UvA-DARE (Digital Academic Repository)

Weighted Guttman errors: Handling ties and two-level data

Koopman, L.; Zijlstra, B.J.H.; van der Ark, L.A.

Link to publication

Weighted Guttman Errors: Handling Ties and Two-Level Data

Letty Koopman, Bonne J. H. Zijlstra, and L. Andries van der Ark

Abstract We provide an introduction to weighted Guttman errors and discuss two problems in computing weighted Guttman errors that are currently not handled correctly by all software: Handling ties—that is, computing weighted Guttman errors when two items have the same estimated popularity—and computing weighted Guttman errors when the data have a two-level structure. Handling ties can be incorporated easily in existing software. For computing weighted Guttman errors for two-level data, we provide an R function.

Keywords Guttman errors • Item ordering • Mokken scale analysis • Multilevel test data • Nonparametric item response theory

1 Introduction

For a pair of dichotomous items in descending order of popularity, a Guttman error (Guttman 1950) occurs if a respondent answers negatively to the first (more popular or easier) item and positively to the second (less popular or more difficult) item. Hence, if item 1 is more popular than item 2, the item-score vector $(0, 1)$ constitutes a Guttman error, whereas $(0, 0)$, $(1, 0)$, and $(1, 1)$ are admissible item-score vectors. Guttman errors are violations of the deterministic Guttman (1950) scale. Guttman errors are used for detecting outliers (e.g., Zijlstra et al. 2007) and aberrant response patterns (e.g., Meijer 1994 Karabatsos 2003) and for computing Mokken's (1971) scalability coefficients in Mokken scale analysis (Sijtsma and Molenaar 2002; also see Sijtsma and Van der Ark 2017; Snijders 2001a). For a pair of polytomous items, multiple item-score vectors can constitute a Guttman error, making both the calculation and the interpretation of Guttman errors more complicated. Molenaar (1991) proposed to weight the Guttman errors to acknowledge that the degree in which item-score vectors are aberrant may differ. For example, consider two polytomous items, each having ordered answer categories $0, 1, 2, 3, 4$. Suppose

L. Koopman (✉) • B.J.H. Zijlstra • L.A. van der Ark
Research Institute of Child Development and Education, University of Amsterdam,
P.O. Box 15776, 1001 NG, Amsterdam, The Netherlands
e-mail: V.E.C.Koopman@uva.nl

item 1 is more popular than item 2, then item-score vector $(0, 4)$ is more aberrant than item-score vector $(0, 1)$.

In recent work on deriving standard errors for two-level scalability coefficients (Koopman 2016), we encountered two problems in estimating the weights of Guttman errors: Estimated weights depend on the value of a random seed when two or more estimated *item popularities* are equal, and estimated weights may be biased for two-level data. In this chapter, we first introduce weighted Guttman errors, then we discuss the two problems and offer a solution for each problem, and finally we discuss some additional features of (two-level) weight computations.

2 Weighted Guttman Errors

2.1 Theory

Let a test consist of J items with $m + 1$ ordered response categories indexed by x $(x = 0, 1, \ldots, m)$. Let X_j denote the item score of item j. Each item score consists of m item steps (Molenaar 1983), binary variables denoted Z_{jx} $(j = 1, \ldots, J; x = 1, \ldots, m)$. $Z_{jx} = 1$ if $X_j \geq x$ (the item step was passed) and $Z_{jx} = 0$ if $X_j < x$ (the item step was failed). It follows that $Z_{j,x-1} \geq Z_{jx}$ and $X_j = \sum_x Z_{jx}$. For example, if $X_j = 1$ and $m = 3$, then $Z_{j1} = 1$, $Z_{j2} = 0$, and $Z_{j3} = 0$. Let the popularity of item step Z_{jx} be the probability of having a score of at least x on item j: $P(Z_{jx}) \equiv P(X_j \geq x)$. Note that by definition, $P(X_j \geq 0) = 1$. Let z_{njx} denote the realization of Z_{jx} for person n, then, in a sample of N respondents, $P(X_j \geq x)$ is estimated by

$$\widehat{P}\left(X_j \geq x\right) = \frac{1}{N} \sum_{n=1}^{N} z_{njx}. \tag{1}$$

Item pair (i, j) has $2m$ item steps: $Z_{i1}, \ldots, Z_{im}, Z_{j1}, \ldots, Z_{jm}$. For the purpose of determining weighted Guttman errors, the $2m$ item steps are put in descending order of their popularity. For example, Table 1 shows $J = 2$ items with $m + 1 = 3$ ordered response categories, for which $P(Z_{i1}) > P(Z_{j1}) > P(Z_{i2}) > P(Z_{j2})$. Hence, the order of the item steps is

$$Z_{i1}, Z_{j1}, Z_{i2}, Z_{j2}. \tag{2}$$

Table 1 Probabilities of item scores X_i and X_j, with $m + 1 = 3$ ordered answer categories

