PSYCHOMETRIKA—VOL. 75, NO. 2, 272–279
JUNE 2010
DOI: 10.1007/S11336-010-9147-7

A NOTE ON STOCHASTIC ORDERING OF THE LATENT TRAIT USING THE SUM OF POLYTOMOUS ITEM SCORES

L. ANDRIES VAN DER ARK

TILBURG UNIVERSITY

WICHER P. BERGSMA

LONDON SCHOOL OF ECONOMICS

In contrast to dichotomous item response theory (IRT) models, most well-known polytomous IRT models do not imply stochastic ordering of the latent trait by the total test score (SOL). This has been thought to make the ordering of respondents on the latent trait using the total test score questionable and throws doubt on the justifiability of using nonparametric polytomous IRT models for ordinal measurement. We show that a broad class of polytomous IRT models has a weaker form of SOL, denoted *weak SOL*, and argue that weak SOL justifies ordering respondents on the latent trait using the total test score and, therefore, the use of nonparametric polytomous IRT models for ordinal measurement.

Key words: Latent trait, monotone likelihood ratio, nonparametric item response theory, ordinal measurement, polytomous item response theory, polytomous items, stochastic ordering, total test score.

In the social and behavioral sciences, tests and questionnaires are frequently used to measure the position of respondents on a latent variable Θ (often called a latent trait). In item response theory (IRT) it is assumed that Θ explains the association between the item scores. An IRT model is used to model the item scores as a function of Θ and to measure the respondents' Θ values. A special class of IRT models consists of nonparametric IRT models (for an overview, see, e.g., Junker & Sijtsma, 2001; Sijtsma & Molenaar, 2002). A nonparametric IRT model consists of a set of weak assumptions about the relation between the item scores and Θ. The idea is to obtain useful measurement properties with as few restrictions on the data as possible. Let a test consist of J items each having $m+1$ ordered answer categories, which are scored $X_j = 0, 1, \ldots, m$ for $j = 1, \ldots, J$. For dichotomous item scores (i.e., $m = 1$), this set of assumptions may be:

Unidimensionality: Θ is unidimensional,
Local independence: The item scores are independent given Θ, and
Monotonicity: The probability of obtaining a score $X_j = 1$ given $\Theta = \theta$, denoted $P(X_j = 1|\theta)$, is a nondecreasing function of θ for all j

(e.g., see Sijtsma & Molenaar, 2002). Nonparametric IRT models that satisfy this set of assumptions include the monotone homogeneity model and the double monotonicity model (Mokken, 1971; also, see Sijtsma & Molenaar, 2002). Also, parametric IRT models, such as the Rasch (1960) model and the two- and three-parameter logistic models (Birnbaum, 1968) satisfy this set of assumptions.

In nonparametric IRT, the total test score, $X_+ = \sum_{j=1}^{J} X_j$, is used to measure a respondent's Θ value. For dichotomous item scores, Grayson (1988), Huynh (1994), Ünlü (2008) (also see

Requests for reprints should be sent to L. Andries van der Ark, Department of Methodology and Statistics, Faculty of Social and Behavioral Sciences, Tilburg University, 5000 LE, Tilburg, The Netherlands. E-mail: a.vdark@uvt.nl

Ghurye & Wallace, 1959) showed that unidimensionality, local independence, and monotonicity imply *monotone likelihood ratio* of X_+ in Θ (MLR), which is defined as

$$\frac{P(X_+ = K|\theta_A)}{P(X_+ = C|\theta_A)} \leq \frac{P(X_+ = K|\theta_B)}{P(X_+ = C|\theta_B)}$$

for $0 \leq C < K \leq Jm$ and for any two respondents A and B with $\theta_A < \theta_B$. Monotone likelihood ratio implies that Θ is stochastically ordered by X_+ (Lehmann, 1959, p. 74); that is,

$$P\big(\Theta > t|X_+ = C\big) \leq P\big(\Theta > t|X_+ = K\big) \quad \forall t, 0 \leq C < K \leq Jm. \tag{1}$$

