JOURNAL OF APPLIED MEASUREMENT, *2*(4), 389–423
Copyright© 2001

The Rasch Model, Additive Conjoint Measurement, and New Models of Probabilistic Measurement Theory

George Karabatsos
LSU Health Sciences Center

This research describes some of the similarities and differences between additive conjoint measurement (a type of fundamental measurement) and the Rasch model. It seems that there are many similarities between the two frameworks, however, their differences are nontrivial. For instance, while conjoint measurement specifies measurement scales using a data-free, non-numerical axiomatic frame of reference, the Rasch model specifies measurement scales using a numerical frame of reference that is, by definition, data dependent. In order to circumvent difficulties that can be realistically imposed by this data dependence, this research formalizes new non-parametric item response models. These models are probabilistic measurement theory models in the sense that they explicitly integrate the axiomatic ideas of measurement theory with the statistical ideas of order-restricted inference and Markov Chain Monte Carlo. The specifications of these models are rather flexible, as they can represent any one of several models used in psychometrics, such as Mokken's (1971) monotone homogeneity model, Scheiblechner's (1995) isotonic ordinal probabilistic model, or the Rasch (1960) model. The proposed non-parametric item response models are applied to analyze both real and simulated data sets.

Requests for reprints should be sent to George Karabatsos, Assistant Professor of Biostatistics, LSU Health Sciences Center, 1600 Canal Street, Suite 1120, New Orleans, LA 70112-1393, e-mail: gkarab@lsuhsc.edu.

Introduction

The Rasch model and additive conjoint measurement

A vast number of social science variables are based on important psychological attributes, such as ability, intelligence, attitudes, preferences, and perception. However, establishing measurement scales for these variables is not a straightforward task, since such variables are not directly observable. Luce and Tukey's (1964) conjoint measurement theory is an important achievement, as it enables the social sciences to verify the construction of fundamental measurement. The basis of such verification is provided by a set of axioms that define the data structure necessary for the existence of measurement scales.

Consider the case where N examinees, $1 \le n \le N$, respond to M test items, $1 \le g \le M$, where each response is scored as either "correct" $X_{ng} = 1$ or "incorrect" $X_{ng} = 0$. The Rasch model (Rasch, 1960) is defined by the logistic function:

$$P_{ng} = P\left[X_{ng} = 1 \mid \beta_n, \delta_g\right] = \left[1 + \exp\left(\delta_g - \beta_n\right)\right]^{-1}, \tag{1}$$

where $0 < P_{ng} < 1$ is the probability that a respondent n with ability β_n obtains the correct response on an item g with difficulty δ_g, $-\infty < \beta_n, \delta_g < +\infty$. It is well known that the total respondent test score X_{n+} and the total item score X_{+g} are sufficient statistics for the model's logit parameters β_n and δ_g, respectively, $0 \le X_{n+} \le M$, $0 \le X_{+g} \le N$. Also, the model assumes that each respondent's item responses are locally independent, conditional on ability:

$$P\left[X_{n1} = x_{n1}, X_{n2} = x_{n2}, \ldots, X_{nM} = x_{nM} \mid \beta_n\right] = \prod_{g=1}^{L}\left[P_{ng} \mid \beta_n\right]^{X_{ng}}\left(1 - \left[P_{ng} \mid \beta_n\right]\right)^{1-X_{ng}}. \tag{2}$$

Researchers have observed connections between the logistic Rasch model and additive conjoint measurement (Keats, 1967; Fischer, 1968; Brogden, 1976). These connections can be shown in a two-way table. Let any set of finite ability values $\beta = (\beta_1,\ldots\beta_n,\ldots,\beta_N)$ be ordered on the rows of the table, and any set of finite values of item easiness $-\delta = (-\delta_1,\ldots,-\delta_g,\ldots,-\delta_M)$ be ordered on the columns. Also, let the cells of the table contain the response expectations $\mathbf{P} = (P_{ng})_{N \times M}$, after plugging in the corresponding values of β_n and the δ_g into

the Rasch model given in (1). A row containing M values of P_{ng} represents a person response function (PRF), and a column vector containing N values of P_{ng} represents an item response function (IRF).

From such a table, it is easy to show that that for any single person n, the Rasch PRF strictly increases over item easiness (decreasing $-\delta$), and for any single item g, the Rasch IRF strictly increases over ability (increasing β_n). Furthermore, any set of N Rasch PRFs do not intersect, as they conform to the order restrictions characterized by conjoint measurement's *row independence* axioms, and also, any set of Rasch IRFs do not intersect, as they conform to the order restrictions characterized by conjoint measurement's *column independence* axioms. The Rasch model also specifies the strictly increasing IRFs to be parallel, which render parallel PRFs, and hence the model conforms to the order restrictions of additive conjoint measurement, characterized together by the row independence, column independence, and a set of cancellation axioms.

The strictly increasing parallel IRFs characterize an invariant item difficulty scale for the entire range of β, where the item scale is on an interval metric. The strictly increasing parallel PRFs characterize an invariant ability scale for the entire range of $-\delta_g$, where the ability scale is also on an interval metric. Non-decreasing parallel IRFs provide conditions for interval scales of person ability and item difficulty, however IRFs need not be parallel for invariant ability and difficulty scales. Invariant ability and dificulty scales are also represented by non-decreasing and non-intersecting IRFs.

The well-known two-parameter logistic model (e.g., Lord and Novick, 1968) is obtained by replacing the term $(\delta_g - \beta_n)$ in (1) with $\alpha_g(\delta_g - \beta_n)$, where $\alpha_g > 0$ is the slope of the IRF. Clearly, allowing the slope to vary across the items renders the possibility for intersecting IRFs, which contradicts the column independence axiom. Such IRFs leads to an item difficulty scale that is not invariant over the range of β.

Differences between the Rasch model and additive conjoint measurement

Indeed, there are strong connections between additive conjoint measurement and the Rasch model. Both the Rasch model and the conjoint measurement axioms specify parallel IRFs, however, each uses a different approach. Whereas the Rasch model specifies parallel IRFs using a numerical function (equation 1) to restrict P_{ng}, conjoint measurement theory defines the shape of the parallel IRFs with non-numerical order restrictions on P_{ng}.

Perline, Wright, and Wainer (1979) correspond Rasch model goodness-of-fit analysis results with the results of certain conjoint measurement axiom tests, and since the authors observed some level of correspondence, they assert that the Rasch model is a "practical realization" (p. 237) of additive conjoint measurement. There is some room for this assertion. The conjoint additivity axioms, although they represent the structural ideals of interval measurement, are deterministic, algebraic formulations. They appear ill-equipped to handle the random sources of variation that characterize fallible data (Cliff, 1973). On the other hand, the Rasch model's numerical formulation of additive conjoint measurement lends naturally to the available tools of standard statistics.

Normal practice of Rasch model analysis employs global, person, or item fit analysis based on the estimated Rasch model residuals $\hat{\mathbf{y}} = (\hat{\mathbf{y}}_{ng})_{N \times M}$, based on the finite Rasch model parameter estimates $\beta = (\beta_1,...\beta_n,...,\beta_N)$ and $-\delta = (-\delta_1,...,-\delta_g,...,-\delta_M)$ obtained from an observed data set. Given that Rasch model IRFs are parallel, and that parallel IRFs conform to the conjoint additivity axioms, the residual, $\hat{y}_{ng} = X_{ng} - \hat{P}_{ng}$, seems to provide a useful index that measures the "distance" between the observed data point and conjoint additivity (including specific objectivity). Examples of fit methods based on y include the family of Outfit mean square and Infit mean square statistics that assess item and respondent fit to the Rasch model (e.g., Wright and Masters, 1982). For instance, the item Outfit mean square, $OMS_g > 0$, simply averages over the response residuals within item g,

$$OMS_g = \tfrac{1}{N} \sum_{n=1}^{N} \left(\hat{y}_{ng}^2 \,/ \hat{W}_{ng} \right),$$

and the item Infit mean square, $IMS_g > 0$, is the variance-weighted average

$$IMS_g = \sum_{n=1}^{N} \hat{y}_{ng}^2 \Big/ \sum_{n=1}^{N} \hat{W}_{ng},$$

the variance given by $\hat{W}_{ng} = \hat{P}_{ng}(1-\hat{P}_{ng})$. Both OMS and IMS have expected values of 1. According to Linacre and Wright (1994) and Smith (1996), a value significantly less than 1 indicates an item that fits the Rasch model better than expected, and a value significantly greater than 1 indicates an item that misfits the model. The measure of significance is given by Outfit ZSTD or Infit ZSTD, where ZSTD standardizes a mean-square statistic (either OMS or IMS) to a unit normal distribution scale (e.g., see Smith, Schumacker, and Bush, 1997). Therefore, given a Type I error rate of .05,

(Outfit or Infit) ZSTD > 1.96 identifies an item that misfits the Rasch model, and (Outfit or Infit) ZSTD < –1.96 identifies an item that fits the Rasch model better than expected. It is generally recommended that the "optimal" item set should include items that meet the criterion ZSTD < 1.96 for both Outfit and Infit (Linacre and Wright, 1994). Items that fit the Rasch model "better than expected" (ZSTD < –1.96) are not seen as inconsistent with the model. As Linacre and Wright (1994, p. 350) state, "Low fit values do not disturb the meaning of a measure" (Linacre and Wright, 1994, p. 350). This is a reasonable interpretation, because a lower fit mean square value for an item g (IMS, OMS, or ZSTD) implies a higher agreement between the set of X_{ng} and $\hat{P}_{ng}$ over $1 \leq n \leq N$.

Hence, within the context of the Rasch model, it appears simple and convenient to check for the consistency between data and measurement axioms, using standard fit statistics. However, this simplicity and convenience comes at a price. It is important to make the basic consideration that the Rasch model parameters, used to define the parallel IRFs, are estimated directly from the data. Of course, data may contain any level of random or systematic noise. Therefore, the degree to which the data contain noise is the degree to which the Rasch model's frame of reference, the *estimated* parallel IRFs, contains noise. Unfortunately, the Rasch model's specification of additive conjoint measurement is data dependent. Such problems of data dependence have been noticed in previous research. Nickerson and McClelland (1984) empirically show that it is possible for a numerical conjoint measurement model, such as the Rasch model, to conclude excellent or perfect data fit, even for data sets containing serious violations of the conjoint measurement axioms. This is because when the parameters of the numerical conjoint measurement model are estimated, they tend to "absorb" data containing measurement disturbances.

