Avaliação Psicológica
ISSN 2175-3431 (versão on-line)

Research report

How to score respondents? A Monte Carlo study comparing three different procedures

Vithor Rosa Franco[1]
Universidade São Francisco, Campinas, SP, Brasil
Marie Wiberg
Umeå Universitet, Umeå, AC, Sweden

ABSTRACT

This study aimed to empirically compare the effectiveness of Likert, Thurstonian, and Expected a Posteriori (EAP) scoring methods. A computational simulation of the two-parameter logistic model was employed under various conditions, including different sample sizes, number of items, scale levels, item extremeness, and varying discriminations. Effectiveness was assessed through correlation with the true score, root mean squared errors, bias, and the accurate recovery of effect sizes. The results indicated that Likert scores exhibit greater bias than EAP and Thurstonian scores for extreme scores, however, they strongly correlate with both the true score and EAP scores. Likert scores were slightly more effective in recovering mean differences between two groups, correlation estimates, and regression parameters. Overall, Likert scores should be avoided when ordering or thresholding individuals at the extremes of scales is the primary objective. However, they are preferable in situations where Thurstonian and EAP scores may fail to converge. The study also recommends that future research explore conditions involving more complex data-generating processes.

Keywords: Factor scores; Psychometric theory; Monte Carlo simulation.

RESUMO – Como pontuar respondentes? Um estudo de Monte Carlo comparando três procedimentos diferentes

O objetivo deste artigo foi comparar empiricamente a eficácia dos métodos de escores de Likert, Thurstoniano e Esperado a Posteriori (EAP). Foi realizada uma simulação computacional de um modelo logístico de dois parâmetros em diversas condições (diferentes tamanhos amostrais, número de itens, níveis de escala, itens extremos e discriminações variadas). A eficácia foi medida pela correlação com o escore verdadeiro, pelo erro quadrático médio, viés e pela recuperação correta dos tamanhos de efeito. Os resultados mostraram que os escores de Likert são mais enviesados do que os escores EAP e Thurstoniano para escores extremos, embora apresentem correlação forte tanto com o escore verdadeiro quanto com os escores EAP. Os escores de Likert também se mostraram ligeiramente mais eficientes na recuperação de diferenças médias entre dois grupos, estimativas de correlação e parâmetros de regressão. De modo geral, os escores de Likert devem ser evitados quando a ordenação ou a definição de limites dos indivíduos nas extremidades das escalas for o principal interesse. No entanto, eles devem ser preferidos sempre que os escores Thurstoniano e EAP possam não convergir. Recomendamos também que estudos futuros testem condições com processos geradores de dados mais complexos.

Palavras-chave: Escores fatoriais; Teoria Psicométrica; simulação de Monte Carlo.

RESUMEN – ¿Cómo calificar a los respondientes? Un estudio de Monte Carlo que compara tres procedimientos diferentes

El objetivo de este artículo fue comparar empíricamente la efectividad de los métodos de medición Likert, Thurstoniano y Esperado a Posteriori (EAP). Se realizó una simulación computacional de un modelo logístico de dos parámetros bajo diversas condiciones (diferentes tamaños de muestra, número de ítems, niveles de escala, ítems extremos y discriminaciones variadas). La efectividad se midió mediante la correlación con la puntuación verdadera, el error cuadrático medio, el sesgo y la correcta recuperación de los tamaños de efecto. Los resultados mostraron que las puntuaciones de Likert son más sesgadas que las puntuaciones EAP y Thurstonianas para puntuaciones extremas, aunque presentan una fuerte correlación tanto con la puntuación verdadera como con las puntuaciones EAP. Las puntuaciones de Likert también se mostraron ligeramente más eficientes en la recuperación de diferencias medias entre dos grupos, estimaciones de correlación y parámetros de regresión. En general, las puntuaciones de Likert deben evitarse cuando la ordenación o el establecimiento de límites de los individuos en los extremos de las escalas sea el principal interés. Sin embargo, deben preferirse siempre que las puntuaciones Thurstonianas y EAP puedan no converger. También recomendamos que estudios futuros prueben condiciones con procesos generadores de datos más complejos.

Palabras clave: Puntuaciones factoriales; Teoría psicométrica; Simulación de Monte Carlo.

Sijtsma et al. (2024a) provide a thorough discussion to sustain that the sum score on a psychological test is, and should continue to be, a central tool in psychometric practice. This paper, and its main arguments, were critiqued by McNeish (2024) and Mislevy (2024), which were also followed by a rejoinder

[1] Endereço para correspondência: Universidade São Francisco, Câmpus de Campinas. Rua Waldemar César da Silveira, 105, Jardim Cura D'Ars, 13045-510, Campinas, SP. Tel.: (19) 3779-3396. *E-mail:* vithorfranco@gmail.com

(Sijtsma et al.; 2024b). In spite of the possible conclusions of this particular interaction, the discussion about the usefulness and overall psychometric robustness of sum scores (also known as Likert scoring, test scores, and others) is not new (e.g., DiStefano et al., 2009; Gorsuch, 1983; Hair et al., 2006). One important aspect to consider is that Likert scoring, factor analysis, Item Response Theory (IRT) and other psychometric models are all implementations of the latent variable theory (LVT; Franco et al., 2022; McDonald 1999). As a consequence, this means that all these models are, in their core, the same, providing only different means to estimate true scores (Wiberg et al., 2019).

