psychometrika* 1995; 60:7–26.
29. Andrich D. Further remarks on nondichotomization of graded responses. *Psychometrika* 1995; 60: 37–46.
30. Andrich D. Distinctive and incompatible properties of two common classes of IRT models for graded responses. *Applied Psychological Measurement* 1995; 19: 101–119.
31. Roskam EE. Graded responses and joining categories: A rejoinder to Andrich's 'Models for measurement, precision, and the nondichotomization of graded responses.' *Psychometrika* 1995; 60, 27–35.
32. Greenfield TK, Nayak MB, Bond J, Ye Y, Midanik LT. Maximum quantity consumed and alcohol-related problems: assessing the most alcohol drunk with two measures. *Alcoholism: clinical and experimental research* 2006; 30, 1576–1582.
33. Andrich D, Luo G. A hyperbolic cosine latent trait model for unfolding dichotomous single-stimulus responses. *Appl Psych Meas* 1993; 17: 253–176.
34. Andrich D. A hyperbolic cosine latent trait model for unfolding polytomous responses: Reconciling Thurstone and Likert methodologies. *Brit J of Math and Stat Psych* 1996; 49: 347–365.
35. Luo G. A general formulation for unidimensional unfolding and pairwise preference models. *J Math Psych* 1998; 42:400–417.
36. Luo G. A class of probabilistic unfolding models for polytomous responses. *J Math Psych* 2001; 45:224–248.
37. MacIntosh R. Global attitude measurement: An assessment of the world values survey postmaterialism scale. *Am Soc Rev* 1998; 63: 452–464.
38. Hagan J, Foster H. Youth violence and the end of adolescence. *Am Soc Rev* 2001; 66: 874–899.
39. Hout M, Greeley AM. The center doesn't hold: Church attendance in the United States, 1940–1984. *Am Soc Rev* 1987; 52: 325–345.
40. Browning CR, Leventhal T, Brooks-Gunn J. Sexual initiation in early adolescence: The nexus of parental and community control. *Am Soc Rev* 2005; 70: 758–778.
41. Oegema D, Klandermans B. Why social movement sympathizers don't participate: Erosion and non-conversion of support. *Am Soc Rev* 1994; 59: 703–722.
42. Kelderman H. Item bias detection using loglinear IRT. *Psychometrika* 1989; 54:681–697.
43. Brualdi RA. *Introductory Combinatorics*. New York: North Holland; 1977.