Table 1 shows a Rasch model item analysis of a data set containing 459 respondents responding to 10 dichotomous-scored items, computed by the Rasch analysis software WINSTEPS (Linacre and Wright, 2001). According to the Outfit ZSTD and Infit ZSTD results of Table 1, items 4, 8, and 10 fit the Rasch model better than expected, items 1, 2, 3, 5 and 6 also fit the model, and items 7 and 9 misfit. On average, the 10 items fit the model (average item Infit ZSTD= –.2, average item Outfit ZSTD= .0). In addition, the set of respondents fit the Rasch model quite well. The mean Infit mean square over all respondents is .99 (s.d. = .18, min = .74, max = 1.3), and the corresponding mean Outfit is 1.00 (s.d. = .31, min = .55, max = 1.51). The Rasch model fit analysis, in general, support that the data

stochastically conform to strictly increasing parallel IRFs (conjoint additivity). From these results, it is tempting to conclude that respondents and items are measured on a common, unidimensional interval scale, where the IRFs define an invariant difficulty scale over the entire range of β.

Table 1.

An output table from WINSTEPS, displaying the analysis of data consisting of 459 individuals responding to a 10-item test, dichotomous response format.

ITEM STATISTICS: ENTRY ORDER

ITEM	RAW SCORE	COUNT	MEASURE	ERROR	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
1	121	459	1.52	.12	1.09	1.5	1.29	1.9
2	350	459	-1.48	.13	1.01	.2	1.03	.2
3	366	459	-1.75	.13	.89	-1.6	.79	-1.3
4	123	459	1.49	.12	.74	-4.8	.55	-3.9
5	270	459	-.37	.11	1.08	1.6	1.13	1.4
6	290	459	-.63	.11	.91	-1.8	.85	-1.6
7	212	459	.33	.11	1.30	5.7	1.43	4.6
8	224	459	.19	.11	.83	-3.8	.74	-3.6
9	267	459	-.34	.11	1.25	4.7	1.51	5.1
10	156	459	1.03	.11	.79	-4.1	.68	-3.3
MEAN	238	459	.00	.12	.99	-.2	1.00	.0
S.D.	82	0	1.08	.01	.18	3.5	.31	3.1

Notes:

RAW SCORE refers to the total item score X_{+i}.

COUNT the number of "non-extreme" respondents scoring $0 < X_{n+} < 10$.

MEASURE refers to the item difficulty estimate $\hat{\delta}_g$.

ERROR refers to the standard error of the estimate.

INFIT MNSQ Infit Mean square fit statistic of the item (expected value = 1).

INFIT ZSTD Unit-normal transformation of the Infit mean square statistic (expected value = 0).

OUTFIT MNSQ Outfit Mean square fit statistic of the item (expected value = 1).

OUTFIT ZSTD Unit-normal transformation of the Outfit mean square statistic (expected value = 0).

However, the results of the Rasch analysis obscure the fact that the analyzed data set was generated by the two-parameter logistic model, from the respondent ability uniform distribution range [–3, 3], and item discriminations a = (0.64, 0.72, 1.06, 1.66, 0.56, 0.85, 0.27,1.06, 0.36, 1.29), for items 1 through 10, respectively. Clearly, the generated data define

crossing IRFs, inconsistent with an invariant item scale over the range of b, and therefore contradict a joint interval scale representation of the persons and items. The Rasch analysis was not able to detect many of the items that explicitly violate parallel IRFs (conjoint additivity).

These findings relate to the fact that residual fit statistics suffer from the "masking" effect (Johnson and Albert, 1999, pp. 98-99), also noticed by Smith (1988) in the context of Rasch model fit analysis (for more criticisms of Rasch residual fit analysis, see Karabatsos, 2000). The response residual statistic $y_{ng} = X_{ng} - P_{ng}$, from which the mean square statistics are based on (and their standardized values ZSTD), suffers from masking because of the dependence between X_{ng} and P_{ng}. While the maximum-likelihood estimates β and δ are those that minimize the response residuals y, the same residuals are also used as a basis for which to measure model fit to the observed data $\mathbf{X} = (X_{ng})_{N \times M}$. Therefore, in the case where the data set contains noise, the estimated residuals $\hat{\mathbf{y}}$ underestimate the "true" residuals y. Although the Rasch model estimated matrix $\mathbf{P}$ always satisfies the order restrictions specified by the conjoint additivity axioms, the observed data X used as input to obtain the estimates $\mathbf{P}$ do not necessarily satisfy these axioms. The Rasch model estimates $\mathbf{P}$ give the illusion that the model can automatically construct additive conjoint measurement from any data set, no matter how noisy the data are. However, there is absolutely no basis to assume that such an automatic construction is possible.

It may be tempting to conclude from that perhaps other fit statistics, not based on residuals, should be employed to test data accordance with the Rasch model. However, non-residual fit statistics can suffer from masking as well. Any fit statistic based on the *estimated* parameters β and δ assumes that they are true parameter values, unspoiled by potentially noisy data.

To circumvent such difficulties, "parameter-free" model fit statistics may serve as useful alternatives to test data accordance to the Rasch model, as they focus on the degree to which observed data accord to an "ideal" data structure, instead of the degree to which observed data conform to data-dependent estimated parameters.

The author recently performed an analysis that compared 36 different person fit statistics in their ability to detect aberrant respondents. The study involved 60 simulated data sets of a fully-crossed 5 x 4 x 3 design, where Rasch model parameters were estimated from each data set. The design consisted of 5 types of aberrant respondents (cheaters, lucky guessers, careless respondents, creative respondents, random respondents), 4

groups each characterized by a certain proportion of aberrant respondents in the data set (5%, 10%, 25%, and 50%), and 3 test length groups (17 items, 33 items, 65 items). The results showed that the top four performing person fit statistics were *non-parametric*, having higher detection power than well-known parametric fit statistics (which were calculated using the estimated Rasch model parameters, not the data-generating parameters). The set of parametric fit statistics includes all the total mean square (e.g., Wright and Masters, 1982) and between-item-group mean square fit statistics (Smith, 1986), all likelihood indices (e.g., Drasgow, Levine, and Williams, 1985), all extended caution indices (Tatsuoka, 1984), and the M statistic (Molenaar and Hoijtink's, 1990; Bedrick, 1997). The non-parametric index H^T (Sijitsma and Meijer, 1992) consistently had the best detection rates over the 60 data sets. Coincidentally, this index specifically relates to conjoint measurement, namely the row and column independence axioms, as the value $0 \leq H^T \leq 1$ confirms non-intersecting IRFs (Sijitsma and Meijer, 1992). However, despite the connections of H^T with conjoint measurement theory, this index may be sample dependent to a certain degree (see Roskam and van den Wollenberg, 1986).

Measurement axioms as a data-free frame of reference

Indeed, within a data-dependent frame of reference, it is difficult to detect measurement inconsistencies from data. It therefore seems desirable to incorporate the conjoint measurement axioms in the practice of Rasch model analysis, since they provide a *data-free* frame of reference. Given the deterministic language of the axioms, and the fact that any data set contains some degree of random or systematic noise, it seems necessary to develop a statistical framework for testing data accordance to the measurement axioms.

Previous research has implemented several statistical tests to perform axiom testing, usually based on some function of the number of measurement axiom violations (e.g., Perline, Wright, and Wainer, 1979; Michell, 1990). However the results of such statistical tests are too simplistic, and difficult to interpret. For instance, they do not capture the degree of each axiom violation, and some do not even consider sample size. It is quite difficult to develop any distribution theory based on such counts (Iverson and Falmagne, 1985).

Iverson and Falmagne (1985) considered a more productive approach, as they formulated order-restricted statistical inference methods for test-

ing measurement axioms, such as weak stochastic transitivity and the quadruple condition. Their work represents the first serious effort to integrate statistical inference with axiomatic measurement theory. They also show the statistical and computational complexities can arise when interfacing statistics and measurement theory. For example, their analysis can get quite complex when testing a measurement axiom that implies a large number of order restrictions, or when testing several axioms simultaneously.

This research introduces non-parametric item response models, based on ideas of order restricted statistical inference (Robertson, Wright, and Dykstra, 1988), that estimate IRFs and PRFs under the null hypothesis that the data accord with a given set of measurement axioms. The axioms specify the form of the IRFs and PRFs with prior order restrictions on the data structure, and the estimated IRFs and PRFs serve as the "expected values" for which to assess data fit to the measurement axioms. The models provide a statistical inference framework for many well-known psychometric models, and can be applied in a straightforward fashion using Markov Chain Monte Carlo methods.

The IRFs and PRFs that define the frame of reference for the non-parametric item response models are *not constrained* by data-dependent numerical estimates such as $\beta = (\beta_1,...,\beta_n,...,\beta_N)$ and $-\delta = (-\delta_1,...,-\delta_g,...,-\delta_M)$, which are functions of marginal respondent and item scores, respectively. Instead, the IRFs and PRFs are *only constrained* by non-numerical order restrictions characterized by data-free axioms that are based on measurement theory. Furthermore, the item response model does not assume that the IRFs and PRFs take any specific, arbitrary functional form, unlike many well-known parametric item response models that arbitrarily assume either logistic or normal-ogive functions. There is little justification that either the logistic or normal-ogive functions can represent latent psychological attributes such as ability and attitudes (Scheiblechner, 1999).

Karabatsos (2001) and Karabatsos and Shev (2001) demonstrates that the non-parametric item response models can be used to test several types of measurement axioms, such as weak stochastic transitivity, the quadruple condition, and the axioms of conjoint measurement theory. In the present context, given the specific focus involving the link between additive conjoint measurement and the Rasch model, this research will formulate item response models and methods of statistical inference with the conjoint measurement axioms. Scheiblechner (1995; 1999) develops an interesting axiomatic framework amenable for the practice of probabilistic conjoint

measurement theory in psychometrics. It therefore seems productive to base the development of the non-parametric item response models on the Scheiblechner formulations.

Probabilistic Axiomatic Conjoint Measurement

Basic framework

Scheiblechner's (1995) basic model framework refers to the probabilistic conjoint system $\langle A \times Q, P \rangle$, detailed by the following:

1. The dimension $A = \{a, b, c, \ldots\}$ contains a finite set of N respondents, and $Q = \{x, y, z, \ldots\}$ the finite set of M items.

2. Denote X_{ax} as the response of some respondent a on some item x, where the response can be any (at least) ordinal value on the real number line, $X_{ax} \in Re$. Hence, a variety of test response formats are accommodated (e.g., dichotomous scored, Likert scale, continuous scale).

3. The set $T = \{t, t', t'', t''', \ldots\}$ contains a finite number $K \geq 2$ of ordinal response categories, where $t < t' < t'' < t''' < \ldots$, and any $t \in Re$.

4. A cumulative IRF corresponds to each respondent-item combination, where $P[X_{ab} \geq t \mid ax] = P(t ; ax)$ refers to the probability that respondent a, on some item x, will give a response that is at least the value of t. Each of the NM respondent-item combinations (ax, ay, bx, …) is represented by K of these cumulative response probabilities.

Axiom framework

Scheiblechner (1995; 1999) formalized a set of probabilistic conjoint measurement theory axioms that characterize several psychometric models. The first two axioms are the following:

Axiom W1: Weak respondent (row) independence.