	X_j				
X_i	0	1	2	$P(X_i = x)$	$P(X_j \geq x)$
0	0.08	0.16	0.00	0.24	1.00
1	0.04	0.04	0.24	0.32	0.76
2	0.36	0.08	0.00	0.44	0.44
$P(X_i = x)$	0.48	0.028	0.24		
$P(X_j \geq x)$	1.00	0.52	0.24		

For notational convenience the subscripts jx in the item steps may be replaced by subscripts $(1), (2), \ldots, (2m)$ indicating the order of the item steps in an item pair. In this notation, Eq. (2) equals $Z_{(1)}, Z_{(2)}, Z_{(3)}, Z_{(4)}$. For each item pair, item-score pattern (x, y) corresponds a specific realization of the ordered item steps. For example, for Eq. (2), item-score pattern $(0,2)$ corresponds to $Z_{i1} = 0, Z_{j1} = 1, Z_{i2} = 0, Z_{j2} = 1$. In a Guttman scale, the ordered item steps are strictly nonincreasing: Once a more popular item step is failed, a less popular item step cannot be passed. For example, in a Guttman scale, admissible values for Eq. (2) are 0,0,0,0; 1,0,0,0; 1,1,0,0; 1,1,1,0; and 1,1,1,1, which correspond to item-score patterns $(0,0)$, $(1,0)$, $(1,1)$, $(2,1)$, and $(2,2)$, respectively. A Guttman error occurs if a less popular item step is passed while a more popular item step is failed. For Eq. (2), realizations 0,1,0,0 ; 0,1,0,1 ; 1,1,0,1; and 1,0,1,0 which correspond to item-score patterns $(0, 1)$, $(0, 2)$, $(1, 2)$, and $(2, 0)$, respectively—are Guttman errors.

The weight of a Guttman error, denoted w_{ij}^{xy}, indicates the degree of deviation from the perfect Guttman scale (Molenaar 1991). Let $z_{(h)}^{xy}$ denote the realization of the hth $(1 \leq h \leq 2m)$ item step corresponding to the item-score pattern (x, y). The weight is computed as

$$w_{ij}^{xy} = \sum_{h=2}^{2m} \left\{ z_{(h)}^{xy} \left[\sum_{g=1}^{h-1} \left(1 - z_{(g)}^{xy} \right) \right] \right\} \tag{3}$$

(see, e.g., Kuijpers et al. 2013). Note that Eq. (3) counts the number of times a more difficult item step was passed, while an easier item step was failed. For admissible item-score patterns, the corresponding weights are zero, whereas for Guttman errors, the weights are positive. For example, assuming the order of the item steps in Eq. (2) is correct, for item-score pattern $(0, 2)$, $z_{(1)}^{02} = 0$, $z_{(2)}^{02} = 1$, $z_{(3)}^{02} = 0$, and $z_{(4)}^{02} = 1$. Hence, following Eq. (3), $w_{ij}^{02} = 1 \times [1] + 0 \times [1 + 0] + 1 \times [1 + 0 + 1] = 3$. Also note that for dichotomous items, the only item-score pattern that constitutes a Guttman error (i.e., either $(0, 1)$ or $(1,0)$) receives a weight 1 by definition. Hence, for dichotomous items weighting the Guttman errors has no effect.

In samples, weights w_{ij}^{xy} are estimated from the order of the item steps in the sample with Eq. (1) and denoted $\widehat{w}_{ij}^{xy}$. Typically, w_{ij}^{xy} and $\widehat{w}_{ij}^{xy}$ are the same, but when the sample is small or when the popularities of two item steps are close, w_{ij}^{xy} and $\widehat{w}_{ij}^{xy}$ may differ (for more information on this topic, we refer to Kuijpers et al. 2016).

