Paidéia
2022, Vol. 32, e3223.doi:https://doi.org/10.1590/1982-4327e3223
ISSN 1982-4327 (online version)

Psychological Evaluation

Theoretical and Practical Foundations of Mokken Scale Analysis in Psychology

Víthor Rosa Franco[1]
Jacob Arie Laros[2]
Rafael Valdece Sousa Bastos[1]

Abstract: Item Response Theory represents one of the major advances in the field of developing valid and reliable measures in psychology. Among the main models used in this perspective are the Rasch model and the logistic models. These parametric models, however, are not suitable for all applications in psychology, since a substantial number of databases in psychology do not satisfy the assumptions of these models: unidimensionality; latent monotonicity; local independence; and, for some models, non-intersecting functions. Given this framework, the objective of this study was to present the theoretical and practical foundations of Mokken Scale Analysis (MSA). We present some historical issues involving the development of MSA, in addition to the main characteristics and assumptions of the two models used in this perspective. After exemplifying a MSA application, limitations and final considerations are presented, supporting the decision-making process for researchers who come to use MSA.

Keywords: nonparametric inference, item response theory, measurement

Fundamentos Teóricos e Práticos da Análise de Escala de Mokken em Psicologia

Resumo: A Teoria de Resposta ao Item representa um dos principais avanços para a construção de medidas válidas e confiáveis em psicologia. Entre os principais modelos utilizados nessa perspectiva estão o modelo de Rasch e os modelos logísticos. Esses modelos paramétricos, no entanto, não podem ser utilizados em todas as aplicações em psicologia, uma vez que um número substancial dos bancos de dados em psicologia não satisfaz os pressupostos desses modelos: unidimensionalidade; monotonicidade latente; independência local; e, para alguns modelos, não-interseção de funções. Dessa forma, o objetivo deste estudo foi apresentar os fundamentos teóricos e práticos da Análise de Escala de Mokken (AEM). São apresentadas questões históricas envolvendo o desenvolvimento da AEM, além das principais características e pressupostos dos dois modelos usados nessa perspectiva. Após exemplificação de uma AEM, limitações e considerações finais são apresentadas, apoiando o processo de tomada decisão para pesquisadores que venham a usar a AEM.

Palavras-chave: inferência não-paramétrica, teoria de resposta ao item, medidas

Fundamentos Teóricos y Prácticos del Análisis de la Escala de Mokken en Psicología

Resumen: La Teoría de Respuesta al Ítem representa uno de los mayores avances en el campo del desarrollo de medidas válidas en psicología. Entre los principales modelos utilizados en esta perspectiva se encuentran los modelos logísticos. Estos modelos no son adecuados para todas las aplicaciones en psicología, ya que algunas bases de datos en psicología no satisfacen las suposiciones de estos modelos: unidimensionalidad; monotonicidad latente; e independencia local; y, para algunos modelos, funciones que no se interceptan. Teniendo en cuenta este marco, el objetivo de este estudio fue presentar los fundamentos teóricos y prácticos del Análisis de la Escala de Mokken (AEM). Presentamos algunas cuestiones históricas relacionadas con el desarrollo de AEM, además de las principales características y suposiciones de los dos modelos utilizados en esta perspectiva. Después de ejemplificar un AEM, se presentan las limitaciones y consideraciones finales, apoyando o procesando la tomada de decisión para investigadores que van a usar el AEM.

Palabras clave: inferencia no parametrica, teoría de respuesta al item, medidas

[1] *Universidade São Francisco, Campinas-SP, Brazil*

[2] *Universidade de Brasília, Brasília-DF, Brazil*

Correspondence address: Víthor Rosa Franco. Universidade São Francisco. R. Waldemar César da Silveira, 105 Jardim Cura D'Ars (SWIFT), Campinas-SP, Brazil. CEP 13045-510. Campinas-SP. E-mail: vithorfranco@gmail.com

Among psychologists and psychometricians, it is quite common to use techniques related to Factor Analysis (FA) and Item Response Theory (IRT) to gather evidence of validity for psychological instruments (Mair, 2018). Most of these techniques are categorized as parametric. This means that

they are based on statistical models that limit what is called a "good measure". An alternative to using parametric models is nonparametric models (Sijtsma & van der Ark, 2017). Nonparametric models are models that do not make specific restrictions on the type of functional relationship expected to be found between the variables included in the model; in the case of IRT, aptitude and response probabilities.