Therefore, while Likert scoring provides an efficient monotonic estimate of the true scores (in ideal scenarios; Coenders et al., 1995), methods derived from factor analysis and IRT take the structure of the data generating process into account. In principle, these "refined" methods should then provide better estimates of theta, even when the data generating process reflects ideal conditions for Likert scoring to be an efficient estimate of the true scores. The present study aims at empirically comparing the effectiveness, defined in terms of the correlation with the true score, error and bias measures, of the estimates of Likert, Thurstonian and EAP scores. To achieve this, computational simulation of two-parameter logistic model (2PLM) in a number of different conditions were used to estimate the different scores.

This study is relevant as Likert scores are far less computationally extensive than other methods, but not useful if its estimates are much biased or error prone. Additionally, a simulation study allows to evaluate properties that are not readily understandable from non-formal treatments of the relations between different forms of estimating the true scores. The rest of the paper is structured as follows. First the general psychometrical framework, which include the different scoring methods, and the methods themselves (Likert, Thurstonian, and EAP scores) are briefly described, followed by a method section which describes the setup of the simulation study. Next, the results from the simulations are presented and the paper ends with a discussion with some concluding remarks.

General Psychometrical Framework

There are three major theories in psychometrics aiming at identifying and measuring response patterns in tests (Franco et al., 2022; McDonald, 1999): Classical Test Theory (CTT); Common Factor Theory (CFT); and Item Response Theory (IRT). Despite usually defined as different theories (e.g., Borsboom, 2005), all of them can be understood as different applications of the same general Latent Variable Theory (LVT). Latent variables are not directly observed – sometimes not even observable –, and inferred from other variables that are directly measured or observed. This inferential step from observed data, after applying different functional forms, is what distinguishes them.

Many textbooks and papers discuss different procedures for estimating the magnitude of latent variables (e.g., van der Linden & Hambleton, 2013). Differences in the results after applying these different methods rely, mostly, in the assumptions relating to the formulas and estimation procedures used. For instance, Likert scores stem from CTT (Crocker & Algina, 1986), which aims to describe psychological outcomes given the general characteristics of a test. CTT relies on the assumptions that each respondent gets an observed score which is composed by the true score on the test and an error term:

$$X_i = T_i + \varepsilon_i \quad (1)$$

where X_i stands for the observed test score of respondent i, T_i for the true, latent, score of respondent i, and ε_i for the random error of the measurement of respondent i's observed score.

For the CFT, on the other hand, items on a test are assumed to be sampled from a population of items that relates to the true score (Borsboom, 2005). Each item then represents, to a different magnitude, the true score. Therefore, the procedures in this theory are mainly aimed at discovering this magnitude, with the general representation of

$$X_{ij} = \lambda_j T_i + \varepsilon_{ij} \quad (2)$$

where X_{ij} stands for the response of respondent i to the j^{th} item, λ_j for the factor loadings of item j, T_i for the true latent score of respondent i, and ε_{ij} for the random error of the measurement of respondent i's observed score in the j^{th} item. An important difference between CFT and CTT is that the factor loadings make CFT a more testable model (Bamber & van Santen, 2000), meaning that factor loadings are usually free parameters. On the other hand, scoring procedures stemming from CTT are usually parameter free and, therefore, not testable.

For IRT, it is accepted that the measurement level of the observed score is ordinal or even nominal (McDonald, 1999). Therefore, models based on IRT, usually called IRMs, assume that mixture effects of items and respondents' latent characteristics generates the probabilities of the response categories of the items. A general representation of this theory is

$$X_{ij} = g(T_{ij}) \ , \ T_{ij} = f(\theta_i - b_j) \quad (3)$$

where X_{ij} stands for the response of respondent i to the j^{th} item, the $g()$ function is some probability mass functions – usually binomial or categorical distributions (van der Linden & Hambleton, 2013) – , T_{ij} is the probability of respondent i giving a certain response to the j^{th} item, the $f()$ function is some link function – usually variations of

the logistic function or the cumulative normal function; (van der Linden & Hambleton, 2013) – , θ_i is the true latent score of respondent *i*, and b_j is the item latent score, usually called the difficulty parameter, of the j^{th} item. The error term ε_{ij} is suppressed in this representation given that it is a natural consequence of the stochastic nature of the *g*() function.

These differences in assumptions about how the true score is related to the observed scores causes differences on how to estimate the true score depending on what theory is held as true (Borsboom, 2005; Franco et al., 2022; McDonald, 1999). If Equation 1 is held as true, the true score is typically estimated by Likert scoring. The true score in Equation 2 is usually estimated by some combination of algebraic and optimization methods applied to models of confirmatory or exploratory factor analysis. The true score in Equation 3 is usually estimated by conditioning on the response patterns on estimated item parameters of some IRM and by assuming a specific distribution for the true scores.

Classical Test Theory and Likert Scores

Following the representation of CTT in Equation 1, a common operationalization is achieved, for instance, whenever deviations from the true score are assumed to follow a normal distribution with 0 mean (Borsboom, 2005). This is a common assumption for many statistical models (Tabachnick & Fidell, 2007) and simply means that different measures of the same thing can result in different values, due to random noise. This means, considering a CTT model, that observed differences in item scores are deemed to be exclusively due to noise in the responding procedure. It is also a consequence from this assumption, and the formal representation of the theory, that observed scores are expected to be correlated. High average correlation between items in a test means lesser magnitude of random errors, which is measured by reliability.