2.2 Applications

Weighted Guttman errors are used to compute scalability coefficients in Mokken scale analysis. Mokken (1971) discussed scalability coefficients for dichotomous items, Molenaar (1983, 1991, 1997) generalized the scalability coefficients to polytomous items, and Snijders (2001a, also, see Crisan et al. 2016) generalized the scalability coefficients to two-level data. The scalability coefficients are

implemented in several software packages, including the stand-alone package MSP (Molenaar and Sijtsma 2000) and the R package mokken (Van der Ark 2012). Mokken's (1971) item-pair scalability coefficient H_{ij} can be written as a function of the Guttman weights and the univariate and bivariate item probabilities:

$$H_{ij} = 1 - \frac{\sum_x \sum_y w_{ij}^{xy} P\left(X_i = x, X_j = y\right)}{\sum_x \sum_y w_{ij}^{xy} P\left(X_i = x\right) P\left(X_j = y\right)}. \tag{4}$$

Note that if unweighted Guttman errors were used, weights w_{ij}^{xy} only take on the values 0 and 1. By using weighted Guttman errors, H_{ij} equals the ratio of the inter-item correlation and the maximum inter-item correlation given the marginal distributions of the two items (Molenaar 1991).

In a sample of size N, the item-pair scalability coefficient is estimated by replacing the weights in Eq. (4) by the estimated weights and replacing the probabilities by the sample proportions, that is,

$$\widehat{H}_{ij} = 1 - \frac{\sum_x \sum_y \widehat{w}_{ij}^{xy} \widehat{P}\left(X_i = x, X_j = y\right)}{\sum_x \sum_y \widehat{w}_{ij}^{xy} \widehat{P}\left(X_i = x\right) \widehat{P}\left(X_j = y\right)} = 1 - \frac{F_{ij}}{E_{ij}}. \tag{5}$$

$F_{ij} = N \sum_x \sum_y \widehat{w}_{ij}^{xy} \widehat{P}\left(X_i = x, X_j = y\right)$ expresses the weighted sum of observed Guttman errors, and $E_{ij} = N \sum_x \sum_y \widehat{w}_{ij}^{xy} \widehat{P}\left(X_i = x\right) \widehat{P}\left(X_j = y\right)$ the weighted sum of Guttman errors expected when the two items are marginally independent.

Weighted Guttman errors are also used as an index to detect outliers and as a person-fit statistic. In these applications, the total of estimated Guttman weights within a response pattern is used. Let x_{ni} denote the observed score of person n on item i, and let y_{nj} denote the observed score of person n on item j. Using the notation of Zijlstra et al. (2007), index G_+ for respondent n equals

$$G_{n+} = \sum \sum_{i<j} \widehat{w}_{ij}^{x_{ni} y_{nj}}. \tag{6}$$

The function check.errors() in the R package mokken provides weighted Guttman errors for each observation.

3 Computational Problems

3.1 Problem 1: Ties

Estimating Guttman weights can be problematic if two estimated item steps have the same popularity. If the estimated item steps pertain to the same item, $\widehat{P}\left(X_j \geq x\right) = \widehat{P}\left(X_j \geq x + 1\right)$, it means that no one in the sample had score x on item j. The ordering of the estimated item steps is not affected because item steps have a fixed order within an item, and estimating Guttman errors is not problematic. However, if

Table 2 Cross-classification of item scores X_i and X_j, with $m + 1 = 3$ ordered answer categories, for $N = 15$ respondents

	X_j				
X_i	0	1	2	Total	$\widehat{P}(X_i \geq x)$
0	2	4	0	6	1.00
1	1	1	0	2	0.60
2	3	2	2	7	0.47
Total	6	7	2	15	
$\widehat{P}(X_j \geq y)$	1.00	0.60	0.13		

Table 3 Observed and expected frequencies, Guttman weights under two possible item-step orderings, and their mean, for each response pattern in Table 2

	Item-score vector								
	00	01	02	10	11	12	20	21	22
$N \times \widehat{P}(X_i = x, X_j = y)$	2	4	0	1	1	0	3	2	2
$N \times \widehat{P}(X_i = x)\widehat{P}(X_j = y)$	2.40	2.80	0.80	0.80	0.93	0.27	2.80	3.27	0.93
$\widehat{w}_{ij}^{xy}$ 1	0	1	3	0	0	1	1	0	0
$\widehat{w}_{ij}^{xy}$ 2	0	0	2	1	0	1	2	0	0
$\bar{\widehat{w}}_{ij}^{xy}$	0	0.5	2.5	0.5	0	1	1.5	0	0

For details, see text

the equally popular estimated item steps pertain to two different items, $\widehat{P}(X_i \geq x) = \widehat{P}(X_j \geq y)$, the item-step ordering cannot be determined. As an example, Table 2 shows the frequencies of the response patterns of $N = 15$ respondents, for two polytomous items with three response categories. For these data, $\widehat{P}(X_i \geq 1) = \widehat{P}(X_j \geq 1) = 0.6$, so the order of the item steps is either $Z_{i1}, Z_{j1}, Z_{i2}, Z_{j2}$ or $Z_{j1}, Z_{i1}, Z_{i2}, Z_{j2}$.