For some item x, and some ordinal response category t,

if $P(t ; ax) < P(t ; bx)$,

then

$P(t' ; ay) \leq P(t' ; by)$, must also be true

for all items $y \in Q$, and all response categories $t' \in T$.

Local independence (LI): The item response vector $\mathbf{t}_{(M)}$ of length M, pertaining to the vector set of items $\mathbf{x}_{(M)}$, satisfies conditional independence, given by:

$$P\left(\mathbf{t}_{(M)}; a\mathbf{x}_{(M)}\right) = \prod_{x \in \mathbf{x}_{(M)}} P\left(t; ax\right) \cdot \tag{3}$$

A unidimensional, ordinal scale representation exists for respondent ability (and the K response categories) when the observed data accord to the restrictions of W1 (and satisfy LI). Figure 1 graphically represents the W1 axiom, which states the following: For any ordinal response category t, if some respondent b is more probable than some respondent a to give a response $X \geq t$ for any item, then respondent a must never be more probable than respondent b to respond $X \geq t$, for all M items and all K response categories.

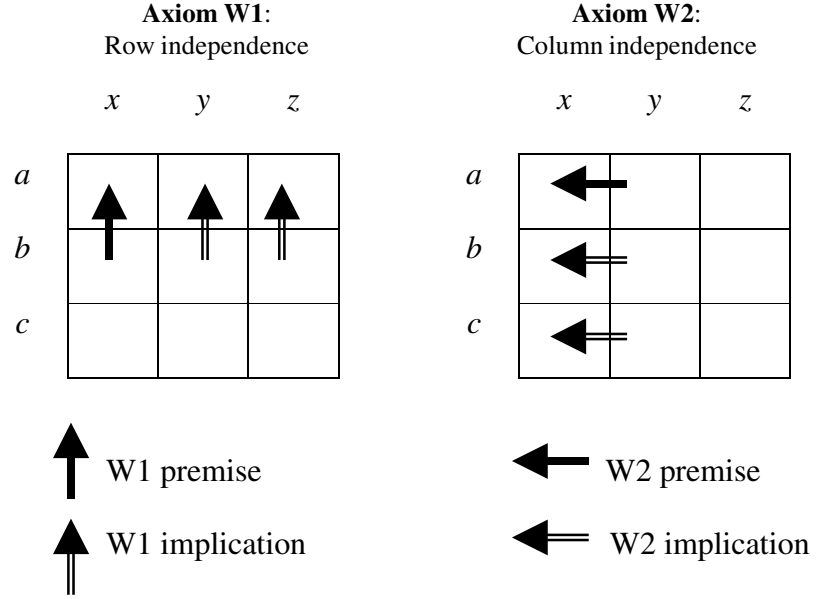

Figure 1. Graphical representation of the row and column independence axioms, pertaining to a single response category t.

A model characterized by W1 (and LI) specifies IRFs (for any particular response category t) to be non-decreasing over respondent ability, but the IFRs are free to intersect. Such a model has been given different names in item response theory (IRT). For example, in non-parametric IRT, it has been referred to as the monotone homogeneity model (Mokken, 1971), the monotone unidimensional latent trait model (Holland and Rosenbaum, 1986), and the strictly unidimensional model (Junker, 1993). Well-known

parametric IRT models characterized by W1 (and LI) include the two- and three-parameter logistic models (Lord and Novick, 1968).

Any model characterized by W1 and LI is a useful and flexible psychometric model, provided that the analyst is only interested in measuring respondents. However, if the analyst is interested in tailored testing, test equating, or item banking, a more restrictive model needs to be implemented. Since an invariant item scale lends to the practice of sample-free person and item measurement, the more restrictive model needs to specify non-decreasing and non-intersecting IRFs. Axiom W2 specifies non-intersecting IRFs:

Axiom W2: Weak item (column) independence.

For some person a, and some response category t,

if $P(t\,;\,ax) < P(t\,;\,ay)$,

then

$P(t'\,;\,bx) \leq P(t'\,;\,by)$, must also be true

for all persons $b \in A$, and all response categories $t' \in T$.

Axiom W2, also shown in Figure 1, is simply axiom W1 interchanging the respondents and items. Axiom W2 states that, the following: For any response category t, if for some respondent, the probability of a $X \geq t$ response is more probable for some item y than for some item x, then item x should never be more probable than y to yield a $X \geq t$ response, for all N persons and all K response categories.

A model characterized by both W1 and W2 (and LI) specifies IRFs (for any particular response category) to be non-decreasing and non-intersecting over respondent ability, where the item difficulty order is invariant over all K response categories. In non-parametric IRT, dichotomous-response special cases of this model include the double monotonicity model (Mokken, 1971), the U-model (Irtel, 1987), and the scalogram model (Guttman, 1950) is a deterministic special case. Scheiblechner's (1995) isotonic probabilistic model (ISOP) is completely characterized by W1, W2 and LI, as it can handle any finite number of ordinal response categories. It is also possible to perform person and test equating with ISOP (see Scheiblechner, 1995).

Figure 2 graphically represents the structure of axioms W1 and W2 simultaneously, for some response category t, where the vertical and horizontal order relations (arrows) refer to axioms W1 and W2, respectively.

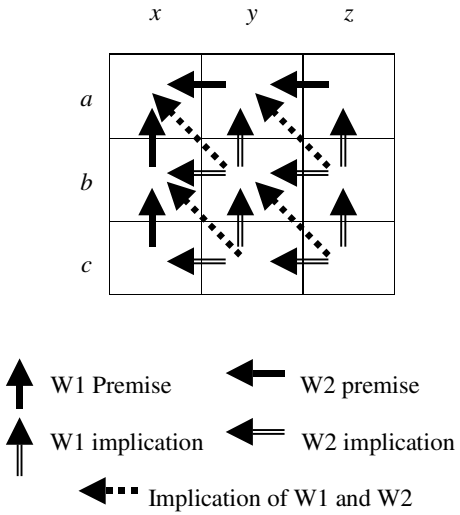

Figure 2. Simultaneous representation of axioms W1 and W2, pertaining to a single response category *t*.

The Figure illustrates a structure consistent with ordinal scale for both the respondents and items, because it implies the order $a < b < c$ for the respondents, and the order $x < y < z$ for the items. The $4 = (N - 1)(M - 1)$ diagonal-left order relations (arrows) are implications of the vertical (W1) and horizontal (W2) order relations. For example, $by \geq bx$ and $bx \geq ax$ implies $by \geq ax$ through transitivity. However, the set of horizontal, vertical, and diagonal-left order relations imply nothing about the directions of the $4 = (N - 1)(M - 1)$ diagonal-right order relations, not shown in Figure 2, but referring to pairs of respondent combinations bx and ay, by and az, cx and by, and cy and bz. A model represented by axioms W1 and W2 (and LI) does not require these pairs to have any particular order, which means that these axioms are not restrictive enough to specify a rank ordering on all $9 = NM$ respondent-item combinations (within any response category t). An axiomatic specification of the $(N - 1)(M - 1)$ diagonal right arrows, in addition W1 and W2, renders it possible to rank order the NM respondent-item combinations, for each of the K ordinal response categories.

Axioms W1, W2, and Co (and LI) characterize Scheiblechner's (1999) additive isotonic probabilistic model, ADISOP, and the completely additive isotonic probabilistic model, CADISOP. This set of axioms provides the necessary conditions for the additive interval scale representation of the respondents and items.

Axiom Co: Cancellation up to the empirically testable finite order o,

$o = \min\{(N - 1), (M - 1)\} - 1.$

Data accordance with these axioms supports the existence of a common "ordered-metric" scale for the respondents and items, which places between the ordinal and interval scale levels. The degree to which an ordered metric scale approximates an interval scale depends on the size of N, M, and the number of distinguishable levels of t (see Scheiblechner, 1999, pp. 300-301). While axioms W1 and W2 specify non-crossing IRFs for each of the K response categories, the cancellation conditions additionally restrict IRFs to be parallel within each response category t.

Axioms W1 and W2 are each also known as "single cancellation," since in each case, the premise consists of one inequality. "Cancellation to the order $o = 2$" refers to the double cancellation axiom, where the premise consists of two inequalities, which imply a third inequality:

Double cancellation:

If $P(t' ; bx) < P(t' ; ay)$ for some t',

and $P(t'' ; cy) < P(t'' ; bz)$ for some t'',

then $P(t ; cx) \leq P(t ; az)$ for all t.

Figure 3 graphically illustrates the double cancellation axiom, also shows cancellation to the order $o = 3$, otherwise known as "triple cancellation," where the premise contains three inequalities, which imply a fourth inequality.

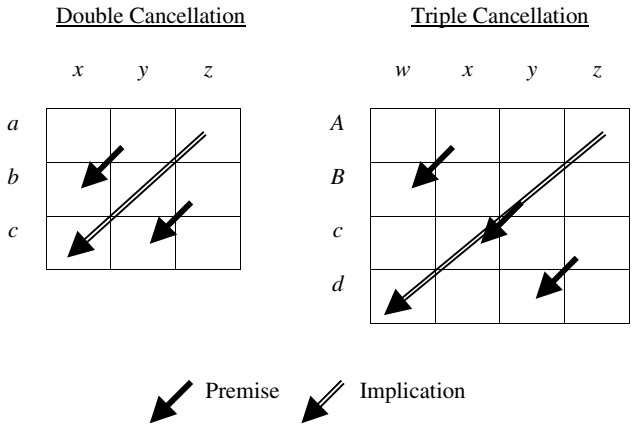

Figure 3. Graphical representation of double and triple cancellation.

Notice that the cancellation axioms specify diagonal-right order relations. Within any given response category *t*, the restrictions specified by W1, W2, and Co render it possible to specify the rank order of all the *NM* respondent-item combinations, *ax*, *ay*, *az*, *bx*, *by*,…, since the vertical, horizontal, and diagonal-left order relations of W1 and W2, in addition to the diagonal-right order relations of Co, provide enough information to infer the binary relations between all pairs of *NM* respondent-item combinations. If any of the *K* rank orders of the *NM* respondent-item combinations contain at least one tie, then the ADISOP results, where the *K* response categories can only be assumed to be measurable on an ordinal scale. If however all *K* rank orders of the *NM* respondent-item combinations contain no ties, then the CADISOP results, in which case the *K* response categories are measurable on a common interval scale with the respondents and items (see Scheiblechner, 1999).

ADISOP is equivalent to CADISOP in the case of dichotomous scored response data ($K = 2$, $T = \{t, t'\}$), where $P(X_{ax} \geq t)$ equals the same constant for all *NM* respondent-item combinations (e.g., 1.0). In such a case, response category *t* may be disregarded, and the analysis only needs to focus on only one response category *t*.