One of the most promising nonparametric psychometric analysis techniques is the Mokken Scale Analysis (MSA; Mokken, 1971). In the mathematical scope of MSA, tests are used to allow a wider range of observations to form psychometric measures (Sijtsma & Molenaar, 2002). In order to present the advantages that can be achieved by using MSA, the objective of this study was to present the theoretical and practical foundations of Mokken Scale Analysis (MSA). The following sections will present the history and conceptualization of this practice, followed by the definition of basic concepts of MSA. Next, the two MSA models are presented. To support the use of MSA, essential information is presented in the report of this analysis, in addition to exemplification of an application. Finally, the main limitations and conclusions are presented.

Historical Conceptualization of MSA

The concept of item response functions emerged around 1950 (Gregory, 2014). It is possible to identify two main authors for the popularization of using item response models. The first is Georg Rasch, creator of the Rasch model, which uses a logistic function, with additive effects between individual aptitude and item difficulty (Bond, Yan, & Heene, 2021). Frederic M. Lord, in 1980, proposed extensions of the Rasch model, generating the family of models known as "logistic models" (Gregory, 2014). Such models are named like this as they use extensions of the logistic function as the item response function. IRT, however, only became more widely used around the 1970s, when both its advantages over traditional psychometric methods and the use of computers to perform the analysis became popular.

In the Brazilian context, IRT receives attention mainly due to its applications in large-scale educational assessments in the National System of Basic Education Assessment (*Sistema Nacional de Avaliação da Educação Básica*, SAEB) and in the National High School Exam (*Exame Nacional do Ensino Médio*, ENEM; Gonçalves & Dias, 2018). While the SAEB is aimed at evaluating the performance in Portuguese and Mathematics of students enrolled in the fifth and ninth grade of elementary school and in the third grade of high school, ENEM is an assessment used to evaluate the competence of students who are completing, or have already completed, high school and wish to enter a university in Brazil. The application of IRT in these contexts allows the comparability of performances over the years, serving as a basis for evaluating academic development at the national level, as well as ensuring, to some extent, evaluation fairness and equality.

IRT models are generally referred to as latent trait models. This denomination is used to emphasize that the item response process is explained by constructs hypothesized from the content of the items, in addition to other processes of validation of the measures (Slaney, 2017). Given the statistical sophistication related to IRT, it is possible to assess characteristics of items and tests more fully, allowing more complex analyses and uses. For example, adaptive testing (Magis, Yan, & von Davier, 2017), a set of procedural techniques that aims to decrease the quantity of items a participant must answer, is basically possible only when using IRT, although nonparametric alternatives exist.

Two main theoretical contributions that preceded and inspired the creation of IRT were factor analysis (FA; Mair, 2018) and the Guttman scale (Sijtsma & van der Ark, 2017). FA is a statistical technique that is used to estimate how well one or more latent variables can explain the variability of observed scores, used in psychology generally with the objective of identifying evidence of structural validity of an instrument. However, one of the main criticisms of using FA in psychological data is that they are usually measured at the ordinal level, while the statistical model of FA assumes that these data are measured at the interval or ratio level (Zhang, Chen, & Liu, 2020). It is worth noting that FA with a polychoric correlation matrix considers that the data are measured at the ordinal level. However, polychoric correlation assumes that the observed variable was generated from a discretization of a latent continuous variable with a normal distribution. Thus, FA starts from the estimation of both the correlation matrix and the factorial parameters, which increases the chance of bias, especially when the assumptions of the polychoric correlation are not met.

On the other hand, the Guttman scale (Sijtsma & van der Ark, 2017) was created to be applied to tests consisting of binary items, assuming that the response pattern of the respondents is deterministic. The Guttman scale consists of a one-dimensional set of items, which are ranked in order of difficulty, from least to the most difficult. Since the response pattern is considered deterministic, the set of possible responses is predictable. This means that, in a given application, any participant who misses a certain item will not be able to hit any of the following items, since these are more difficult ones. However, empirical data show that it is not uncommon to find patterns in the data that contradict this assumption (Engelhard, 2008). Hitting a more difficult item after giving a wrong answer on an easier item is called a Guttman error. In the Guttman model, Guttman errors are not allowed. Thus, IRT models can be considered stochastic versions of the Guttman scale, being relevant to the present context the Rasch model and the Mokken Scale Analysis.