Reliability is the overall consistency of a measure. One of the most well-known techniques for estimating overall consistency is Cronbach's alpha (Dunn et al., 2014). This method follows from the assumption that items equally represent the same true score and, therefore, higher average correlation between items means lower error variance. Following the instantiation of a highly reliable test, which thus serves as a test of the overall magnitude of error, respondents true scores can be estimated by Likert scores (Norman, 2010). Likert scoring can be calculated as

$$L_i = \sum_{j=1}^{k} x_{ij} \tag{4}$$

where L_i represents the Likert score for respondent *i*, *k* is the quantity of items in the test and X_{ij} is the observed score given by the i^{th} respondent to the j^{th} item. The calculation of Likert scores is sound from a CTT perspective, but problematic from a statistical perspective. Its assumptions of equally representing the true score and of being measured on an interval level of measurement have long been criticized by psychometricians, whom found several examples when not all items equally represent what is meant to be measured (Borsboom, 2005; Crocker & Algina, 1986). This started the practice of factor analysis (Thompson, 2004) and IRM (van der Linden & Hambleton, 2013).

Common Factor Theory and Thurstonian Factor Scores

Factor analysis (Thompson, 2004) is typically used to reduce the dimensionality in a dataset and describe the variability among the observed variables. Assuming that respondents' observable scores $\boldsymbol{x}$ are due to a linear combination of their latent trait and a specific skill to each item, one can simply recover the true score *f* using

$$\boldsymbol{x} = \boldsymbol{\Lambda} f \boldsymbol{\Lambda}^{'} + \boldsymbol{u}^{2} , \tag{5}$$

where

$$\boldsymbol{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}, \boldsymbol{\Lambda} = \begin{pmatrix} \lambda_1 \\ \vdots \\ \lambda_k \end{pmatrix}, \ f = \begin{pmatrix} f_1 \\ \vdots \\ f_n \end{pmatrix} \quad \boldsymbol{u} = \begin{pmatrix} u_{11} & \cdots & u_{1k} \\ \vdots & \ddots & \vdots \\ u_{n1} & \cdots & u_{nk} \end{pmatrix},$$

and $\boldsymbol{\Lambda}$ is the factor loadings for each k^{th} item, f_n is the factor score of the n^{th} respondent, and $\mathbf{u}$ is the unique property of each k^{th} item. Thurstone (1935) was the first to propose the use of this model, where $\boldsymbol{\Lambda}$ is a set of free parameters, to estimate the true scores of respondents.

For the proper estimation of Thurstonian scores, one must first choose an extraction (or factoring) procedure (Tabachnick & Fidell, 2007). The minimum residual procedure, which minimizes the off-diagonal residuals of the correlation matrix, is one of the most effective extraction procedures (Harman & Jones, 1966) and can, therefore, be used as a solver for Equation 5. After estimating factor loadings, the Thurstonian procedure can be used to obtain standardized true score estimates (Grice, 2001):

$$\mathbf{W} = \mathbf{R}_{jj}^{-1}\mathbf{S}_j \ , \ f_i = \mathbf{Z}_{ij}\mathbf{W} \tag{6}$$

where $\mathbf{W}$ is the matrix of factor score coefficients, $\boldsymbol{R}_{jj}$ is the matrix of items' correlations, $\boldsymbol{S}_j$ is the vector of structure coefficients and $\boldsymbol{Z}_{ij}$ is the matrix of standardized observed scores. However, the use of traditional factor analysis models in psychology has been criticized as the general model assumes observed variables to be measured on a interval level, when they are, at best, measured on an ordinal level (Wirth & Edwards, 2007). IRT has been proposed to overcome this limitation (Lord, 1980).

Item Response Theory and EAP scores

IRT (Lord, 1980) is an item centered theory which relies on the assumptions that each observed response is conditioned on a single respondent's ability θ, that the items have local independence (as in the CFT framework), and that responses can be modelled with an IRM. There are several IRMs, but the Rasch and the family of logistic models are the most famous models (van der Linden & Hambleton, 2013). If we assume 2PLM, then we have that if a respondent i has ability (or true score) θ_i, then the probability of answering an item correctly can be modelled by

$$Pr\,(X_{ij} = 1 \mid \theta_i, b_j, a_j) = \frac{exp[a_j(\theta_i - b_j)]}{1 + exp[a_j\,(\theta_i - b_j)]'} \qquad (7)$$

where b_j is the item difficulty and a_j is the item discrimination of item j. This model is different from the ones used for CTT or CFT given that, when using IRMs, latent score f is thought of a mixture of the θ_i and b_j parameters.

After obtaining item difficulty and item discrimination estimates, Bayesian statistical principles of parameter updating can be used to estimate the respondents true scores using the EAP estimator (Bock & Aitkin, 1981). Let $f(\theta)d\theta$ be a prior distribution, then we can compute the EAP score estimates by

$$\theta_{EAP} = E[p(\theta|X,\pi)]\ ,\ p(\theta|X,\pi) = \frac{p(X|\theta,\pi)\,f(\theta)d\theta}{\int p(X|\theta,\pi)\,f(\theta)d\theta'} \qquad (8)$$

which simply computes the expected value (θ_{EAP}) of the posterior probability, given a prior distribution of the true score, conditioned on respondents' observed scores, X, and a set of items parameters, π. For the 2PLM this means that $a_j, b_j \in \pi$. Usually, this prior distribution is defined as a normal distribution with mean 0 and standard deviation of 1 (Bock & Aitkin, 1981).

What are the Expected Differences Between These Methods?

