Nonparametric item response theory for multilevel test data

Koopman, L.

Publication date
2022

Chapter 2

Weighted Guttman Errors – Handling Ties and Two-Level Data

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 if two items have the same estimated popularity—and computing weighted Guttman errors if 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.

Chapter 2 is published as: Koopman, L., Zijlstra, B. J. H., & Van der Ark, L. A. (2017). Weighted Guttman errors: Handling ties and two-level data. In L. A. van der Ark, M. Wiberg, S. A. Culpepper, J. A. Douglas, & W.-C. Wang (Eds.), *Quantitative psychology: The 81st annual meeting of the Psychometric Society, Asheville, North Carolina, 2016* (pp. 183–190). Springer. doi: 10.1007/978-3-319-56294-0_17

2.1 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 & Molenaar, 2002; also see Sijtsma & 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 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, Zijlstra, & Van der Ark, 2020), we encountered two problems in estimating the weights of Guttman errors: Estimated weights depend on the value of a random seed if 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.

2.2 Weighted Guttman Errors

2.2.1 Theory

Let a test consist of I items with $m+1$ ordered response categories indexed by x ($x = 0, 1, \ldots, m$). Let X_i denote the item score of item i. Each item score consists of m item steps (Molenaar, 1983), binary variables denoted Z_{ix} ($i = 1, \ldots, I; x = 1, \ldots, m$). $Z_{ix} = 1$ if $X_i \geq x$ (the item step was passed) and $Z_{ix} = 0$ if $X_i < x$ (the item step was failed). It follows that $X_{i,x-1} \geq Z_{ix}$ and $X_i = \sum_x Z_{ix}$. For example, if $X_i = 1$ and $m = 3$, then $Z_{i1} = 1$, $Z_{i2} = 0$,, and $Z_{i3} = 0$. Let the popularity of item step Z_{ix} be the probability of having a score of at least x on item i: $P(Z_{ix}) \equiv P(X_i \geq x)$. Note that by definition, $P(X_i \geq 0) = 1$. Let z_{nix} denote the realization of Z_{ix} for person n, then, in a sample of

N respondents, $P(X_i \geq x)$ is estimated by

$$\widehat{P}(X_i \geq x) = \frac{1}{N} \sum_{n=1}^{N} z_{nix}. \tag{2.1}$$

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 2.1 shows $I = 2$ 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

$$Z_{i1}, Z_{j1}, Z_{i2}, Z_{j2} \tag{2.2}$$

Table 2.1: Probabilities of item scores X_i and X_j, with $m + 1 = 3$ ordered answer categories.

	X_j				
X_i	0	1	2	$P(X_i = x)$	$P(X_i \geq x)$
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
$P(X_j = x)$	0.48	0.28	0.24		
$P(X_j \geq x)$	1.00	0.52	0.24		

For notational convenience the subscript jx in the item steps may be replaced by the subscripts (1), (2), …, ($2m$) indicating the order of the item steps in an item pair. In this notation, Equation 2.2 equals $Z_{(1)}$, $Z_{(2)}$, $Z_{(3)}$, $Z_{(4)}$. For each item pair, item-score pattern (x, y) corresponds to a specific realization of the ordered item steps. For example, for Equation 2.2, 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 Equation 2.2 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 Equation 2.2, 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

$$w_{ij}^{xy} = \sum_{h=2}^{2m} \left\{ z_h^{xy} \times \left[\sum_{g=1}^{h-1} \left(1 - z_g^{xy}\right) \right] \right\} \tag{2.3}$$

(see, e.g., Kuijpers et al., 2013). Note that Equation 2.3 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 Equation 2.2 is correct, for item-score pattern $(0, 2)$, $z^{02}_{(1)} = 0$, $z^{02}_{(2)} = 1$, $z^{02}_{(3)} = 0$, and $z^{02}_{(4)} = 1$. Hence, following Equation 2.3, $w^{xy}_{ij} = 1 \times [1] + 0 \times [1 + 0] + 1 \times [1 + 0 + 1] = 3$. 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^{xy}_{ij} are estimated from the order of the item steps in the sample with Equation 2.1 and denoted $\widehat{w}^{xy}_{ij}$. Typically, w^{xy}_{ij} and $\widehat{w}^{xy}_{ij}$ are the same, but if the sample is small or if the popularities of two item steps are close, w^{xy}_{ij} and $\widehat{w}^{xy}_{ij}$ may differ (for more information on this topic, we refer to Kuijpers et al., 2016).