The Rasch model was one of the first stochastic item response models. It predicts that the chance that a respondent will answer an item correctly, $P(X = 1)$, is described by an item response function (IRF), determined by the following equation:

$$P(X = 1) = \frac{e^{\theta-\delta}}{1+e^{\theta-\delta}} \qquad (1)$$

where θ is the latent aptitude of the individual and δ the difficulty of the item. It is important to point out that despite the applicability of the Rasch model in different contexts, its use is also criticized. The first issue is that the model assumes that all items present the same level of discrimination, defined as the degree to which the item differentiates between individuals with different levels of the latent trait (Bond et al., 2021). Other parametric models, such as the two-parameter logistic model (2PLM), were created to overcome this problem. However, another criticism is related precisely to the fixed form of these models, which will give rise to a "S" shaped IRF, which does not always adequately fit the data (Wiberg, Ramsay, & Li., 2018).

Seeking to solve this problem, Mokken developed his Mokken Scale Analysis procedure (MSA; Mokken, 1971). The MSA is similar to the Rasch model and the 2PLM in that they are all probabilistic models of the Guttman scale. However, the MSA is a model described as "nonparametric", since it does not assume the exact form of the IRF, sustaining more flexible versions of the assumptions of unidimensionality, monotonicity, and local independence, in addition to the assumption of non-intersection for one of its models (Sijtsma & van der Ark, 2017). Thus, there are two models that are derived from the MSA: the more severe model of Double Monotonicity, in which the items can differ in their difficulty, but cannot intersect, as in the Rasch model; and the less severe model of Monotone Homogeneity, in which items differ in difficulty and may intersect, which resembles the 2PLM.

Assumptions and Models of the MSA

MSA, being an IRT model, is based on the idea that the combined effects of a latent variable, called item difficulty, with another latent variable, called aptitude of individuals, affects the probability of response of individuals to a set of items on an instrument (Andrade, Laros, & Lima, 2021). However, as MSA is nonparametric, the logistic model cannot be used to estimate the values of these latent variables. Thus, MSA models use Likert scores (Sijtsma & Molenaar, 2002), also known as sum scores, to generate estimates of the aptitudes of the respondents in the sample. This procedure is considered acceptable because, asymptotically, the sum scores tend to approach the true score (Sijtsma & Molenaar, 2002). Based on these scores, different procedures are used to test the four general assumptions of the MSA: unidimensionality; monotonicity; local independence; and non-intersection of the IRFs. The first three are general assumptions common to parametric IRT, while the fourth is generally an implicit assumption in many parametric IRT models.

An important fundamental difference between MSA and traditional IRT modeling is highlighted. As it is nonparametric, MSA does not establish a "S" shaped IRF like the ones seem in graphical representations that relate the latent variables to the probabilities of response for parametric models. For this reason, MSA cannot be used to estimate the latent scores of individuals. Therefore, tunning of MSA models is conducted differently. In a parametric IRT model, the dimensionality of the items is usually tested using some technique such as Parallel Analysis (Irwing, Booth, & Hughes, 2018; Mair, 2018), which is then followed by the adjustment of the desired IRT model. The individual adjustment indices of items and respondents are used to exclude or even discard models. In MSA, as it does not present a specific model to be adjusted, the approach used is to validate the estimates made using sum scores by testing the assumptions of the models (Sijtsma & Molenaar, 2002).

Unidimensionality is the idea that a single latent trait from the individuals interacts with latent characteristics of the items, expressed, for instance, in a parametric model of IRT such as Equation 1, in which the symbol θ (theta) represents the aptitude of the respondents and the symbol δ (delta) represents the difficulty of the items. Local independence is the idea that the observed correlation, or dependence, among items is explained exclusively by θ, and multidimensional models apply an extension of this assumption. Latent monotonicity (as distinguished from observed monotonicity) represents the idea that if an individual has more of the latent trait, then his probability of giving a correct answer, or to use a higher ordered category on a scale, should also increase. In parametric models, the function that represents this assumption is "S" shaped. However, in nonparametric models, given that there is no specific function to relate latent traits with the probability of correctness of the items, any function that is positively increasing can be used. Finally, non-intersecting means that item-item response functions must not intersect.