It should be noted that, in some cases, these procedures will result in equivalent estimates for the true scores. For instance, Likert scoring is a valid procedure for estimating the magnitude of respondents' true scores when the data generating process is equivalent to a parallel factor analytical model (Coenders et al., 1995). In parallel factor analytical models, all items are considered to be equally influenced by the theorized latent variable (i.e., all factor loadings are equal). This is also equivalent to an IRT model where all the discriminations are equal, such as the Rasch model or the one-parameter logistic model (Alphen et al., 1994).

Therefore, despite of the fact that many researchers defend that respondents' estimates stemmed from FA (i.e., factor scores; Grice 2001) and IRT (e.g., expected a posteriori, EAP; Bock & Aitkin 1981) are superior to the ones stemmed from CTT (Likert scores; e.g., DiStefano et al., 2009), there are cases where they are quantitatively the same. This endless discussion can be attributed to several reasons, with at least two of them worth to discuss. First, the fact that, in CTT, reliability estimates were, and sometimes still are, improperly used for dimensionality analysis (Flake & Fried, 2020; Schmitt, 1996). Reliability estimates, such as Cronbach's alpha (Cronbach, 1951), are sometimes used to justify the inference of measurement by means of Likert scores to the particular trait. The second reason is that IRT and FA models have free parameters while CTT is a parameter-free model. Therefore, IRT and FA should be able to provide a better fit to the observed data.

Additionally, some previous studies have found that, overall, both items and respondents' estimates are highly correlated, despite from which model they were estimated from (Lawson, 1991). For instance, Fan (1998) showed that items and respondents' estimates derived from IRT and CTT are quite correlated, when assumptions regarding CTT hold. The author also showed that the degree of invariance of item statistics across samples, usually considered as the theoretical superiority IRT models, only differed for some particular cases of discriminations' estimates. These results are probably because of factor scores indeterminacy (Grice, 1991): an infinite number of sets of latent variables score can be created for the same analysis.

Some other authors (e.g., Franco et al., 2023; Ramsay & Wiberg, 2017; Ramsay et al., 2020; Wallmark et al., 2024) took this matter further and demonstrated that any strictly monotonic transformation can be used as a function for respondents scores' estimates. Therefore, for unidimensional scales, any estimates for respondents' true scores which are monotonically related to factor or EAP scores, such as the Likert scores, should provide good approximations of the true scores. However, they also showed that extreme scores are usually more biased for Likert scores than for other estimates of factor scores. Therefore, despite of the fact that previous research has shown that Likert scores are reliable estimates of the true score (Henson et al., 2007; Hinz et al., 2012), there are no study which compares the effectiveness of Likert, Thurstonian, and EAP scoring methods in recovering the true score using simulations as it is done here.

In sum, some psychometricians (e.g., Borsboom, 2005) consider CTT and CFT to be unsatisfying because these theories do not adequately represent the attribute to be measured in the model. This limitation should be surpassed by IRT, which proposes IRMs of the data-generating processes, explicitly stating the relation between observed and latent variables. As a consequence, it is expected that scoring methods esteemed from IRMs will be more effective than other scoring procedure that does not adequately represents the

attribute to be measured and the data to be described, such as Likert and Thurstonian scores. This should be true especially when discriminations are very different for each item.

Method

Study Design and Data Simulation

A Monte Carlo simulation study was designed to test the effectiveness of the three discussed scoring methods. To those interested in further examining the results or conducting further simulations, our codes are available from the corresponding author. The simulation used the 2PLM as the data-generating process. The θ parameter was drawn from an evenly spaced vector in the closed interval [-3,3], depending on the sample size, which was drawn from the set {60,100,500}. The difficulty parameter was distributed as a truncated normal distribution with three possible means, {-3,0,3}, two possible standard deviations, {.1,1}, with lower and upper bounds equal to the minimum and maximum of the means' set, respectively. Similarly, the discrimination parameter was also distributed with three possible means, {.5,1,2}, two possible standard deviations, {.1,1}, with lower and upper bounds equal to the minimum and maximum of the means set, respectively. The size of the drawn for both difficulty and discrimination parameters were dependent on the "test size" (i.e., number of items) parameter, which was drawn from the set {5,10,20}. The last parameter was scale size, fixed in dichotomous or 5-point scale response patterns.

These specifications resulted in 648 crossed conditions, which were iterated 100 times. These conditions were selected to replicate common practices in psychological literature (e.g., Colman et al., 1997; Marszalek et al., 2011; Morgado et al., 2018) and also to test the effects of weighting (i.e., varying discriminations) and extremeness (i.e., overall test difficulty, high or low) on the scoring procedures. Also, the θ parameter was drawn from a linearly spaced vector to avoid the necessity of using any equating procedure given distributional differences between difficulties and aptitudes (González & Wiberg, 2017). The 2PLM was used as the link function for getting the probabilities of respondents answering each item correctly. Finally, simulated dichotomous scores were generated using a binomial distribution and simulated polytomous scores were generated using a multinomial distribution, following recommendations from Agresti (2003). These steps make the 2PLM always true in the sample and, therefore, scores estimated using IRT procedures should be the most effective (Borsboom, 2005).

Data Analysis

Effectiveness of each method was assessed by their capacity to recover the true score – by means of Pearson's correlation between the observed scores and the known true score – and the use of root mean squared errors (RMSE) and bias estimates. It is expected that more effective methods will correlate more with the true score (i.e. values closer to 1) and also present lower RMSE and bias (i.e., values closer to 0). Using these measures of effectiveness, the levels of the seven conditions – sample size, number of items, number of scale levels, discrimination mean and variance and difficulty mean and variance – were compared so an optimal crossed condition could be used for the following tests. We define the optimal condition as the one that makes the best estimates of the true parameters, minimizing the costs (e.g., smaller sample size) for doing so.