Currently, the software program `mokken` (Van der Ark 2012) adds a small random value to the estimated popularities to avoid equal item steps. There are two downsides to this approach. First, one item step is randomly assigned to be more popular than the other item step without theoretical justification. Second, analyzing the same data twice may result in different weights and, thus, different scalability coefficients.

Molenaar (1991) suggested computing the weights for all combinations of equivalent item-step orderings. For each item-score vector in Table 2, Table 3 shows the observed frequencies ($N \times \widehat{P}(X_i = x, X_j = y)$), the expected frequencies under marginal independence ($N \times \widehat{P}(X_i = x)\widehat{P}(X_j = y)$), the resulting weights given item-step ordering $Z_{i1}, Z_{j1}, Z_{i2}, Z_{j2}$ ($\widehat{w}_{ij}^{xy}$ 1), the resulting weights given item-step ordering $Z_{j1}, Z_{i1}, Z_{i2}, Z_{j2}$ ($\widehat{w}_{ij}^{xy}$ 2), and the average of the two weights. For both item-step orderings, the weighted sum of Guttman errors results in $F_{ij} = 7$ and $E_{ij} = 8.27$ (yielding $\widehat{H}_{ij} \approx 0.15$). Therefore, for scalability coefficients, the item-step order does not affect the outcome (Molenaar 1991). However, for individual-level statistics, such as the person-fit index G_+ (Eq. (6)), the item-step order matters. For example, a person with item-score vector (0,2) has value $G_+ = 3$ for the first item-step ordering and $G_+ = 2$ for the second item-step ordering. Because both item-step

orderings are equally likely in the population, the average weight (Table 3, last row) is considered more appropriate as opposed to randomly favouring one ordering over the other, and results in a value of $G_+ = 2.5$.

3.2 Problem 2: Estimating the Item Ordering for Two-Level Test Data

In Mokken scale analysis for two-level data, X_{srj} denotes the response of subject s ($s = 1, \ldots, S$) to item j ($j = 1, \ldots J$) scored by rater ($r = 1, \ldots, R_s$). As with one-level data, item step $Z_{jx} = 1$ if $X \geq x$, and $Z_{jx} = 0$, otherwise. The problem is that the order of the item steps, and hence the value of the Guttman weights, depends on the estimation method for $P(X_j \geq x)$. $P(X_j \geq x)$ can be estimated in two ways (Snijders 2001a), possibly yielding different estimates. Let Z_{srjx}, with realization z_{srjx}, be a binary variable that takes on the value one if $X_{srj} \geq x$, and zero otherwise. First, $P(X_j \geq x)$ can be estimated by averaging the relative frequencies for all subjects, that is,

$$\widehat{P}(X_j \geq x) = \frac{1}{S} \sum_{s=1}^{S} \frac{1}{R_s} \sum_{r=1}^{R_s} z_{srjx}, \tag{7}$$

and, second, $P(X_j \geq 1)$ can be estimated by averaging the absolute frequencies for all subjects, that is,

$$\widehat{P}(X_j \geq x) = \frac{1}{\sum_{s=1}^{S} R_s} \sum_{s=1}^{S} \sum_{r=1}^{R_s} z_{srjx}. \tag{8}$$

The example in Table 4 (last two rows) shows that the estimation methods do not only result in different estimates but also in different ordering of item steps. When averaging the relative frequencies of all subjects in Eq. (7), the ordering of the item steps is $Z_{i1}, Z_{j1}, Z_{i2}, Z_{j2}$, and when averaging the absolute frequencies of

Table 4 Values of $\sum_{r=1}^{R_s} z_{srjx}$ for $J = 2$ items, each having three ordered response categories, $S = 3$ subjects who are rated by $R_s = 10, 3, 10$ raters, respectively, and the values of $\widehat{P}(X_j \geq x)$ using Eqs. (7) and (8), respectively

	X_i			X_j			
s	$x \geq 0$	$x \geq 1$	$x \geq 2$	$x \geq 0$	$x \geq 1$	$x \geq 2$	R_s
1	10	4	2	10	3	3	10
2	3	2	2	3	3	2	3
3	10	4	2	10	3	3	10
Equation (7)	1.00	0.49	0.36	1.00	0.43	0.26	
Equation (8)	1.00	0.53	0.42	1.00	0.39	0.35	

all subjects in Eq. (8), the ordering of the item steps is Z_{i1}, Z_{i2}, Z_{j1}, Z_{j2}. Snijders (2001a) argued that averaging the relative frequencies in Eq. (7) is the preferred method, as averaging the absolute frequencies is biased under certain conditions.