Equation (1) is referred to as a *stochastic ordering of the latent trait by the total test score* X_+ (SOL; Hemker, Sijtsma, Molenaar, & Junker, 1997). Grayson's result implies that if unidimensionality, local independence, and monotonicity hold, it is reasonable to order respondents on the latent variable Θ using the observable test score X_+. For example, it follows from (1) that

$$E\big(\Theta|X_+ = C\big) \leq E\big(\Theta|X_+ = K\big).$$

In general, Grayson's result does not hold for polytomously scored items ($m > 1$). Hemker, Van der Ark, and Sijtsma (2001) provided the Venn diagram in Figure 1, showing the hierarchical relationships among 17 IRT models for polytomously scored items. In Figure 1, the nonparametric graded response model (np-GRM; Hemker et al., 1997; a.k.a. the monotone homogeneity model for polytomously scored items; Molenaar, 1997) is the most general model; it assumes unidimensionality, local independence, and a special form of monotonicity stating that $P(X_j \geq x|\theta)$ is nondecreasing in θ for $j = 1, \ldots, J$ and $x = 1, \ldots, m$. All other models depicted in Figure 1 imply these assumptions as well but they also have additional assumptions.

Only the partial credit model (Masters, 1982) and special cases of this model (e.g., the rating scale model, Andrich, 1978) imply SOL (Hemker et al., 1997, 2001). All other IRT models for polytomously scored items do not imply SOL. Hence, under well-known models such as the generalized partial credit model (Muraki, 1992), the graded response model (Samejima, 1969), and the np-GRM, there was no theoretical justification to order respondents on Θ using X_+. Sufficient conditions for SOL have been formulated for the generalized partial credit model (Van der Ark, 2005), but these conditions are so restrictive that they are unlikely to hold in practice. Van der Ark (2005) and DeMars (2008) used simulations to study conditions under which SOL is violated.

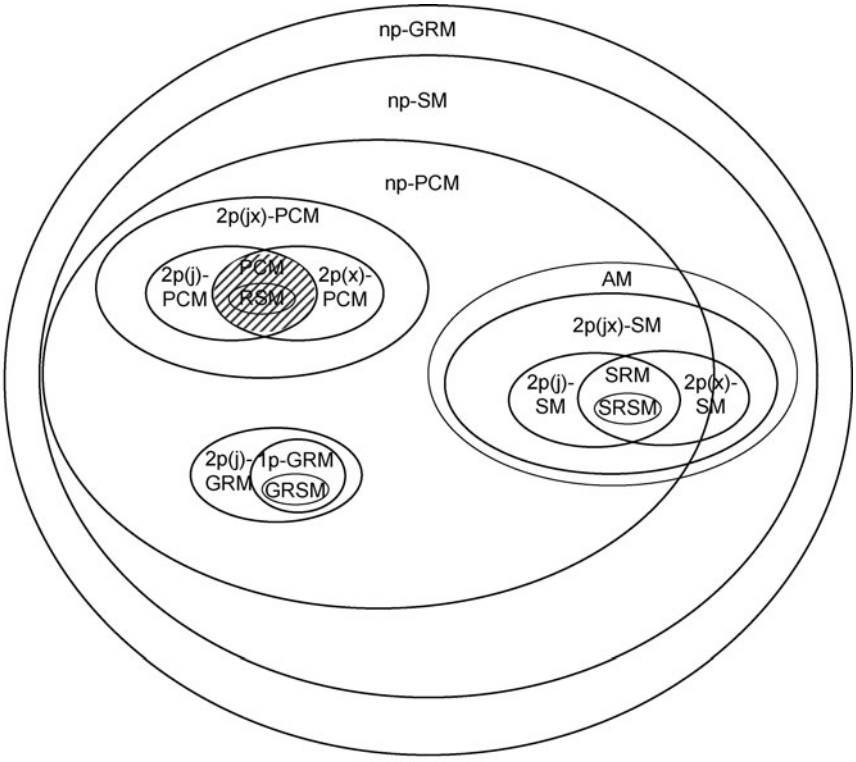
To alleviate these problems, we propose to modify SOL (1) to a weaker version, denoted weak SOL. Weak SOL holds if

$$P\big(\Theta > t|X_+ < K\big) \leq P\big(\Theta > t|X_+ \geq K\big) \quad \text{for all } t \text{ and } 0 < K \leq Jm. \tag{2}$$