Using only the optimal crossed condition, the methods were compared in their capacity to properly return estimates of mean differences in independent sample *t* tests, linear regression parameters and correlation between latent variables. Their effectiveness to recover standard mean differences between two independent latent variables was assessed by the average estimates of Cohen's *d* (Gignac & Szodorai, 2016). Their effectiveness to recover slopes and intercepts of linear regression parameters were tested by the average of estimated parameters of linear regressions, without using corrections for observed reliabilities (Bacon 2004). Finally, their effectiveness to recover correlation between two latent variables was examined through average Pearson's correlation on the estimated scores. All the analyses were conducted with R version 4.4.1 (R Core Team, 2024) and packages psych version 2.4.6.26 (Revelle, 2014) and ltm version 1.2-0 (Rizopoulos, 2007).

Results

Table 1 presents the results for each condition. In general, EAP (with a standardized normal as the prior for the true scores) and Likert scores always present higher correlations with the true score than the Thurstonian Scores, but Likert scores always present the highest RMSE. From the overall mean of effectiveness measures, EAP score has the best correlation performance, followed with a difference of .001 by the Likert score and with a difference of .042 by the Thurstonian score. In terms of RMSE, Thurstonian score always performs better, with overall difference of .142 to EAP and of .583 to Likert scores.

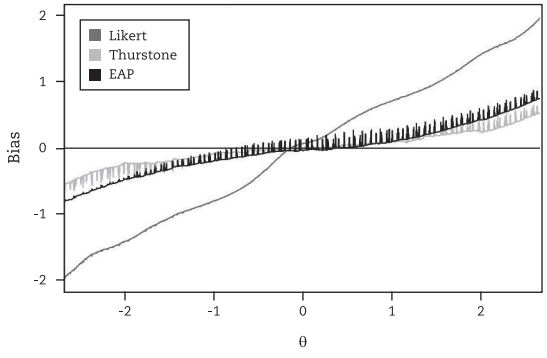
These results can be completed with the bias in Figure 1. It is evident that all the methods have similar performances when true scores are close to 0, which can also be numerically verified by the RMSE in Table 1. When the scores are below average, all the methods underestimate the scores. When the scores are above average, all the methods overestimate the scores. Nevertheless, the Likert procedure is clearly the most biased, especially for more extreme scores.

Table 1
Mean correlation with the true score and RMSE given the testing conditions for each scoring method

	Correlation with the True Score			RMSE		
	LS	TS	EAP	LS	TS	EAP
Average performance	.757	.716	**.758**	1.460	**.735**	.877
Sample Size						
60	**.751**	.705	.750	1.460	**.823**	.925
100	.751	.713	**.752**	1.462	**.785**	.912
500	.769	.731	**.772**	1.457	**.713**	.862
Number of items						
5	**.754**	.719	.743	1.462	**.753**	.788
10	.756	.721	**.758**	1.462	**.738**	.825
20	.762	.709	**.774**	1.459	**.720**	.922
Scale Levels						
2	.685	.650	**.690**	1.400	**.778**	1.026
5	**.829**	.783	.826	1.394	**.671**	.673
Difficulty						
M=-3; *SD*=.1	.594	.569	**.613**	.878	**.660**	.987
M=-3; *SD*=1	.784	.723	**.790**	.826	**.461**	.672
M=0; *SD*=.1	**.865**	.848	.845	.909	**.329**	.519
M=0; *SD*=1	**.909**	.868	.884	.968	**.234**	.380
M=3; *SD*=.1	.615	.570	**.630**	1.067	**.726**	.979
M=3; *SD*=1	.777	.715	**.788**	.951	**.474**	.738
Discrimination						
M=.5; *SD*=.1	**.713**	.657	.671	1.475	**.516**	.795
M=.5; *SD*=1	**.769**	.726	.766	1.433	**.705**	.805
M=1; *SD*=.1	**.768**	.729	.768	1.425	**.671**	.804
M=1; *SD*=1	.771	.731	**.774**	1.429	**.732**	.823
M=2; *SD*=.1	.752	.724	**.785**	1.444	**.898**	1.024
M=2; *SD*=1	.771	.731	**.786**	1.430	**.775**	.859

Note. LS=Likert score; TS=Thurstonian Score; EAP=Estimated a Posteriori. The most effective method in each condition is in bold. The optimal conditions are in italic

Figure 1
The bias for each procedure at estimating the true score



The results shown in Table 1 and Figure 1 suggest that the optimal condition is the one with 500 cases, 10 items when using a 5 points scale, with items' mean difficulty equal to 0, with standard deviation equal to 1, and with items' mean discriminations equal to 2, with standard deviation equal to 1. From Table 2, it is evident that when comparing the same measure of two independent samples, Likert and Thurstonian scores will give on average the same result, which is close to the real difference for the latent variables. The EAP score will usually underestimate them. The mid part of Table 2 shows the performance of the three methods when examining correlations of two observed variables. It is evident that all methods have similar performance, with Likert scores as the most effective, followed by EAP and then by the Thurstonian scores. The lower part of Table 2 displays when estimating regression's parameters, scoring methods will display diverging results. For the intercept, none of the methods gave a reliable estimate. For the slope, all the methods underestimated values above 1, with Likert scoring being the least biased one.