The axioms of the ADISOP model closely correspond to Rasch's (1960) definition of *specific objectivity*. In the context of the logistic Rasch model for dichotomous scored responses defined in (1), let

$$l_{ni} = \ell n\left(P_{ng} \Big/ \left(1 - P_{ng}\right)\right).$$

Since the model specifies a matrix $(P_{ng})_{N \times M}$ characterized by parallel IRFs, it is therefore conjoint additive, and therefore the matrix $(l_{ng})_{N \times M}$ is also conjointly additive. Specific objectivity specifies the ability distance between any arbitrary respondent pair *m* and *n*, $1 \leq m < n \leq N$, to equal the constant $l_{ng} - l_{mg} = l_{ng}$, for all items $1 \leq g \leq M$. In other words, the ability distance between any pair of persons is invariant over the test items used to compare them. Specific objectivity also specifies the difficulty distance of any arbitrary item pair *u* and *v*, $1 \leq u < v \leq L$, to equal the constant $l_{nu} - l_{nv} = l_{uv}$, for all respondents $1 \leq n \leq N$. This means that the difficulty distance between any item pair is invariant over all persons in the sample.

Although in the case of dichotomous response data, ADISOP and the logistic Rasch model closely correspond, they are not equivalent models. In fact, the Rasch model is a logistic special case of ADISOP. Along the same lines, it can be shown that the Rasch rating scale model (Andrich,

1978) is a special case of CADISOP (see Scheiblechner, 1999). Again, the conjoint measurement axioms that represent the ISOP, ADISOP, and CADISOP models does not require the shape of the IRFs to be represented by any specific function, such as the logistic or normal-ogive.

Irtel's (1987) defines *ordinal specific objectivity*, formulated by Schieblechner (1995) in the context of the ISOP model. Characterized by axioms *W1* and *W2* (ACM row and column independence axioms), ordinal specific objectivity formalizes specific objectivity of the person and items at the level of ordinal relations (instead of distance relations). For instance, consider again the framework of dichotomous scored test responses, and let P_{ng} be the probability of a correct response. Ordinal specific objectivity states that, for the ability comparison between all respondent pair m and n, $1 \leq m < n \leq N$, if $P_{ng} < P_{mg}$ for some item g, then $P_{ng} \leq P_{mg}$ for all remaining items $1 \leq h \leq M$, $h \neq g$. Also, for the difficulty comparison of all item pairs u and v, $1 \leq u < v \leq M$, if $P_{nu} < P_{nv}$ for some person n, then $P_{nu} \leq P_{nv}$ for all remaining persons $1 \leq m \leq N$, $n \neq m$. Although strictly increasing parallel IRFs, specified by the Rasch model, provide a framework for specific objectivity, non-intersecting and non-decreasing IRFs, specified by ISOP, provide a framework for specific objectivity that is *both more general and flexible*. In other words, the logistic Rasch model (either the dichotomous or rating scale model), relative to ISOP, is an additive conjoint special case, *and a special case of specific objectivity* (Scheiblechner, 1995).

Data framework

It is possible to relate the abstract ideas of probabilistic conjoint measurement theory to basic paradigms of standard categorical data analysis. Consider the context of an $I \times J \times K$ contingency table, where the rows $1 \leq i \leq I$ refer to the set of persons $A = \{a, b, c, \ldots\}$, the columns $1 \leq j \leq J$ refer to the set of items $Q = \{x, y, z, \ldots\}$, and the K layers, $1 \leq k \leq K$, refer to the set of K ordinal response categories $T = \{t, t', t'', t''', \ldots\}$, where $k = 1$ corresponds to category t, $k = 2$ refers to category t', $k = 3$ refers to category t'', and so forth. Each row $1 \leq i \leq I$ and column $1 \leq j \leq J$ can contain at least one respondent and item, respectively. Therefore, $I \leq N$ and $J \leq M$. The conjoint system

$$\langle A \times Q, P \rangle$$

can be related to the observed data structure with a *regular $I \times J \times K$ contingency table*, having the following definition.

> *Definition 1.* A *regular $I \times J \times K$ contingency table*, for each and every category system $1 \leq k \leq K$ (ordinal response category

t), specifies the order of the item response function $P(t; a, x)$ to be non-decreasing over the row index $1 \leq i \leq I$, and non-decreasing over the column index $1 \leq j \leq J$.

Note that since $P(t\,; a, x)$ is a cumulative item response function, the order $P(t; a, x) \geq P(t'\,; a, x) \geq P(t''\,; a, x) \geq \ldots$ automatically holds over the K category systems $k = 1, 2, 3, \ldots$, respectively, for all a (and hence $1 \leq i \leq I$) and all x (and hence $1 \leq j \leq J$).

One method to create a regular $I \times J \times K$ contingency table involves grouping and ordering the respondents and items by the simple scores X_{n+} and X_{+g}, respectively, since these scores are unbiased estimates of the ordinal position of a respondent and an item, respectfully (Scheiblechner, 1999). Furthermore, according to basic assumptions of unidimensional measurement, for any ordinal response category t, $P(t\,; a, x)$ should be a non-decreasing function of X_{n+} , and X_{+g} (e.g., Junker, 2000). This particular grouping and ordering scheme for the regular table can be considered provisional. For instance, the results of an analysis may suggest a slightly different grouping of the subjects or items in a subsequent analysis.

Each element of the *empirical regular $I \times J \times K$ contingency table* corresponds to the number of observed binomial "successes," $0 \leq n_{ijk} \leq N_{ijk}$, where $N_{ijk} \geq 1$ is the total number of binomial observations, and $\mathbf{n} = (n_{ijk})_{I \times J \times K}$. For instance, n_{ij2} is the observed number of $X \geq t'$ responses observed involving ij, and $N_{ij2} - n_{ij2}$ is the number of $X \ngeq t'$ responses. Of course, the maximum likelihood estimate (MLE) for the observed proportion of successes is given by $\hat{p}_{ijk} = n_{ijk} / N_{ijk}$, $0 \leq p_{ijk} \leq 1$. The set of proportions is given by $\hat{\mathbf{p}} = (\hat{p}_{ijk})_{I \times J \times K}$.

A statistical framework for axiomatic measurement theory

The task of estimating a non-parametric item response model, representing any set of measurement theory axioms, involves testing the hypothesis that the set of proportions $\hat{\mathbf{p}}$, based on the observed counts $\mathbf{n}$, stochastically conforms to the order restrictions implied by the axioms. This section formulates Markov Chain Monte Carlo (MCMC) methods of inference for estimating non-parametric item response models that can perform such a hypothesis test. Order restricted statistical inference can be routinely implemented with MCMC (e.g., Gelfand, Smith, and Lee, 1992), as it can provide a flexible framework to estimate even highly complex models (e.g., Carlin and Louis, 1998).

In the current context, the MCMC inference task involves estimating the posterior distribution of $\Theta = (\theta_{ijk})_{I \times J \times K}$. The quantity $0 \leq \theta_{ijk} \leq 1$ is the

expected binomial proportion for cell *ijk*, under the hypothesis that the data accord to particular axioms of the measurement theory. The posterior distribution of Θ, conditional on the observed data **n**, is defined by:

$$p\left(\Theta \mid \mathbf{n}\right) = \frac{L\left(\mathbf{n} \mid \Theta\right)\pi(\Theta)}{\int L\left(\mathbf{n} \mid \Theta\right)\pi(\Theta)}, \tag{4}$$

where L refers to the likelihood of the model, and p the prior distribution defined on $(\theta_{ijk})_{I \times J,\, k}$. For any response category k, the likelihood of the model, assuming local independence of item responses (Axiom LI), is given by:

$$L\left(\mathbf{n}_k \mid \Theta_k\right) = \prod_{i=1}^{I}\prod_{j=1}^{J} \theta_{ijk}^{n_{ijk}} \left(1 - \theta_{ijk}\right)^{N_{ijk} - n_{ijk}} . \tag{5}$$

Holding k constant, the observed proportions $\hat{\mathbf{p}}_k = (\hat{p}_{ijk})_{I \times J,\, k}$ are free to occupy any part of the IJ dimensional space $[0, 1]^{IJ}$, and the prior distribution $\pi(\Theta)$ is employed to constrain $\Theta_k = (\theta_{ijk})_{I \times J,\, k}$ to lie in a subset of this space, for $1 \leq k \leq K$. The prior constraints represent order restrictions on the θ_{ijk}, specified by a set of measurement axioms. The value of θ_{ijk} is constrained to have some minimum value $\min(\theta_{ijk})$, and some maximum value $\max(\theta_{ijk})$, $\min(\theta_{ijk}) \leq \max(\theta_{ijk})$. The minimum value $\min(\theta_{ijk})$ can either be 0 or one of the remaining $IJK - 1$ elements of Θ (excluding θ_{ijk}), and the maximum value $\max(\theta_{ijk})$ can either be 1 or one of the remaining IJK-1 elements of Θ (excluding θ_{ijk}), with the condition that $\min(\theta_{ijk})$ and $\max(\theta_{ijk})$ do not refer to the same element. Then borrowing ideas from Gelfand, Smith, and Lee (1992), the prior probability $\pi(\theta_{ijk})$ can be represented by the identifier function:

$$\pi\left(\theta_{ijk}\right) = \begin{cases} 1 & \text{if } \min\left(\theta_{ijk}\right) \leq \theta_{ijk} \leq \max\left(\theta_{ijk}\right) \\ 0 & \text{otherwise} \end{cases} . \tag{6}$$

There is no closed-form solution available for the complex integral in (4), the normalizing constant. However, this difficulty is easily circumvented through the implementation of a Metropolis-Hastings step in the MCMC estimation scheme, which renders the calculation of the integral unnecessary. For the estimation of the non-parametric item response models, the MCMC scheme simply involves the iterative sampling of each θ_{ijk}, conditional on previously sampled elements of Θ. A general MCMC algorithm is presented below for the estimation of the non-parametric item

response theory models of axiomatic measurement theory. The following algorithm assumes a framework of the regular $I \times J \times K$ contingency table.

Algorithm 1:

Let iteration t sample the IJK elements of Q in the order of:

$\theta_{111},...,\theta_{1J1};.....;\theta_{i11},...,\theta_{iJ1};.....;\theta_{I11},...,\theta_{IJ1};.....;\theta_{I1K},...,\theta_{1Jk};.....;\theta_{I1K},...,\theta_{IJK}$.

For any element, two steps decide $\theta^{(t)}_{ijk}$:

1. Generate $r_{ijk} \sim Unif[0,1]$ and $\theta^{*}_{ijk} \sim Unif\left[\min\left(\theta_{ijk}\right), \max\left(\theta_{ijk}\right)\right]$.