We have some remarks on the relation of weak SOL to SOL and other ordering properties. First, the stronger property SOL (1) implies weak SOL (Lemma 1; Appendix). Second, weak SOL implies that $E(\Theta|X_+ < K) \leq E(\Theta|X_+ \geq K)$ for $K = 1, \ldots, Jm$ (e.g., Shaked & Shantikumar, 1994, p. 4). Third, weak SOL is equivalent to positive dependence in terms of global odds ratios, that is,

$$\frac{P(\Theta > t, X_+ \geq K)P(\Theta \leq t, X_+ < K)}{P(\Theta \leq t, X_+ \geq K)P(\Theta > t, X_+ < K)} \geq 1 \quad \text{for all } t \text{ and } 0 < K \leq Jm \tag{3}$$

(Lemma 2, Appendix). Positive dependence in terms of global odds ratios was studied by Douglas, Fienberg, Lee, Sampson, and Whitaker (1990) in the context of contingency tables with or-



np-GRM	nonparametric graded response model
np-SM	nonparametric sequential model
np-PCM	nonparametric partial credit model
AM	acceleration model
2p(jx)-SM	2pjx-sequential model
2p(j)-SM	2pj-sequential model
2p(x)-SM	2px-sequential model
SRM	sequential Rasch model
SRSM	sequential rating scale model
2p(j)-GRM	graded response model
1p-GRM	one parameter graded response model
GRSM	one parameter graded response model with rating scale restrictions
2p(jx)-PCM	2pjx-partial credit model
2p(j)-PCM	2pj-partial credit model (generalized partial credit model)
2p(x)-PCM	2px-partial credit model
PCM	partial credit model
RSM	rating scale model

FIGURE 1.
Venn diagram showing the hierarchical relationships among 17 polytomous IRT models. The least restrictive model is the nonparametric graded response model (np-GRM), the most restrictive models are the rating scale model (RSM), the sequential rating scale model (SRSM), and a rating scale version of the restricted graded response model (GRSM). Only the partial credit model (PCM) and the rating scale model (RSM), which have been depicted with a shaded background, imply SOL.

dinal variables. Fourth, a concept somewhat related to weak SOL was introduced by Scheiblechner (2002) (also, see Scheiblechner, 2007). He proposed the property of monotone likelihood ordering (MLO). Let X_{iA} and X_{iB} denote the score of respondents A and B on item i, respec-

tively; then MLO is defined as

$$P\big(\theta_A < \theta_B | X_{iA} < X_{iB}\big) > P\big(\theta_A > \theta_B | X_{iA} < X_{iB}\big)$$

for all pairs of respondents A and B and for $i = 1, \ldots, J$.

The main result of this note is a theorem stating that the most general IRT model, the np-GRM (see Figure 1), implies weak SOL (2). All other IRT models in Figure 1 are a special case of the np-GRM (see Van der Ark, 2001, for an overview of the proofs), and, therefore, a corollary of the theorem is that all IRT models in Figure 1 imply weak SOL.

Theorem. *The np-GRM implies weak SOL.*

Proof: Hemker et al. (1997, Theorem 1) showed that the np-GRM implies stochastic ordering of the manifest variable X_+ by Θ (abbreviated SOM). SOM means that

$$P\big(X_+ \geq K | \theta\big) \quad \text{is nondecreasing in } \theta \text{ for } 0 \leq K \leq Jm. \tag{4}$$

With $I(.)$ the indicator function, let I_K denote the binary random variable $I(X_+ \geq K)$, and let $f_{I_K,\Theta}$ denote the joint density of (I_K, Θ); this is a density with respect to the product of counting measure and Lebesgue measure. Also, let $f_{I_K|\Theta}$ denote the conditional density of I_K given Θ. Then