Table 2
Comparison of methods performance on standardized mean differences of two means, regression coefficients, and correlation estimates for the optimal crossed condition

Parameters		Likert		Thurstonian		EAP	
d		Mean differences between two independent samples					
.1		**.094**		.091		.090	
.5		**.472**		.462		.457	
1		**.959**		.950		.923	
1.5		**1.451**		1.443		1.392	
ρ		Correlation coefficients					
.100		.097		**.099**		.106	
.300		.284		.276		**.286**	
.500		**.472**		.460		.469	
.800		**.765**		.748		.754	
β_0 and β_1		Intercepts and slopes					
β_0	β_1	β_0	β_1	β_0	β1	β_0	β_1
0	1	.034	**.988**	-.0009	.986	.009	.983
.5	1.5	-.066	**1.172**	.006	.926	.067	.973
1	2	.067	**1.222**	.008	.876	.190	.958
1.5	2.5	.267	**1.216**	.009	.823	.238	.930

Note. *d* is the standard mean difference. ρ is the true correlation. β_0 and β_1 are the true intercept and slope, respectively

Discussion

Likert scoring is the most commonly used method for scoring respondents. Nevertheless, it is also the most criticized one, given that its assumptions are not deemed as reasonable for many psychometricians. Given that IRMs have a good reputation among psychometric theories, and that logistic models are the most famous IRMs, the present study used 2PLM as the basis for creating simulated data so true scores could be estimated. Overall, the results show that it makes almost no difference if the scoring method is based on CTT, CFT or IRT. The estimates given by the methods tested in the present study – Likert, Thurstonian, and EAP – have similar magnitudes of correlation with the simulated true scores. On the other hand, Thurstonian scores almost always presented the lowest RMSE values and EAP scores almost always presented the highest correlations with the true score. In general, Likert scores presented largest bias than the other methods.

On the other hand, if one was to consider the overall results, despite bias and solely from an efficiency in scoring respondents' point of view, Likert scoring should always be preferred over Thurstonian or EAP scoring, given two main reasons. First, computing a Likert score is far less intensive and intuitive then computing Thurstonian or EAP scores. Thurstonian and EAP scores usually will be estimated using some optimization method, such as maximum likelihood estimation by Expectation–Maximization algorithm (EM; Bock & Aitkin, 1981), which means sometimes the algorithm may fail to converge. The second reason is that, despite bias, Likert score performs at least as well as the other methods in several conditions, especially in the optimal crossed condition. Nevertheless, scoring respondents in dynamical test sets, such as computerized adaptive and multistage testing (Magis et al., 2017), can benefit by other kind of estimates, such as maximum Fisher information, that can be calculated only by applying IRMs.

An important practical consequence from the results follows for researchers who use Structural Equation Modeling (SEM; Byrne, 2016) to test regression models between latent variables. Most SEM's softwares, for example AMOS (Arbuckle, 2012) and lavaan, (Rosseel, 2012) use Thurstonian scores (or some variation) to estimate the latent variables. Therefore, SEMs will probably give biased slope estimates, such as the ones found in the present study. Following from the result that all scores are very biased when estimating linear regression parameters, but quite efficient for estimating correlation parameters, non-parametric SEM (Pearl, 2009) or Likert scores' associations modeled with Markovian Networks (Liu et al., 2017) may be used to reduce this bias.

From a theoretical point of view, the results found in Table 1 also partially refute the idea that IRT provides great practical advantages over CTT in recovering item and respondents estimates, as similarly found by Fan (1998). "Partially" because, as stated before, using a scoring method that is IRT based presented less bias. On the other hand, Likert scores will give at least as good estimates for regression parameters, mean differences and correlation as the other methods, even taking into account the amount of bias for scoring respondents with very high or very low true scores. It should then be remarked that the problem of measurement goes well beyond the simpler problem of making good estimates of a latent variable, given that experimentation, prior theorization and proper validation procedures can be more important than the use of some specific statistical model based on the general psychometrical framework (Sijtsma, 2012).

The results and conclusions from the present study are limited by the number and types of conditions tested. We only used a single process, based on the 2PLM, for the data generation process, as well as only unidimensional data generating processes. Different data generation processes, based on factor analysis, Item Factor Analysis, and other IRMs, may reveal different patterns that could change the overall results found in the present study. Also, there are a number of different IRT methods for estimating the respondents' true scores, such as maximum a posteriori estimation (MAP; Bock & Aitkin, 1981) and scoring methods that directly use Bayesian modeling (Fox, 2010), which may have different effectiveness when compared to conventional implementations of Likert, Thurstonian, and EAP scores.

We emphasize that, although our results could seem like encouragement for abandoning latent variable methods, we echo Sijtsma et al. (2024b) in saying that the type of evidence and arguments we provided should not be understood like that. Rather, given the limitations of our simulations, in an ideal psychometric framework, the main implication of our study is that sum score should be used only after one has established the fit of an IRT or CFT model to the data. However, we also add to this idea that if the intended use involves selection processes (such as large-scale assessment) or thresholding (such as establishing clinical criteria), maybe Likert scoring should be avoided. The bias identified in the extreme values could add noise to the estimates that could result in injustices and unreliable estimates that could be very harmful in high-stake conditions (Franco et al., 2023).

Acknowledgments

There are no mentions.

Funding

This research received no source of financing being funded with resources of the authors themselves

Authors' contributions

We declare that all the authors participated in the elaboration of the manuscript. Specifically, all authors participated in the initial wording of the study – conceptualization, investigation, visualization, all authors participated in the data analysis, and all the authors participated in the Final Writing of Work – Review and Editing.