2. Decide: $\theta^{(t)}_{ijk} = \begin{cases} \theta^{*}_{ijk} & \text{iff } r_{ijk} \le \left(L\left[\mathbf{n}_k \mid \theta^{*}_{ijk}, rest\right] \Big/ L\left[\mathbf{n}_k \mid \theta^{(t-1)}_{ijk}, rest\right]\right) \\ \theta^{(t-1)}_{ijk} & \text{otherwise.} \end{cases}$

where $rest = \left\{\theta^{(t)}_{<i(1\le j\le J)k}, \theta^{(t)}_{i(<j)k}, \theta^{(t-1)}_{i(>j)k}, \theta^{(t-1)}_{>i(1\le j\le J)k}\right\}$.

To generate samples from the posterior distribution $p(\Theta \mid \mathbf{n})$, the algorithm is repeated for T iterations, $1 \le t \le T$, where T is a number in the thousands. In MCMC estimation theory, it is well known that, under certain regularity conditions, the degree to which the total number of iterations T approaches infinity is the degree to which the generated samples converge towards the true posterior distribution of the model (e.g., Carlin and Louis, 1998, Theorems 5.3, 5.4). After generating samples from the posterior distribution, it is recommended to discard the samples from the first B "burn-in" iterations, as they depend on potentially arbitrary starting values used for iteration $t = 0$.

In Algorithm 1, the two-step procedure characterizes the Metropolis-Hastings algorithms. In step 1, a random number r_{ijk} is sampled from the domain [0,1], and a candidate value θ^{*}_{ijk} is sampled from the order restrictions defined by the uniform prior distribution $Unif[\min(\theta_{ijk}),\max(\theta_{ijk})]$. Step 2 decides on a value $\theta^{(t)}_{ijk}$, which is either equal to θ^{*}_{ijk} or $\theta^{(t-1)}_{ijk}$, depending on the result of the likelihood ratio. The ratio actually is simplified from a ratio of posterior distributions

$$p\left(\theta^{*}_{ijk}, rest \,\middle|\, \left(\mathbf{n}_{ijk}\right)_{I\times J,k}\right) \Big/ p\left(\theta^{(t-1)}_{ijk}, rest \,\middle|\, \left(\mathbf{n}_{ijk}\right)_{I\times J,k}\right),$$

where the normalization constant (the integral) cancels out of the numerator and denominator, and the prior factor π is disregarded because Step 1 already specifies candidate values to be sampled within the constraints of the priors.

After the MCMC algorithm generates $T - B$ samples from the posterior distribution $p(\Theta \mid \mathbf{n})$, posterior moments are readily obtainable, such as the posterior mode $\hat{\theta}_{ij}$, mean $\tilde{\theta}_{ijk}$, standard deviation $\sigma(\tilde{\theta}_{ijk})$, and the 95% posterior interval of θ_{ijk}, defined by the 2.5 percentile lower bound $\sigma(\tilde{\theta}_{ijk})_{.025}$ and the 97.5 percentile upper bound $\sigma(\tilde{\theta}_{ijk})_{.975}$. The measurement axioms are then tested, simply, by checking whether the observed $\hat{p}_{ijk}$ are contained within the corresponding lower bound $\sigma(\tilde{\theta}_{ijk})_{.025}$ and upper bound $\sigma(\tilde{\theta}_{ijk})_{.975}$.

The inferential task of an "axiomatic" item response model is the reverse of standard IRT inference. Whereas an IRT model first estimates the person and item parameters from the observed, potentially noisy data to derive the IRFs and PRFs, a axiomatic non-parametric item response model uses data-free order restrictions to estimate the IRFs and PRFs.

The next section introduces non-parametric item response models that represent Scheiblechner's (1995; 1999) probabilistic conjoint measurement theory axioms. The models are also applied to analyze published and simulated data sets. The analyses were performed with a computer program written by the author in S-PLUS code (S-PLUS, 1995). The program can be directly obtained from the author.

Statistical Inference for Measruement Theory

Model IRT-W1

Model IRT-W1 specifies IRFs to be non-decreasing over ability, but allows IRFs to intersect. Given a $I \times J \times K$ regular contingency table of any finite size, the item response model under axiom W1 (and LI) is given by $0 \le \theta_{(i\text{-}1)jk} \le \theta_{ijk} \le \theta_{(i+1)jk} \le 1$, for all $1 \le i \le I$, $1 \le j \le J$, and $1 \le k \le K$, setting $\theta_{0jk} \equiv 0$ and $\theta_{(I+1)jk} \equiv 1$. Stated formally:

Item Response Theory Model W1 (IRT-W1):

$$0 \le \theta_{(i\text{-}1)jk} \le \theta_{ijk} \le \theta_{(i+1)jk} \le 1, \quad \forall i, \forall j, \forall k, \theta_{0jk} \equiv 0 \text{ and } \theta_{(I+1)jk} \equiv 1.$$

Model IRT-W1 is estimated by defining

$$\min\left(\theta_{ijk}\right) = \theta^{(t)}_{(i-1)\,jk} \text{ and}$$

$$\max\left(\theta_{ijk}\right) = \theta^{(t-1)}_{(i+1)\,jk}$$

in the first step of Algorithm 1.

		Items Hard ⟵⟶ Easy			
		4 **(*j* = 1)**	***9*** **(*j* = 2)**	***2*** **(*j* = 3)**	***N* per group**
Test Score Group	***3*** **(*i* = 1)**	.18	.33	.61	61
	4 **(*i* = 2)**	.52	.51	.64	84
	5 **(*i* = 3)**	.73	.68	.68	82

Note: Data obtained from Perline, Wright, and Wainer (1979), Table 2, p. 244.

Figure 4. An example 3 x 3 data set, containing the proportion of positive responses, by three respondent score groups and three items.

Perline, Wright, and Wainer (PWW, 1979, p.244) analyzed data from a 10-item dichotomous scored 0/1 test administered to 2500 released convicts (data collected by Hoffman and Beck, 1974), where the test inquires about the subject's criminal history. Figure 4 presents a 3 x 3 subset of the data set, in the form of a regular I x J regular table, where the rows order the score groups from lowest score to highest score from top to bottom, and the columns order the three items from most difficult to easiest from left to right. Each of the $9 = IJ$ cells contains p_{ij}, the proportion of positive responses, scored $X \geq 1$. Of course, the proportion of responses scored $X \geq 0$ equals 1 for all $1 \leq i \leq I$ and $1 \leq j \leq J$ for the dichotomous-scored questionnaire. Therefore, the estimated proportions of $X \geq 0$ responses is disregarded from subsequent data analysis, which explains the current focus on a 2-dimensional (I x J) regular contingency table, containing estimated proportions of $X \geq 1$ responses.

Model IRT-W1 was employed to analyze the data summarized in Figure 4, and Table 2 presents the results generated by the computer program (based on 2,000 MCMC iterations, 500 burn-in iterations). The table shows that all $9 = IJ$ cell observed proportions $\hat{p}_{ij}$ fall within the respective 95% posterior intervals, indicating that the data do not violate axiom W1, and therefore fit Model IRT-W1. Hence, the data clearly support a unidimensional ordinal-scale representation of the total test score (ability). Also observe that under Model IRT-W1, the data led to posterior modes $\hat{\theta}_{ij}$ characterized by intersecting IRFs, as shown in Figure 5. Under Model IRT-

Table 2

The IRT-W1 model estimated from the data set summarized in Figure 4.

CELL *i,j*	OBSERVED DATA n_{ij}	N_{ij}	$\hat{p}_{ij}$	POSTERIOR DISTRIBUTION ESTIMATES $\hat{\theta}_{ij}$	$\tilde{\theta}_{ij}$	$\sigma(\tilde{\theta}_{ij})$	2.5%ile	97.5%ile	Axiom Violation?
1,1	11	61	.18	.19	.23	.10	.05	.43	No
2,1	44	84	.52	.52	.52	.07	.39	.65	No
3,1	60	82	.73	.73	.73	.06	.61	.82	No
1,2	20	61	.33	.32	.33	.08	.18	.50	No
2,2	43	84	.51	.51	.51	.07	.38	.64	No
3,2	56	82	.68	.68	.68	.06	.57	.78	No
1,3	37	61	.61	.56	.55	.06	.41	.66	No
2,3	54	84	.64	.64	.64	.05	.55	.72	No
3,3	56	82	.68	.71	.71	.05	.62	.80	No

Notes:

n_{ij} is the number of positive responses in cell *ij*.

N_{ij} is the number of observed responses in cell *ij*.

$\hat{p}_{ij}$ is the observed proportion.

$\tilde{\theta}_{ij}$ the posterior mean, the expected proportion of cell *ij* under axiom W1.

$\sigma(\tilde{\theta}_{ij})$ is the posterior standard deviation.

2.5%ile, 97.5%ile the estimated lower and upper bounds of the 95% posterior interval of θ_{ij} .

Axiom Violation? "No" if $\hat{p}_{ij}$ is contained in the posterior interval of θ_{ij} , "Yes" otherwise.

W1, the estimated IRFs are always non-decreasing over the total test score, where it is possible for these IRFs to intersect.

A statistical framework for ISOP: Model IRT-W1W2

Scheiblechner's (1995) ISOP model specifies IRFs to be non-decreasing and non-intersecting over ability. Model IRT-W1, to be presented, provides a statistical framework for Scheiblechner's (1995) ISOP model. Given a $I \times J \times K$ regular contingency table of any finite size, the item response model under axioms W1 and W2 (and LI) is given by:

Item Response Theory Model W1W2 (IRT-W1W2):

$$0 \leq \max\{\theta_{(i-1)jk}, \theta_{i(j-1)k}\} \leq \theta_{ijk} \leq \min\{\theta_{(i+1)jk}, \theta_{i(j+1)k}\} \leq 1,$$
$$\forall i, \forall j, \forall k, \theta_{0jk} \equiv \theta_{i0k} \equiv 0 \text{ and } \theta_{(I+1)jk} \equiv \theta_{i(J+1)k} \equiv 1.$$

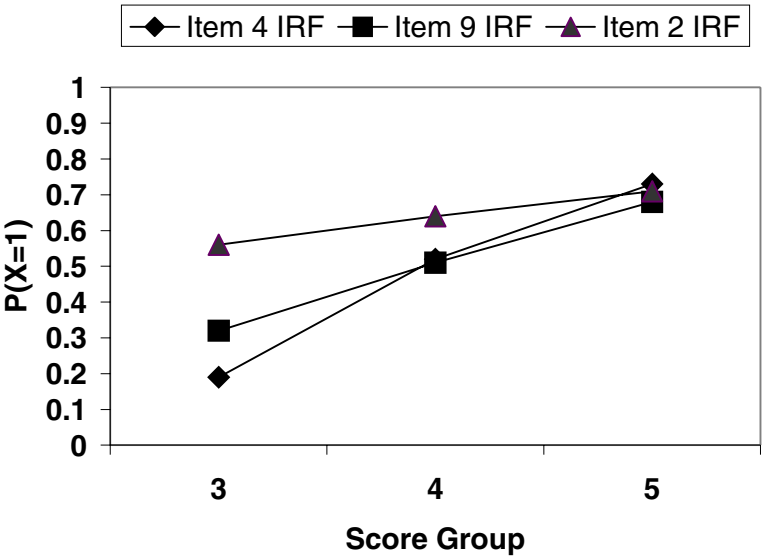

Figure 5. Item response functions of Model IRT-W1, estimated from the data set summarized in Figure 4, based on the posterior modes given in Table 2.