$$\begin{aligned}
&\text{Equation (4)} \\
\iff \quad & f_{I_K|\Theta}(1|\theta_B) \geq f_{I_K|\Theta}(1|\theta_A) \quad \forall \theta_A < \theta_B, 0 < K \leq Jm \\
\iff \quad & \frac{f_{I_K|\Theta}(1|\theta_B)}{f_{I_K|\Theta}(0|\theta_B)} \geq \frac{f_{I_K|\Theta}(1|\theta_A)}{f_{I_K|\Theta}(0|\theta_A)} \quad \forall \theta_A < \theta_B, 0 < K \leq Jm \\
\iff \quad & \frac{f_{I_K,\Theta}(1,\theta_B)}{f_{I_K,\Theta}(0,\theta_B)} \geq \frac{f_{I_K,\Theta}(1,\theta_A)}{f_{I_K,\Theta}(0,\theta_A)} \quad \forall \theta_A < \theta_B, 0 < K \leq Jm \\
\iff \quad & f_{I_K,\Theta}(1,\theta_B) f_{I_K,\Theta}(0,\theta_A) \geq f_{I_K,\Theta}(1,\theta_A) f_{I_K,\Theta}(0,\theta_B) \\
& \quad \forall \theta_A < \theta_B, 0 < K \leq Jm.
\end{aligned} \tag{5}$$

By integrating both sides over $\theta_A \leq t$ and $\theta_B > t$, (5) yields

$$\begin{aligned}
&P(X_+ \geq K, \Theta > t) P(X_+ < K, \Theta \leq t) \geq P(X_+ < K, \Theta > t) P(X_+ \geq K, \Theta \leq t) \\
&\quad \text{for all } t \text{ and for } 0 < K \leq Jm,
\end{aligned} \tag{6}$$

from which (3) immediately follows. It follows from Lemma 2 (Appendix) that (3) is equivalent to weak SOL. □

A numerical example illustrates that under particular item response theory models SOL can be violated whereas weak SOL holds.

Example (The graded response model implies weak SOL but does not imply SOL). Assume that the response probabilities of two trichotomous items are given by a graded response model; that is,

$$P\big(X_j \geq x | \theta\big) = \frac{\exp(\alpha_j(\theta - \beta_{jx}))}{1 + \exp(\alpha_j(\theta - \beta_{jx}))}$$