4 Discussion

Two problems with the weighted Guttman errors have been addressed and described in this chapter. The solution to the problem of ties can be incorporated in the software easily. The software program MSP prints a warning when ties are present. As far as we know, the DOS program TWOMOK (Snijders 2001b) is the only software for two-level scalability coefficients. Because it pertains to dichotomous items only, weighted Guttman errors are not an issue. A new R function to compute weighted Guttman errors for dichotomous and polytomous two-level item scores is called `MLweight()`. The function is described in Koopman (2016). The main goal of `MLweight()` is to allow the computation of Mokken's scalability coefficients for two-level data in the function `MLcoefH()`. Both functions have been implemented in the R package `mokken`.

References

D.R. Crisan, J.E. Van de Pol, L.A. Van der Ark, Scalability coefficients for two-level polytomous item scores: an introduction and an application, in *Quantitative Psychology Research: The 80th Annual Meeting of the Psychometric Society, Beijing, China, 2015*, ed. by L.A. van der Ark, D.M. Bolt, W.-C. Wang, J.A. Douglas, M. Wiberg (Springer, New York, 2016), pp. 139–153. doi: 10.1007/978-3-319-38759-8_11

L. Guttman, The basis for scalogram analysis, in *Measurement and Prediction*, ed. by S.A. Stouffer, L. Guttman, E.A. Suchman, P.F. Lazarsfeld, S.A. Star, J.A. Clausen (Princeton University Press, Princeton, 1950), pp. 60–90

G. Karabatsos, Comparing the aberrant response detection performance of thirty-six person-fit statistics. Appl. Meas. Educ. **16**, 277–298 (2003). doi:10.1207/S15324818AME1604_2

L. Koopman, Standard errors of scalability coefficients in two-level Mokken scale analysis. Unpublished master's thesis, Research Institute of Child Development and Education, University of Amsterdam, The Netherlands, 2016

R.E. Kuijpers, L.A. Van der Ark, M.A. Croon, Standard errors and confidence intervals for scalability coefficients in Mokken scale analysis using marginal models. Sociol. Methodol. **43**, 42–69 (2013). doi:10.1177/0081175013481958

R.E. Kuijpers, L.A. Van der Ark, M.A. Croon, K. Sijtsma, Bias in estimates and standard errors of Mokken's scalability coefficients. Appl. Psychol. Meas. **40**, 331–345 (2016). doi:10.1177/0146621616638500

R.R. Meijer, The number of Guttman errors as a simple and powerful person-fit statistic. Appl. Psychol. Meas. **18**, 311–314 (1994). doi:10.1177/014662169401800402

R.J. Mokken, *A Theory and Procedure of Scale Analysis: With Applications in Political Research* (De Gruyter, Berlin, 1971)

I.W. Molenaar, *Item Steps. Heymans Bulletin HB-83-630-EX* (University of Groningen, Groningen, 1983)

I.W. Molenaar, A weighted Loevinger H-coefficient extending Mokken scaling to multicategory items. Kwantitatieve Methoden **12**(37), 97–117 (1991)

I.W. Molenaar, Nonparametric models for polytomous responses, in *Handbook of Modern Item Response Theory*, ed. by W.J. van der Linden, R.K. Hambleton (Springer, New York, 1997), pp. 369–380. doi: 10.1007/978-1-4757-2691-6_21

I.W. Molenaar, K. Sijtsma, *User's Manual MSP5 for Windows* (iec ProGAMMA, Groningen, 2000)

K. Sijtsma, I.W. Molenaar, *Introduction to Nonparametric Item Response Theory* (Sage, Thousand Oaks, CA, 2002)

K. Sijtsma, L.A. Van der Ark, A tutorial on how to do a Mokken scale analysis on your test and questionnaire data. Br. J. Math. Stat. Psychol. **70**(1), 137–158 (2017). doi:10.1111/bmsp.12078

T.A.B. Snijders, Two-level nonparametric scaling for dichotomous data, in *Essays on Item Response Theory*, ed. by A. Boomsma, M.A.J. van Duijn, T.A.B. Snijders (Springer, New York, 2001a), (pp. 319–338). doi: 10.1007/978-1-4613-0169-1_17

T.A.B. Snijders, TWOMOK [computer software]. Retrieved from https://www.stats.ox.ac.uk/~snijders/ (2001b)

L.A. Van der Ark, New developments in Mokken scale analysis in R. J. Stat. Softw. **48**(5), 1–27 (2012). doi:10.18637/jss.v048.i05

W.P. Zijlstra, L.A. Van der Ark, K. Sijtsma, Outlier detection in test and questionnaire data. Multivar. Behav. Res. **42**, 531–555 (2007). doi:10.1080/00273170701384340