Model IRT-W1W2 is estimated by defining

$$\min\left(\theta_{ijk}\right) = \max\left\{\theta^{(t)}_{(i-1)jk}, \theta^{(t)}_{i(j-1)k}\right\} \text{and}$$

$$\max\left(\theta_{ijk}\right) = \min\left\{\theta^{(t-1)}_{(i+1)jk}, \theta^{(t-1)}_{i(j+1)k}\right\}$$

in the first step of Algorithm 1.

Notice the observed data summarized in Figure 4 in relation to Axiom W2. For instance, $\hat{p}_{11} < \hat{p}_{12} < \hat{p}_{13}$ is consistent with non-intersecting IRFs, while $\hat{p}_{21} > \hat{p}_{22} < \hat{p}_{23}$ and $\hat{p}_{31} > \hat{p}_{32} = \hat{p}_{33}$ appear slightly inconsistent with non-intersecting IRFs. However, judgements that declare these inconsistencies as definite violations to axiom W2 seem unreasonable and deterministic, because the inconsistencies may have only occurred by chance alone. Unfortunately, many well-known conjoint measurement studies have based their results on axiom tests that do not sufficiently assess the degree of stochastic approximation to the measurement axioms (e.g., Perline, Wright, and Wainer, 1979; Michell, 1990). These studies have either used the sign test, Kendall's rank correlation, and calculated confidence intervals around the observed proportions, to count the number of axiom violations. However, such statistics are arbitrary and have nothing to do with measurement theory axioms. The asymptotic distributions of these statis-

tics are largely disconnected with the order restrictions implied by any given set of measurement axioms.

It seems more reasonable to implement an analysis that can determine whether the observed data stochastically approximate non-intersecting IRFs, i.e., the set of order restrictions implied by axioms W1 and W2. A visual inspection Figure 4 suggests that the data stochastically approximate W1 and W2, however, since human judgement can be subjective, statistical inference is needed that does not depend on such subjectivity.

Model IRT-W1W2 provides this statistical inference in a straightforward fashion. The model analyzed the data summarized in Figure 4, and output of this analysis is given in Table 3 (based on 2,000 MCMC iterations, 500 burn-in iterations). The data demonstrated one minor axiom violation, related to cell 1,3 of the contingency table (group with score 5,

Table 3

The IRT-W1W2 (ISOP) model estimated from the data set summarized in Figure 4.

	OBSERVED DATA			***POSTERIOR DISTRIBUTION ESTIMATES***					
CELL i,j	n_{ij}	N_{ij}	$\hat{p}_{ij}$	$\hat{\theta}_{ij}$	$\tilde{\theta}_{ij}$	$\sigma(\tilde{\theta}_{ij})$	**2.5%ile**	**97.5%ile**	**Axiom Violation?**
1,1	11	61	.18	.15	.18	.08	.04	.36	No
2,1	44	84	.52	.47	.47	.06	.36	.57	No
3,1	60	82	.73*	.66	.66	.04	.57*	.725*	Yes
1,2	20	61	.33	.34	.35	.08	.19	.51	No
2,2	43	84	.51	.55	.55	.05	.45	.64	No
3,2	56	82	.68	.70	.70	.04	.63	.77	No
1,3	37	61	.61	.57	.57	.06	.44	.68	No
2,3	54	84	.64	.66	.66	.05	.57	.74	No
3,3	56	82	.68	.74	.74	.04	.66	.82	No

Notes:

n_{ij} is the number of positive responses in cell *ij*.

N_{ij} is the number of observed responses in cell *ij*.

$\hat{p}_{ij}$ is the observed proportion.

$\tilde{\theta}_{ij}$ the posterior mean, is the expected proportion of cell *ij* under axioms W1 and W2.

$\sigma(\tilde{\theta}_{ij})$ is the posterior standard deviation.

2.5%ile, 97.5%ile the estimated lower and upper bounds of the 95% posterior interval of θ_{ij} .

Axiom Violation? "No" if $\hat{p}_{ij}$ is contained in the posterior interval of θ_{ij} , "Yes" otherwise.

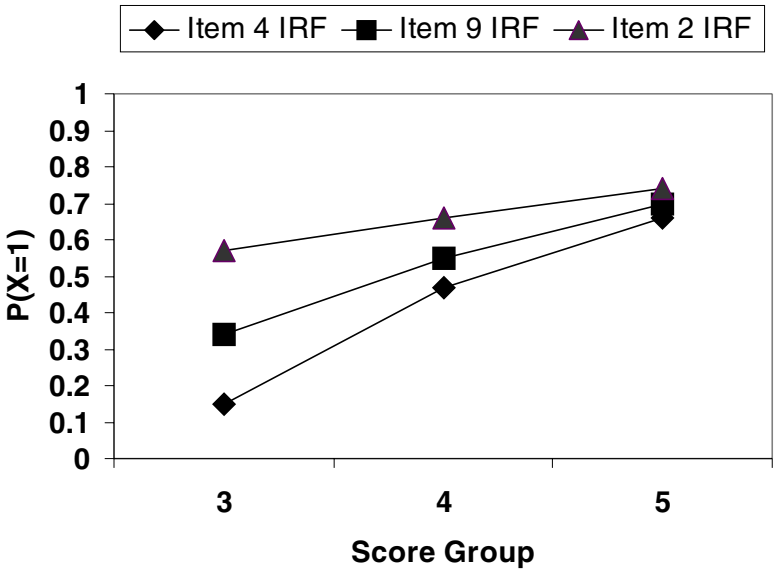

Figure 6. Item response functions of Model IRT-W1W2, estimated from the data set summarized in Figure 4, based on the posterior modes given in Table 3.

item 4). Note that, according to the posterior modes $\hat{\theta}_{ij}$ given in Table 3, Model IRT-W1W2 estimated non-intersecting IRFs. These IRFs are graphically represented in Figure 6. The estimated IRFs under Model IRT-W1W2 are always non-intersecting, and non-decreasing over the total test score.

Since Model IRT-W1 is more flexible than Model IRT-W1W2, i.e., has less data constraints, the former, of course, will have superior fit on any data set. Naturally, it is of interest to determine whether the additional constraints imposed by Model IRT-W1W2 significantly impacts model fit. To facilitate the fit comparison between the two models, the MCMC algorithm, in the previous two analyses involving Model IRT-W1 and IRT-W1W2, had recorded the model log-likelihood per iteration, which generated a distribution of log-likelihood values for each of the two models. The log-likelihood distribution of Model IRT-W1 had a mean of –246.53, while the log-likelihood distribution of Model IRT-W1W2 had a mean of –246.94. These results imply that, given the data set in Figure 4, Model IRT-W1 roughly provides about the same level of fit as Model IRT-W1W2.

Hence, relative to Model IRT-W1, the axiom violation under Model IRT-W1 was not severe enough to reject Model IRT-W1W2. The data seem to stochastically support the existence of ordinal specific objectivity, which also implies support for an ordinal representation of the total test score (ability) and the total item score (item easiness or difficulty).

A statistical framework for ADISOP: Model IRT-W1W2Co

Given a 3 x 3 x *K* regular contingency table, one example of an item response model under axioms W1, W2, and Co (and LI) is given by:

Item Response Theory Model W1W2Co (IRT-W1W2Co):

$0 \leq \mathrm{B} \leq \theta_{ijk} \leq \mathrm{C} \leq 1$, $\forall i$, $\forall j$, $\forall k$, where

$\mathrm{B} = \max\{\theta_{(i\text{-}1)jk}, \theta_{i(j\text{-}1)k}, \theta_{(i+1)(j\text{-}1)k}\}$, setting $\theta_{0jk} \equiv \theta_{i0k} \equiv \theta_{(I+1)jk} \equiv \theta_{i(J+1)k} \equiv 0$, and

$\mathrm{C} = \min\{\theta_{(i+1)jk}, \theta_{i(j+1)k}, \theta_{(i\text{-}1)(j+1)k}\}$, setting $\theta_{0jk} \equiv \theta_{i0k} \equiv \theta_{(I+1)jk} \equiv \theta_{i(J+1)k} \equiv 1$.

This model is estimated by setting

$$\min\left(\theta_{ijk}\right) = \max\left\{\theta^{(t)}_{(i-1)\,jk}, \theta^{(t)}_{i(\,j-1)\,k}, \theta^{(t-1)}_{(i+1)(\,j-1)\,k}\right\} \text{and}$$

$$\max\left(\theta_{ijk}\right) = \min\left\{\theta^{(t-1)}_{(i+1)\,jk}, \theta^{(t-1)}_{i(\,j+1)\,k}, \theta^{(t)}_{(i-1)(\,j+1)\,k}\right\}$$

in the first step of Algorithm 1. This statistical model, particular to $I = J = 3$, is an example of Scheiblechner's ADISOP model (1999). The model above can be adapted to handle another instance of double cancellation, by moving $\theta_{(i+1)(j\text{-}1)k}$ into set C, and moving $\theta_{(i\text{-}1)(j+1)k}$ into set B.

Unfortunately, in the more general case of $I, J > 3$, there doesn't seem to be a straightforward general formulation for order restrictions based on cancellation conditions up to the order $o = \min\{(I-1), (J-1)\} - 1$. For example, in an I x J table where $I, J \geq 3$, there are $\binom{I}{3}\binom{J}{3}$ 3 x 3 submatrices to consider, only for double cancellation. It therefore remains an open problem to develop a model IRT-W1W2Co that handles a I x J x K regular contingency table of any size.

Figure 7 presents a different 3 x 3 submatrix of the same dataset examined in Perline, Wright, and Wainer (1979, p. 244). Notice in Figure 7 that the relations between the observed proportions $\hat{p}_{11} < \hat{p}_{21} > \hat{p}_{31}$ and $\hat{p}_{12} < \hat{p}_{22} > \hat{p}_{32}$ appear inconsistent with Axiom W1. The relations between the observed proportions $\hat{p}_{11} < \hat{p}_{12} < \hat{p}_{13}$ and $\hat{p}_{21} < \hat{p}_{22} < \hat{p}_{23}$ appear inconsistent with Axiom W2. And given $\hat{p}_{21} < \hat{p}_{12}$ and $\hat{p}_{32} < \hat{p}_{23}$, the relation $p_{31} > \hat{p}_{13}$ is slightly inconsistent with double cancellation. Model IRT-W1W2Co analyzed the data summarized in Figure 7 to determine if the observed proportions stochastically approximate the IRFs specified by axioms W1, W2, and double cancellation. The results of this analysis are presented in

		ITEMS Hard ⟵⟶ Easy			
		6 **(*j* = 1)**	***1*** **(*j* = 2)**	***8*** **(*j* = 3)**	***N* per group**
Test Score Group	***4*** **(*i* = 1)**	.18	.24	.12	84
	5 **(*i* = 2)**	.13	.33	.30	82
	6 **(*i* = 3)**	.13	.28	.64	86

Note: Data obtained from Perline, Wright, and Wainer (1979), Table 2, p. 244.