Availability of data and materials

All data and syntax generated and analyzed during this research will be treated with complete confidentiality due to the Ethics Committee for Research in Human Beings requirements. However, the dataset and syntax that support the conclusions of this article are available upon reasonable request to the principal author of the study.

Competing interests

The authors declare that there are no conflicts of interest.

References

Agresti, A. (2003). *Categorical data analysis.* John Wiley & Sons.

Alphen, A., Halfens, R., Hasman, A., & Imbos, T. (1994). Likert or Rasch? Nothing is more applicable than good theory. *Journal of Advanced Nursing, 20*(1), 196-201. https://doi.org/10.1046/j.1365-2648.1994.20010196.x

Arbuckle, J. L. (2012). *IBM SPSS Amos 21 user's guide.* Amos Development Corporation, SPSS Inc.

Bacon, D. (2004). The contributions of reliability and pretests to effective assessment. *Practical Assessment, Research & Evaluation, 9*(3), 1-8. https://core.ac.uk/download/pdf/239583870.pdf

Bamber, D., & van Santen, J. P. (2000). How to assess a model's testability and identifiability. *Journal of Mathematical Psychology, 44*(1), 20-40. https://doi.org/10.1006/jmps.1999.1275

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika, 46*(4), 443-459. https://doi.org/10.1007/BF02293801

Borsboom, D. (2005). *Measuring the mind: Conceptual issues in contemporary psychometrics.* Cambridge University Press.

Byrne, B. M. (2016). *Structural equation modeling with AMOS: Basic concepts, applications, and programming.* Routledge.

Coenders, G., Batista-Foguet, J. M., & Satorra, A. (1995). Scale dependence of the true score MTMM model. *The Multitrait-Multimethod Approach to Evaluate Measurement Instruments. Eötvös University Press*, Budapest, 71-87.

Colman, A. M., Norris, C. E., & Preston, C. C. (1997). Comparing rating scales of different lengths: Equivalence of scores from 5-point and 7-point scales. *Psychological Reports, 80*(2), 355-362. https://doi.org/10.2466/pr0.1997.80.2.355

Crocker, L. M., & Algina, J. (1986). *Introduction to classical and modern test theory.* Holt, Rinehart, and Winston.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*(3), 297-334. https://doi.org/10.1007/BF02310555

DiStefano, C., Zhu, M., & Mindrila, D. (2009). Understanding and using factor scores: Considerations for the applied researcher. *Practical Assessment, Research & Evaluation, 14*(20), 1-11. https://core.ac.uk/download/pdf/239584512.pdf

Dunn, T. J., Baguley, T., & Brunsden, V. (2014). From alpha to omega: A practical solution to the pervasive problem of internal consistency estimation. *British Journal of Psychology, 105*(3), 399-412. https://doi.org/10.1111/bjop.12046

Fan, X. (1998). Item response theory and classical test theory: An empirical comparison of their item/person statistics. *Educational and Psychological Measurement, 58*(3), 357-381. https://doi.org/10.1177/0013164498058003001

Flake, J. K., & Fried, E. I. (2020). Measurement schmeasurement: Questionable measurement practices and how to avoid them. *Advances in Methods and Practices in Psychological Science, 3*(4), 456-465. https://doi.org/10.1177/2515245920952393

Fox, J. P. (2010). *Bayesian item response modeling: Theory and applications.* Springer.

Franco, V. R., Laros, J. A., Wiberg, M., & Bastos, R. V. S. (2022). How to think straight about psychometrics: Improving measurement by identifying its assumptions. *Trends in Psychology, 32*, 786-806. https://doi.org/10.1007/s43076-022-00183-6

Franco, V. R., Wiberg, M., & Bastos, R. V. S. (2023). Nonparametric Item Response Models: A Comparison on Recovering True Scores. *Psico-USF, 28*(4), 685-696. https://doi.org/10.1590/1413-82712023280403

Gignac, G. E., & Szodorai, E. T. (2016). Effect size guidelines for individual differences researchers. *Personality and Individual Differences, 102*, 74-78. https://doi.org/10.1016/j.paid.2016.06.069

González, J., & Wiberg, M. (2017). *Applying test equating methods: using R.* Springer.

Gorsuch, R. (1983). *Factor analysis.* Erlbaum Associates.

Grice, J. W. (2001). Computing and evaluating factor scores. *Psychological Methods, 6*(4), 430-450. https://doi.org/10.1037/1082-989X.6.4.430

Harman, H., & Jones, W. (1966). Factor analysis by minimizing residuals (minres). *Psychometrika, 31*(3), 351-368. https://doi.org/10.1007/BF02289468

Hair, J. F., Black, W. C., Babin, B. J., Anderson, R. E., & Tatham, R. L. (2006). *Multivariate Data Analysis.* Pearson Education Inc.

Henson, R., Templin, J., & Douglas, J. (2007). Using efficient model based sum-scores for conducting skills diagnoses. *Journal of Educational Measurement, 44*(4), 361-376. https://doi.org/10.1111/j.1745-3984.2007.00044.x

Hinz, A., Einenkel, J., Briest, S., Stolzenburg, J. U., Papsdorf, K., & Singer, S. (2012). Is it useful to calculate sum scores of the quality of life questionnaire EORTC QLQ-C30? *European Journal of Cancer Care, 21*(5), 677-683. https://doi.org/10.1111/j.1365-2354.2012.01367.x

Lawson, S. (1991). One parameter latent trait measurement: Do the results justify the effort. In B. Thompson (Ed.), *Advances in educational research: Substantive findings, methodological developments* (pp. 159-168). JAI.