2.2.2 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; see also Crişan 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 & Sijtsma, 2000) and the R package `mokken` (Van der Ark, 2007, 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:

$$H_{ij} = 1 - \frac{\sum_x \sum_y w^{xy}_{ij} P(X_i = x, X_j = y)}{\sum_x \sum_y w^{xy}_{ij} P(X_i = x) P(X_j = y)}. \tag{2.4}$$

Note that if unweighted Guttman errors were used, weights w^{xy}_{ij} 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 Equation 2.4 by the estimated weights and replacing the probabilities by sample proportions, that is,

$$\widehat{H}_{ij} = 1 - \frac{\sum_x \sum_y \widehat{w}^{xy}_{ij} \widehat{P}(X_i = x, X_j = y)}{\sum_x \sum_y \widehat{w}^{xy}_{ij} \widehat{P}(X_i = x) \widehat{P}(X_j = y)} = 1 - \frac{F_{ij}}{E_{ij}}. \tag{2.5}$$

$F_{ij} = N \sum_x \sum_y \widehat{w}^{xy}_{ij} \widehat{P}(X_i = x, X_j = y)$ expresses the weighted sum of observed Guttman errors, and $E_{ij} = N \sum_x \sum_y \widehat{w}^{xy}_{ij} \widehat{P}(X_i = x) \widehat{P}(X_j = y)$ the weighted sum of

expected Guttman errors under marginal independence of the two items.

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

$$G_{n+} = \sum_{i<j} \sum \widehat{w}_{ij}^{x_{ni}y_{nj}}. \tag{2.6}$$

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

2.3 Computational Problems

2.3.1 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, $\widehat{P}(X_i \geq x) = \widehat{P}(X_i \geq x+1)$, it means that no one in the sample had score x on item i. 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 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 2.2 shows the frequencies of the response patterns for $N = 15$ 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 $Z_{j1}, Z_{i1}, Z_{i2}, Z_{j2}$.

Table 2.2: 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	
$\widehat{P}(X_j \geq x)$	1.00	0.60	0.13		

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 2.2, Table 2.3 shows the observed frequencies ($N \times \widehat{P}(X_i = x, X_j = y)$), the expected frequencies ($N \times \widehat{P}(X_i = x)\widehat{P}(X_j = y)$), 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 ($\overline{w}_{ij}^{xy}$). 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_+ (Equation 2.6), 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 2.3, last row) is considered more appropriate as apposed to randomly favouring one ordering over the other, and results in a value of $G_+ = 2.5$.

Table 2.3: Observed and expected frequencies, Guttman weights under two possible item-step orderings, and their mean, for each response pattern in Table 2.2.

	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
$\overline{w}_{ij}^{xy}$	0	0.5	2.5	0.5	0	1	1.5	0	0

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

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

be estimated by averaging the relative frequencies for all subjects, that is,

$$\widehat{P}(X_i \geq x) = \frac{1}{S} \sum_{s=1}^{S} \frac{1}{R_s} \sum_{r=1}^{R_s} z_{srix}, \tag{2.7}$$

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

$$\widehat{P}(X_i \geq x) = \frac{1}{\sum_{s=1}^{S} R_s} \sum_{s=1}^{S} \sum_{r=1}^{R_s} z_{srix}. \tag{2.8}$$

The example in Table 2.4 (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 Equation 2.7, the ordering of the item steps is $Z_{i1}, Z_{j1}, Z_{i2}, Z_{j2}$, and when averaging the absolute frequencies of all subjects in Equation 2.8, the ordering of the item steps is $Z_{i1}, Z_{i2}, Z_{j1}, Z_{j2}$. Snijders (2001a) argued that averaging the relative frequencies in Equation 2.7 is the preferred method, as averaging the absolute frequencies is biased under certain conditions.

Table 2.4: Observed and expected frequencies, Guttman weights under two possible item-step orderings, and their mean, for each response pattern in Table 2.2.

	Item-score vector			Item-score vector			
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 2.7	1.00	0.49	0.36	1.00	0.43	0.26	
Equation 2.8	1.00	0.53	0.42	1.00	0.39	0.35	

2.4 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 main goal of `MLweight()` is to allow the computation of two-level scalability coefficients for two-level data in the function `MLcoefH()`. Both functions have been implemented in the R package `mokken`.