Figure 7. An example 3 x 3 data set, containing the proportion of positive responses, by three respondent score groups and three items.

Table 4, and they conclude one axiom violation, involving cell 1, 3 (score group 4, item 8). Figure 8 plots the IRFs based on the posterior modes, and it can be seen that the functions are approximately parallel (additive) and increasing over the test score.

Models IRT-W1 and IRT-W1W2 also analyzed the data summarized in Figure 7. The likelihood distribution of IRT-W1 had a mean of -117.08, while IRT-W1W2 had a mean of –117.13, and IRT-W1W2Co had a mean of –117.33. Hence, it seems that the axiom violation under IRT-W1W2Co was not "severe" enough to convincingly reject this model in favor of the two more flexible models, IRT-W1W2 and IRT-W1.

Analyzing data generated by the two-parameter logistic model

The two-parameter logistic model was used to generate a test data set containing 100 respondents and 6 dichotomous test items. The generating respondent abilities θ_n had a uniform distribution in [–3, 3], and the six generating items had difficulties $\delta_1 = -2$, $\delta_2 = 1$, $\delta_3 = 0$, $\delta_4 = -1$, $\delta_5 = 2$, $\delta_6 = 0$, and discriminations $\alpha_1 = 1.5$, $\alpha_2 = 2$, $\alpha_3 = .5$, $\alpha_4 = 1$, $\alpha_5 = .1$, $\alpha_6 = 2.5$. Eighty-one of the 100 generated respondents had non-perfect or non-zero test scores, $0 < X_{n+} < 6$, and therefore are the focus of analysis. The data set was deliberately generated to have a wide range in the item discriminations, so that the data unambiguously violate axioms W2, i.e., and therefore unambiguously misfit the logistic Rasch model (recall that the Rasch model specifies all item discriminations to be equal to some constant, such

Table 4

The IRT-W1W2Co (ADISOP) model estimated from the data set summarized in Figure 7.

CELL i,j	OBSERVED DATA n_{ij}	N_{ij}	$\hat{p}_{ij}$	POSTERIOR DISTRIBUTION ESTIMATES $\hat{\theta}_{ij}$	$\tilde{\theta}_{ij}$	$\sigma(\tilde{\theta}_{ij})$	2.5%ile	97.5%ile	Axiom Violation?
1,1	15	84	.18	.09	.11	.04	.03	.19	No
2,1	11	82	.13	.14	.15	.05	.07	.24	No
3,1	11	86	.13	.19	.21	.05	.11	.31	No
1,2	20	84	.24	.21	.22	.05	.13	.32	No
2,2	27	82	.33	.27	.28	.05	.19	.37	No
3,2	24	86	.28	.33	.34	.06	.23	.45	No
1,3	10	84	.12*	.32	.33	.06	.23*	.46*	Yes
2,3	25	82	.30	.41	.42	.07	.30	.56	No
3,3	55	86	.64	.65	.64	.07	.50	.76	No

Notes:

n_{ij} is the number of positive responses in cell *ij*.

N_{ij} is the number of observed responses in cell *ij*.

$\hat{p}_{ij}$ is the observed proportion.

$\tilde{\theta}_{ij}$ the posterior mean, the expected proportion of cell *ij* under axioms W1, W2, and Co.

$\sigma(\tilde{\theta}_{ij})$ is the posterior standard deviation.

2.5%ile, 97.5%ile the estimated lower and upper bounds of the 95% posterior interval of θ_{ij} .

Axiom Violation? "No" if $\hat{p}_{ij}$ is contained in the posterior interval of θ_{ij} , "Yes" otherwise.

as 1). A summary of the generated data is given in Figure 9. This Figure clearly shows non-trivial violations of axiom W2, and therefore provide strong evidence of intersecting IRFs. The data even suggest some strong contradictions to Axiom W1. Hence, the data do not seem to support even an ordinal scale representation for the respondents (ability) and the items (difficulty).

The WINSTEPS program performed a Rasch analysis of the generated data. The Rasch item analysis is given in Table 5, which concludes that the set of six items generally *fit* the Rasch model. The analysis concludes that only item 5 misfits, while items 2 and 6 fit the Rasch model better than expected. Also, the same analysis concludes that overall, the

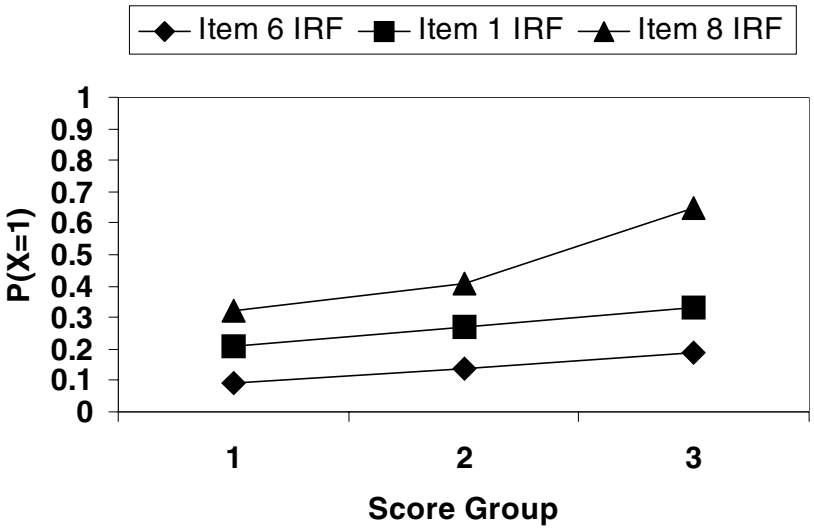

Figure 8. Item response functions of Model IRT-W1W2Co, estimated from the data set summarized in Figure 7, based on the posterior modes given in Table 4.

ITEMS

Hard ⟵ ⟶ Easy

		2 (*j*=1)	**5 (*j*=2)**	**6 (*j*=3)**	**3 (*j*=4)**	**4 (*j*=5)**	**1 (*j*=6)**	***N* per group**
Test Score Group	***1*** **(*i* = 1)**	.00	.28	.00	.06	.17	.50	18
	2 **(*i* = 2)**	.00	.36	.00	.43	.50	.71	14
	3 (*i* = 3)	.08	.46	.38	.69	.77	.62	13
	4 (*i* = 4)	.38	.50	.81	.44	.94	.94	16
	5 **(*i* = 5)**	.80	.45	1.0	.80	.95	1.0	20

Notes: The data set ($N = 81$ respondents with non-perfect or non-zero test scores, and $M = 6$ items) were generated under the uniform distribution of respondent abilities β_n in [-3,3]. The six items had difficulties $\delta_1 = -2$, $\delta_2 = 1$, $\delta_3 = 0$, $\delta_4 = -1$, $\delta_5 = 2$, $\delta_6 = 0$, and discriminations $\alpha_1 = 1.5$, $\alpha_2 = 2$, $\alpha_3 = .5$, $\alpha_4 = 1$, $\alpha_5 = .1$, $\alpha_6 = 2.5$.

Figure 9. A summary of data generated by the two-parameter logistic model.

81 respondents fit the model (mean person Infit = 1, s.d. = .34, mean person Outfit = .96, s.d. = .6). So, despite the fact that the generated data set, summarized in Figure 9, is explicitly inconsistent with parallel IRFs, and even explicitly consistent with non-intersecting IRFs, the Rasch model analysis concludes that, in general, the data fit the model quite well. Once again, the Rasch analysis obscured the severe measurement inconsistencies contained in the data structure, and concluded an overoptimistic view as to the scalability of the respondents and items. The over-optimism is a result of the Rasch model's data dependent frame of reference.

Model IRT-W1W2, a statistical framework for the ISOP model, ana-

Table 5

An output table from WINSTEPS, displaying the analysis of data consisting of 81 individuals responding to a 6-item test, dichotomous-response format.

ITEM STATISTICS: ENTRY ORDER

ITEM	RAW SCORE	COUNT	MEASURE	ERROR	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
1	62	81	-1.58	.30	1.10	.7	.89	-.3
2	23	81	1.39	.28	.72	-2.4	.51	-2.0
3	39	81	.20	.27	1.17	1.2	1.11	.6
4	54	81	-.91	.28	.83	-1.3	.81	-.8
5	33	81	.63	.27	1.64	4.2	1.99	3.8
6	38	81	.27	.27	.55	-4.2	.44	-3.9
MEAN	42.	81.	.00	.28	1.00	-.3	.96	-.5
S.D.	13.	0.	.98	.01	.36	2.7	.51	2.4

Notes:

RAW SCORE refers to the total item score X_{+i}.

COUNT the number of "non-extreme" respondents scoring $0 < X_{n+} < 10$.

MEASURE refers to the item difficulty estimate $\hat{\delta}_i$.

ERROR refers to the standard error of the estimate.

INFIT MNSQ Infit Mean square fit statistic of the item (expected value = 1).

INFIT ZSTD Unit normal transformation of the Infit mean square statistic (expected value = 0).

OUTFIT MNSQ Outfit Mean square fit statistic of the item (expected value = 1).

OUTFIT ZSTD Unit-normal transformation of the Oufit mean square statistic (expected value = 0).

lyzed the same data set. Recall that the frame of reference of this model, the order restrictions implied by measurement theory axioms W1 and W2, are not data dependent. Model IRT-W1W2 analyzed the data set summarized in Figure 9. The results of IRT-W1W2 analysis demonstrated that the model cor-

rectly identified the existence of several measurement inconsistencies (violations of axioms W1 and W2), as shown by the 10 axiom violations summarized in Table 6. Hence, even though ISOP is a much more flexible model than the logistic Rasch model, the results of the IRT-W1W2 analysis shows that ISOP was far better at identifying stochastic violations to the measurement axioms.

Table 6

The IRT-W1W2 (ISOP) model estimated from data generated by the 2-parameter logistic model.