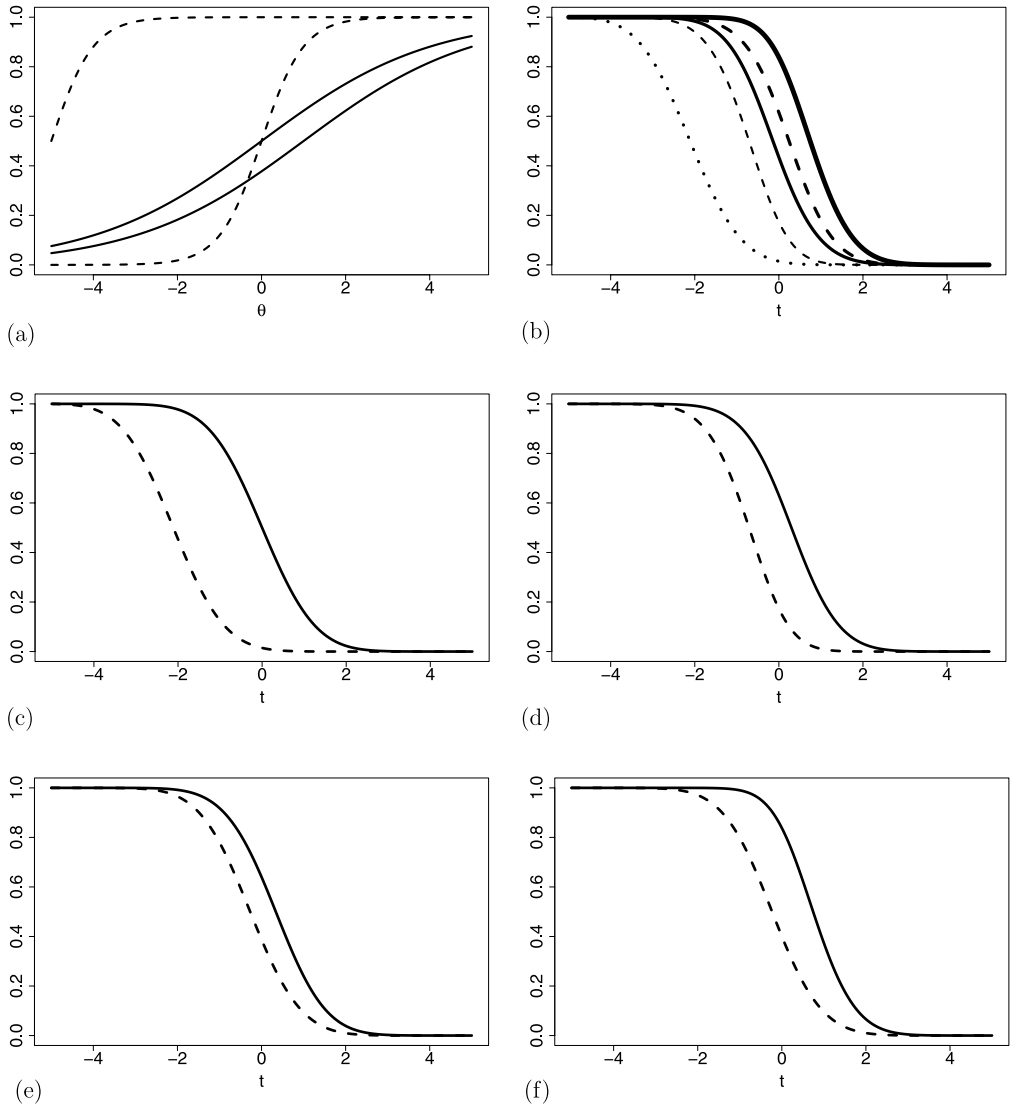


FIGURE 2.
Six plots illustrating weak SOL and a violation of SOL for two trichotomous items under the graded response model. For details, see text. (**a**) $P(X_j \geq x|\theta)$ as a function of θ for $x = 1, 2$. (**b**) $P(\Theta > t|X_+ = K)$ as a function of t for $K = 0, \ldots, 4$. $P(\Theta > t|X_+ < K)$ and $P(\Theta > t|X_+ \geq K)$ as a function of t for $K = 1$ (**c**), $K = 2$ (**d**), $K = 3$ (**e**), and $K = 4$ (**f**).

for $j = 1, 2$ and $x = 1, 2$, with discrimination parameters $\alpha_1 = \frac{1}{2}$ and $\alpha_2 = 2$, and location parameters $\beta_{11} = \beta_{22} = 0$, $\beta_{12} = -1$, and $\beta_{21} = -5$. Also, assume that Θ has a standard normal density (we approximated the standard normal density by a histogram of 10001 equally sized intervals of Θ in the range $[-5; 5]$). Figure 2a shows the two item step response functions $P(X_j \geq x|\theta)$, $x = 1, 2$, for item 1 (solid line) and item 2 (dashed line). Figure 2b shows the conditional probabilities $P(\Theta > t|X_+ = x_+)$ as a function of t for $x_+ = 0$ (dotted line), $x_+ = 1$ (dashed thin line), $x_+ = 2$ (dashed line), $x_+ = 3$ (solid line), and $x_+ = 4$ (solid thick line). The lines in Figure 2b are nonincreasing by definition. An incorrect ordering of the lines in terms of (1) for at least some values of t indicates a violation of SOL. Figure 2b shows that SOL is violated because

TABLE 1.
Values of $E(\Theta|X_+ = K)$, $E(\Theta|X_+ \leq K)$, and $E(\Theta|X_+ > K)$ for $K = 0, \ldots, 4$ for the graded response model in the Example, rounded to three decimals. Violations of SOL are printed in boldface.

K	$E(\Theta\|X_+ = K)$	$E(\Theta\|X_+ < K)$	$E(\Theta\|X_+ \geq K)$
0	−2.103	NA	0.000
1	−0.734	−2.103	0.001
2	**0.233**	−0.736	0.295
3	**−0.125**	−0.266	0.333
4	0.773	−0.226	0.773