Liu, S., Suzuki, T., Relator, R., Sese, J., Sugiyama, M., & Fukumizu, K. (2017). Support consistency of direct sparse-change learning in Markov networks. *The Annals of Statistics, 45*(3), 959-990. https://doi.org/10.1214/16-AOS1470

Lord, F. M. (1980). Applications of item response theory to practical testing problems. Lawrence Erlbaum.

Magis, D., Yan, D., & von Davier, A. A. (2017). *Computerized adaptive and multistage testing with R: Using packages catR and mstR.* Springer.

Marszalek, J. M., Barber, C., Kohlhart, J., & Cooper, B. H. (2011). Sample size in psychological research over the past 30 years. *Perceptual and Motor Skills, 112*(2), 331-348. https://doi.org/10.2466/03.11.PMS.112.2.331-348

McDonald, R. P. (1999). *Test theory: A unified treatment.* Psychology Press.

McNeish, D. (2024). Practical implications of sum scores being psychometrics' greatest accomplishment. *Psychometrika, 89,* 1148-1169. https://doi.org/10.1007/s11336-024-09988-z

Mislevy, R. J. (2024). Are Sum Scores a Great Accomplishment of Psychometrics or Intuitive Test Theory? *Psychometrika, 89*, 1170-1174. https://doi.org/10.1007/s11336-024-10003-8

Morgado, F. F., Meireles, J. F., Neves, C. M., Amaral, A. C., & Ferreira, M. E. (2018). Scale development: ten main limitations and recommendations to improve future research practices. *Psicologia: Reflexão e Crítica, 30*(1), 1-20. https://doi.org/10.1186/s41155-016-0057-1

Norman, G. (2010). Likert scales, levels of measurement and the "laws" of statistics. *Advances in Health Sciences Education, 15*(5), 625-632. https://doi.org/10.1007/s10459-010-9222-y

Pearl, J. (2009). *Causality.* Cambridge University Press.

R Core Team (2024). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/

Ramsay, J. O., & Wiberg, M. (2017). A strategy for replacing sum scoring. *Journal of Educational and Behavioral Statistics, 42*(3), 282-307. https://doi.org/10.3102/1076998616680841

Ramsay, J., Wiberg, M., & Li, J. (2020). Full information optimal scoring. *Journal of Educational and Behavioral Statistics, 45*(3), 297-315. https://doi.org/10.3102/1076998619885636

Revelle, W. (2014). psych: Procedures for psychological, psychometric, and personality research. *Northwestern University, Evanston, Illinois*, 165.

Rizopoulos, D. (2007). ltm: An R package for latent variable modeling and item response analysis. *Journal of Statistical Software, 17*, 1-25. https://doi.org/10.18637/jss.v017.i05

Rosseel, Y. (2012). lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical Software, 48*(2), 1-36. https://doi.org/10.18637/jss.v048.i02

Schmitt, N. (1996). Uses and abuses of coefficient alpha. *Psychological Assessment, 8*(4), 350-353. https://psycnet.apa.org/doi/10.1037/1040-3590.8.4.350

Sijtsma, K. (2012). Psychological measurement between physics and statistics. *Theory & Psychology, 22*(6), 786-809. https://doi.org/10.1177/0959354312454353

Sijtsma, K., Ellis, J. L., & Borsboom, D. (2024a). Recognize the Value of the Sum Score, Psychometrics' Greatest Accomplishment. *Psychometrika, 89*(1), 84-117. https://doi.org/10.1007/s11336-024-09964-7

Sijtsma, K., Ellis, J. L., & Borsboom, D. (2024b). Rejoinder to McNeish and Mislevy: What Does Psychological Measurement Require? *Psychometrika*, 1-11. https://doi.org/10.1007/s11336-024-10004-7

Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics.* Pearson.

Thompson, B. (2004). *Exploratory and confirmatory factor analysis: Understanding concepts and applications.* American Psychological Association.

Thurstone, L. L. (1935). *The vectors of mind: Multiple-factor analysis for the isolation of primary traits.* University of Chicago Press.

van der Linden, W. J., & Hambleton, R. K. (Eds.) (2013). *Handbook of modern item response theory.* Springer.

Wallmark, J., Ramsay, J. O., Li, J., & Wiberg, M. (2024). Analyzing polytomous test data: A comparison between an information-based IRT model and the generalized partial credit model. *Journal of Educational and Behavioral Statistics, 49*(5), 753-779. https://doi.org/10.3102/10769986231207879

Wirth, R. J., & Edwards, M. C. (2007). Item factor analysis: Current approaches and future directions. *Psychological Methods, 12*(1), 58-79. https://doi.org/10.1037/1082-989X.12.1.58

recebido em abril de 2021
aprovado em novembro de 2024

Sobre os autores

Víthor Rosa Franco is an assistant professor in the Post-graduate Program in Psychology at University of San Francisco. His research interests include measurement theory and quantitative modeling, being especially interested in Bayesian methods.

Marie Wiberg is a full professor in Statistics at the Department of Statistics of the Umeå School of Business, Economics and Statistics, Umeå University. Her research interests include educational measurement and psychometrics in general.

Como citar este artigo

Franco, V. R. & Wiberg, M. (2025). How to Score Respondents? A Monte Carlo Study Comparing Three Different Procedures. *Avaliação Psicológica, 24*, e22224, 1-10. http://doi.org/10.15689/ap.2025.24.e22224