	OBSERVED DATA			***POSTERIOR DISTRIBUTION ESTIMATES***					
CELL i,j	n_{ij}	N_{ij}	$\hat{p}_{ij}$	$\hat{\theta}_{ij}$	$\tilde{\theta}_{ij}$	$\sigma(\tilde{\theta}_{ij})$	2.5%ile	97.5%ile	**Axiom Violation?**
2,1	0	14	.00	.10	.13	.07	.02	.27	Yes
5,1	16	20	.80	.64	.64	.08	.46	.78	Yes
5,2	9	20	.45	.72	.70	.07	.54	.84	Yes
1,3	0	18	.00	.18	.21	.08	.07	.36	Yes
2,3	0	14	.00	.35	.36	.08	.21	.51	Yes
5,3	20	20	1.0	.86	.85	.05	.75	.93	Yes
1,4	1	18	.06	.27	.29	.09	.13	.46	Yes
4,4	7	16	.44	.74	.73	.07	.59	.85	Yes
1,5	3	18	.17	.38	.39	.09	.20	.55	Yes
3,6	8	13	.62	.82	.80	.06	.69	.91	Yes

Notes:

n_{ij} is the number of positive responses in cell *ij*.

N_{ij} is the number of observed responses in cell *ij*.

$\hat{p}_{ij}$ is the observed proportion.

$\tilde{\theta}_{ij}$ the posterior mean, the expected proportion of cell *ij* under axioms W1 and W2.

$\sigma(\tilde{\theta}_{ij})$ is the posterior standard deviation.

2.5%ile, 97.5%ile the estimated lower and upper bounds of the 95% posterior interval of θ_{ij}.

Axiom Violation? "No" if $\hat{p}_{ij}$ is contained in the posterior interval of θ_{ij}, "Yes" otherwise.

The more flexible Model IRT-W1, which allows intersecting IRFs, was also employed to analyze the same data set, and uncovered 3 axiom violations, therefore concluding that the data stochastically violates axiom W1. The mean log-likelihood of model IRT-W1W2 is –120.55, and the mean log-likelihood model IRT-W1 is –116.84, therefore, the data seem to provide more evidence in support of model IRT-W1 than model IRT-W1W2.

Conclusions

The Rasch model is a form of conjoint measurement, because the model's IRFs conform to the order restrictions implied by the set of conjoint additivity axioms. However, by definition, the IRFs under the Rasch model can only be data dependent, which leads to overoptimistic conclusions with regards to the scalability of observed data sets.

Therefore, statistical models need to be developed that explicitly use the data-free frame of reference granted by the conjoint measurement axioms. Since the axioms are expressed in deterministic language, the models need to be stochastic in the sense that they perform axiom testing while taking into account the fact that many data sets contain at least some degree of random noise.

To meet this need, this research introduced non-parametric item response models, namely IRT-W1, IRT-W1W2 to represent Scheiblechner's (1995) ISOP model, and IRT-W1W2Co to represent a case of Scheiblechner's (1999) ADISOP model. The non-parametric item response models are probabilistic measurement theory models, which integrate the ideas of axiomatic measurement theory, order restricted statistical inference, and MCMC inference. The presented models seem to provide useful stochastic tests of the conjoint measurement axioms.

Models IRT-W1, IRT-W1W2, and IRT-W1W2Co are not only specific to the task of testing measurement axioms. Simple modifications of these models can be made to perform other types of analyses common to psychometrics. For instance, item bias analysis can be performed between I different respondent subgroups among J items, where the respondent groups can be defined by gender, race, or income, for example. A model that performs item bias analysis is given by:

$$0 \leq \theta_{1jk} = \ldots = \theta_{ijk} = \ldots = \theta_{Ijk} \leq 1, \forall j, \forall k.$$

From the analysis results of this model, sources of item bias are simply uncovered by identifying the observed proportions $\hat{p}_{ijk}$ that are not contained in the respective 95% posterior intervals of θ_{ijk}. On the other hand, the psychometrician may be interested in determining whether scores among I respondent score groups are invariant over J item a-priori defined subscales, where each subscale j can contain two or more items. In this case, the following model can be applied:

$$0 \leq \theta_{i1k} = \ldots = \theta_{ijk} = \ldots = \theta_{iJk} \leq 1, \forall i, \forall k.$$

From the analysis results of this model, violations of score invariance over the J item subgroups are uncovered simply by identifying the observed proportions p_{ijk} that are not contained in the respective 95% posterior intervals of θ_{ijk}.

IRT-W1, IRT-W1W2, and IRT-W1W2Co are also useful models for the measurement of respondents and items, provided that the data stochastically approximate the relevant axioms. IRT-W1 is a flexible model when it is only of interest to measure the respondents on an ordinal scale. The ISOP model, with the statistical framework provided by model IRT-W1W2, can now be applied as a method for jointly scaling the respondents and items on a common ordinal scale. The specific objectivity of ISOP is more generally defined than the specific objectivity defined by the conventional logistic Rasch model (i.e., Rasch model specific objectivity is a special case of ISOP specific objectivity), and therefore model IRT-W1W2 can be applied in a straightforward fashion for the practice of sample-free respondent and item measurement.

Acknowledgements

This study was supported by Spencer Foundation research grant SG200100020, George Karabatsos, Principal Investigator. The statements made, and the data analyses presented, are solely the responsibility of the author.

References

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, *43*, 561-573.

Brogden, H. E. (1977). The Rasch model, the law of comparative judgement and additive conjoint measurement. *Psychometrika*, *42*, 631-634.

Bedrick, E. J. (1997). Approximating the conditional distribution of person-fit indexes for checking the Rasch model. *Psychometrika*, *62*, 191-199.

Brogden, H. E. (1976). The Rasch model, the law of comparative judgement, and additive conjoint measurement. *Psychometrika*, *24*, 473-505.

Carlin, B. P., and Louis, (1998). *Bayes and empirical Bayes methods for data analysis*. Boca Raton, FL: Chapman and Hall/CRC Press.

Cliff, N. (1973). Scaling. *Annual Review of Psychology*, *24*, 473-506.

Drasgow, F., Levine, M. V., and Williams, E. A. (1985). Appropriateness measurement with polychotomous item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology*, *38*, 67-86.

Fischer, G. (1968). *Psychologische testtheorie*. Bern: Huber.

Gefland, A. E., Smith, A. F. M., and Lee, T. M. (1992). Bayesian analysis of constrained parameter and truncated data problems using Gibbs sampling. *Journal of the American Statistical Association*, *87*, 523-532.

Guttman, L. (1950). *The basis for scalogram analysis.* In Stouffer, et al., *Measurement and Prediction, The American Soldier, Vol IV*. New York: Wiley.

Hoffman, P. B., and Beck, J. L. (1974). Parole decision making: A salient factor score. *Journal of Criminal Justice*, *2*, 195-206.

Holland, P. W., and Rosenbaum, P. R. (1986). Conditional association and unidimensionality in monotone latent variable models. *The Annals of Statistics*, *14*, 1523-1543.

Irtel, H. (1987). *On specific objectivity as a concept in measurement*. In E.E. Roskam and R. Suck (Eds.), Progress in Mathematical Psychology-1 (pp. 35-45). Amsterdam North, Holland: Elsevier.

Iverson, G., and Falmagne, J.-C. (1985). Statistical issues in measurement. *Mathematical Social Sciences*, *10*, 131-153.

Johnson, V. E., and Albert, J. H. (1999). *Ordinal data modeling*. New York: Springer.

Junker, B. W. (1993). Progess in characterizing strictly unidimensional IRT representations. *The Annals of Statistics*, *21*, 1359-1378.

Junker, B. W.(2000). *Monotonicity and conditional independence in models for student assessment and attitude measurement*. Technical report, Carnegie Mellon University , Department of Statistics. Website: http://www.stat.cmu.edu/~brian/bjtrs.html.

Karabatsos, G. (2000). A critique of Rasch residual fit statistics. *Journal of Applied Measurement*, *1*, 152-176.

Karabatsos, G. (2001). *Testing item response theory models with measurement theory axioms and order restricted inference*. Poster presentation, annual meeting of the International Society for Bayesian Analysis, Laguna Beach, CA, April.

Karabatsos, G., and Shev, C.-F. (2001). Testing measurement theory axioms using Markov Chain Monte Carlo. Paper presented at the 34th annual meeting of the Society for Mathematical Psychology, Brown University.

Keats, J. (1967). Test theory. *Annual Review of Psychology*, *18*, 217-238.

Linacre, J. M., and Wright, B. D. (1994). Chi-square fit statistics. *Rasch Measurement Transactions*, *8*, 350.

Linacre, J. M., and Wright, B. D. (2001). *A user's guide to WINSTEPS, Rasch measurement program*. Chicago: MESA Press.

Lord, F. M., and Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Luce, R. D., and Tukey, J. W. (1964). Additive conjoint measurement: A new type of fundamental measurement. *Journal of Mathematical Psychology*, *1*, 1-27.

Michell, J. (1990). *Introduction to the logic of psychological measurement*. Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Mokken, R. J. (1971). *A theory and procedure of scale analysis*. Paris/Den Haag: Mouton.

Molenaar, I. W., and Hoijtink, H. (1990). The many null distributions of person fit indices. *Psychometrika*, *55*, 75-106.

Nickerson, C. A., and McClelland, G. H. (1984). Scaling distortion in numerical conjoint measurement. *Applied Psychological Measurement*, *8*, 183-198.

Perline, R., Wright, B. D., and Wainer, H. (1979). The Rasch model as additive conjoint measurement. *Applied Psychological Measurement*, *3*, 237-255.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: The Danish Institute for Educational Research.

Robertson, T., Wright, F. T., and Dykstra, R. L. (1988). *Order Restricted Statistical Inference*. New York: John-Wiley and Sons.

Roskam, E. E., van den Wollenberg, A. L., and Jansen, G. W. (1986). The Mokken scale: A critical discussion. *Applied Psychological Measurement*, *10*, 265-277.

Scheiblechner, H. (1995). Isotonic ordinal probabilistic models. *Psychometrika*, *60*, 281-304.

Scheiblechner, H. (1999). Additive conjoint isotonic probabilistic models. *Psychometrika*, *64*, 295-316.

Sijitsma, K. and Meijer, R. R. (1992). A method for investigating the intersection of of item response functions in Mokken's non-parametric IRT model. *Applied Psychological Measurement*, *16*, 149-157.

Smith, R. M. (1988). The distributional properties of Rasch standardized residuals. *Educational and Psychological Measurement*, *48*, 657-667.

Smith, R.M. (1986). Person fit in the Rasch model. *Educational and Psychological Measurement*, *46*, 359-372.

Smith, R. M. (1996). Polytomous mean-square fit statistics. *Rasch Measurement Transactions*, *10*, 516-517.

Smith, R. M., Schumacker, R. E., and Bush, M. J. (1998). Using item mean squares to evaluate fit to the Rasch model. *Journal of Outcome Measurement*, *2*, 66-78.

S-PLUS (1995). *S-PLUS documentation*. Seattle: Statistical Sciences, Inc.

Tatsuoka, K. K. (1984). Caution indices based on item response theory. *Psychometrika*, *49*, 95-110.

Wright, B. D., and Masters, G. N. (1982). *Rating scale analysis*. Chicago: MESA Press.