for almost all values of t (i.e., $t \in [-4.658; 4.993]$), $P(\Theta > t|X_+ = 2) > P(\Theta > t|X_+ = 3)$. The lines in Figures 2c, d, e, and f show $P(\Theta > t|X_+ < K)$ (dashed line) and $P(\Theta > t|X_+ \geq K)$ (solid line) for $K = 1, \ldots, 4$, respectively, as functions of t. A violation of weak SOL would be indicated by an intersection. Because the graded response model implies weak SOL, there are no intersections. Table 1 shows the values of $E(\Theta|X_+ = K)$, $E(\Theta|X_+ < K)$, and $E(\Theta|X_+ \geq K)$. The expected latent trait value is less for a respondent with $X_+ = 3$ than for a respondent with $X_+ = 2$ indicating a violation of SOL. Using weak SOL means comparing $E(\Theta|X_+ < K)$ and $E(\Theta|X_+ \geq K)$ for $K = 0, \ldots, 4$. Note that $E(\Theta|X_+ \geq 0) = E(\Theta) = 0$. Also note that in this particular example, $E(\Theta|X_+ < K)$ and $E(\Theta|X_+ \geq K)$ are increasing in K. In general, this need not be true.

The theorem shows that all popular nonparametric IRT models for polytomously scored items can be used for ordinal person measurement; yet the ordering properties are weaker than SOL or monotone likelihood ratio. The papers of Hemker et al. (1996, 1997, 2001), in which it was shown that nonparametric IRT models do not imply SOL and monotone likelihood ratio, may have led to the belief that there is no justification for nonparametric IRT models for polytomous item scores. The theorem provides this justification. The difference between SOL and weak SOL in applications was illustrated in the example. Whereas SOL allows ordering of the respondents' expected latent trait values based on individual total test scores, weak SOL allows ordering of the expected latent trait values for a high total test score group on the one hand and a low total test score group on the other hand.

Acknowledgements

We would like to thank three anonymous reviewers for their careful reading and useful suggestions.

Appendix

Lemma 1. *SOL implies weak SOL.*

Proof: Starting with SOL (1), we obtain:

$$\text{SOL} \iff \frac{P(\Theta > t|X_+ = C)}{P(\Theta \leq t|X_+ = C)} \leq \frac{P(\Theta > t|X_+ = K')}{P(\Theta \leq t|X_+ = K')} \quad \forall t, 0 \leq C < K' \leq Jm$$

$$
\begin{aligned}
&\iff \quad \frac{P(\Theta > t, X_+ = C)}{P(\Theta \leq t, X_+ = C)} \leq \frac{P(\Theta > t, X_+ = K')}{P(\Theta \leq t, X_+ = K')} \quad \forall t, 0 \leq C < K' \leq Jm \\
&\iff \quad P(\Theta > t, X_+ = C) P\big(\Theta \leq t, X_+ = K'\big) \\
&\qquad\quad \leq P\big(\Theta > t, X_+ = K'\big) P(\Theta \leq t, X_+ = C) \quad \forall t, 0 \leq C < K' \leq Jm \\
&\iff \quad P\big(X_+ = K', \Theta > t\big) P(X_+ = C, \Theta \leq t) \\
&\qquad\quad \geq P(X_+ = C, \Theta > t) P\big(X_+ = K', \Theta \leq t\big) \quad \forall t, 0 \leq C < K' \leq Jm.
\end{aligned}
$$

Summing both sides of the last inequality over $C < K$ and $K' \geq K$ yields (6), which implies weak SOL (see the lines below (6)). □

Lemma 2. *Weak SOL and* (3) *are equivalent.*

Proof: We have

$$
\begin{aligned}
\text{Equation (3)} \quad &\iff \quad \frac{P(\Theta > t, X_+ \geq K)}{P(\Theta \leq t, X_+ \geq K)} \geq \frac{P(\Theta > t, X_+ < K)}{P(\Theta \leq t, X_+ < K)} \quad \forall t, 0 < K \leq Jm \\
&\iff \quad \frac{P(\Theta > t | X_+ \geq K)}{P(\Theta \leq t | X_+ \geq K)} \geq \frac{P(\Theta > t | X_+ < K)}{P(\Theta \leq t | X_+ < K)} \quad \forall t, 0 < K \leq Jm \\
&\iff \quad \frac{P(\Theta > t | X_+ \geq K)}{1 - P(\Theta > t | X_+ \geq K)} \geq \frac{P(\Theta > t | X_+ < K)}{1 - P(\Theta > t | X_+ < K)} \quad \forall t, 0 < K \leq Jm \\
&\iff \quad P(\Theta > t | X_+ \geq K) \geq P(\Theta > t | X_+ < K) \quad \forall t, 0 < K \leq Jm,
\end{aligned}
$$

which is weak SOL. □

References

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

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In Lord, F.M., & Novick, M.R. (Eds.) *Statistical theories of mental test scores* (pp. 395–480). Reading: Addison-Wesley.

DeMars, C.E. (2008). Polytomous differential item functioning and violations of ordering of the expected latent trait by the raw score. *Educational and Psychological Measurement*, *68*, 379–396.

Douglas, R., Fienberg, S.E., Lee, M.L.T., Sampson, A.R., & Whitaker, L.R. (1990). Positive dependence concepts for ordinal contingency tables. In Block, H.W., Sampson, A.R., & Savits, T.H. (Eds.), *Topics in statistical dependence* (pp. 189–202). Hayward: Institute of Mathematical Statistics. Retrieved September 13, 2009, from http://projecteuclid.org/DPubS?verb=Display&version=1.0&service=UI&handle=euclid.lnms/1215457559&page=record.

Grayson, D.A. (1988). Two-group classification in latent trait theory: scores with monotone likelihood ratio. *Psychometrika*, *53*, 383–392.

Ghurye, S.G., & Wallace, D.L. (1959). A convolutive class of monotone likelihood ratio families. *Annals of Mathematical Statistics*, *30*, 1158–1164.

Hemker, B.T., Sijtsma, K., Molenaar, I.W., & Junker, B.W. (1996). Polytomous IRT models and monotone likelihood ratio of the total score. *Psychometrika*, *61*, 679–693.

Hemker, B.T., Sijtsma, K., Molenaar, I.W., & Junker, B.W. (1997). Stochastic ordering using the latent trait and the sum score in polytomous IRT models. *Psychometrika*, *62*, 331–347.

Hemker, B.T., Van der Ark, L.A., & Sijtsma, K. (2001). On measurement properties of continuation ratio models. *Psychometrika*, *66*, 487–506.

Huynh, H. (1994). A new proof for monotone likelihood ratio for the sum of independent Bernoulli random variables. *Psychometrika*, *59*, 77–79.

Junker, B.W., & Sijtsma, K. (2001). Nonparametric item response theory in action: an overview of the special issue. *Applied Psychological Measurement*, *25*, 211–220.

Lehmann, E.L. (1959). *Testing statistical hypotheses*. New York: Wiley.

Masters, G. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*, 149–174.

Mokken, R.J. (1971). *A theory and procedure of scale analysis*. The Hague: Mouton/De Gruyter.

Molenaar, I.W. (1997). Nonparametric models for polytomous responses. In van der Linden, W.J., & Hambleton, R.K. (Eds.), *Handbook of modern item response theory* (pp. 369–380). New York: Springer.

Muraki, E. (1992). A generalized partial credit model: application of an EM algorithm. *Applied Psychological Measurement*, *16*, 159–177.
Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Nielsen & Lydiche.
Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph, 17*.
Scheiblechner, H. (2002). *Nonparametric IRT: scoring functions and ordinal parameter estimation of isotonic probabilistic models (ISOP)*. Unpublished manuscript. Retrieved September 13, 2009, from http://www.staff.uni-marburg.de/~scheible/Isoscore2.pdf
Scheiblechner, H. (2007). A unified nonparametric IRT model for d-dimensional psychological test data (d-ISOP). *Psychometrika*, *72*, 43–67.
Shaked, M., & Shantikumar, J.G. (1994). *Stochastic orders and their applications*. San Diego: Academic Press.
Sijtsma, K., & Molenaar, I.W. (2002). *Introduction to nonparametric item response theory*. Thousand Oaks: Sage.
Ünlü, A. (2008). A note on monotone likelihood ratio of the total score variable in unidimensional item response theory. *British Journal of Mathematical and Statistical Psychology*, *61*, 179–187.
Van der Ark, L.A. (2001). Relationships and properties of polytomous item response theory models. *Applied Psychological Measurement*, *25*, 273–282.
Van der Ark, L.A. (2005). Stochastic ordering of the latent trait by the sum score under various polytomous IRT models. *Psychometrika*, *70*, 283–304.

Manuscript Received: 30 MAR 2009
Final Version Received: 11 SEP 2009
Published Online Date: 30 JAN 2010