From these assumptions, two models of MSA are derived. The first model respects the first three assumptions (unidimensionality, local independence, and latent monotonicity). This model is called the monotone homogeneity model (MHM; Mokken, 1971), and is also known as the nonparametric gradual response model. When the four assumptions of unidimensionality, local independence, latent monotonicity, and non-intersection are respected, the double monotonicity model (DMM) can be used. The main difference between these two models is that the MHM allows ranking only the respondents, while the DMM allows ranking both respondents and items. This feature of the DMM is known as invariant item ordering (IIO), which means that the ordering of items according to their average score is the same for all values of the latent scale, thus allowing the ordering of items by their difficulty levels (Sijtsma & Molenaar, 2002).

To test the assumptions of the models, the main index used is the Loevinger scalability coefficient, H (Loevinger, 1948). There are three scalability indices: the item pairs index (H_{ij}); the item index (H_j); and the general index of the test (H). Equations (2), (3), and (4) respectively represent the ways to calculate such indices.

$$H_{ij} = \frac{\mathrm{COV}(X_i, X_j)}{\mathrm{COV}(X_i, X_j)^{\max}} \quad (2)$$

$$H_j = \frac{\mathrm{COV}(X_j, R_{-j})}{\mathrm{COV}(X_j, R_{-j})^{\max}} \quad (3)$$

$$H = \frac{\Sigma_{j=1}^{J}\mathrm{COV}(X_j, R_{-j})}{\Sigma_{j=1}^{J}\mathrm{COV}(X_j, R_{-j})^{\max}} \quad (4)$$

In this notation, X_i is the sum score of item i, X_j is the sum score of item j, and R_{-j} is the rest score of the test when item j is disregarded, which is simply the sum score of all the items minus item j. It is possible to observe that the scalability indices depend on the covariance between the items (for the pairs of items indices), on the covariance between some item j and the rest score (for the individual item index) and on the sum of these covariance of the items with the total score. The superscript *max* indicates the maximum covariance that two items could have if there were no Guttman errors.

According to the Guttman scale, an individual with an aptitude greater than the difficulty of the item will necessarily always hit or mark the answer to such an item positively. Correspondingly, an individual with an aptitude lower than the item's difficulty will necessarily always make a mistake or mark the answer to that item negatively. From this, the expected scalogram can be estimated by the empirical scalogram, which is defined by the collected data. Keeping the marginal distributions constant (i.e., the sums of the rows and columns do not change), the cell in which a difficult item is correct ($X_D = 1$), but an easy item is wrong ($X_F = 0$) must be set to zero and the other cells must be modified accordingly. Thus, MSA determines that the more similar the empirical scalogram is to the expected scalogram, the more strongly the items are related and, therefore, must represent the same construct. Deviations from the empirical scalogram from the expected scalogram are called Guttman errors.

Finally, it is worth noting that the theoretical values of the H indices can vary between –1 and +1, and the assumptions of unidimensionality, local independence, and latent monotonicity imply: $0 \leq H_{ij} \leq 1$, for all $i \neq j$; $0 \leq H_j \leq 1$, for all j; and $0 \leq H \leq 1$. This means that, if the assumptions are respected, the observed values of the H indices should not be less than 0, although it is possible to observe negative values when the items are not suitable for the scale (Sijtsma & Molenaar, 2002). This means that the calculation of scalability coefficients, besides being descriptive, also serves predictive purposes of the quality of the measures, allowing more robust inferences.

How to Conduct an MSA? Application and Exemplification in Four Steps

The use of MSA does not differ much from the use of traditional psychometric models. This means that, first, the dimensionality of the scale is assessed and then the quality of the model adjustment is tested (Sijtsma & Molenaar, 2002). The main difference lies in the fact that, while parametric IRT models test the quality of items according to fixed assumptions, MSA tests these assumptions directly. For example, while a Rasch model will always impose the same level of discrimination on all items, MSA tests the intersection of the item response functions and, if no intersection is desired, items with such a characteristic are discarded. Thus, the quality of the model depends on which assumptions are used and how well those assumptions are met by the data.

The example presented below was fully analyzed using the mokken package (van der Ark, Koopman, Straat, & van den Bergh, 2021) of the R software (R Core Team, 2022). Currently, as far as we know, this is the only free software alternative to perform MSA. We used a database available in the mokken package, with responses to 12 dichotomous items administered to 425 children from 2nd to 6th grade in The Netherlands (Verweij, Sijtsma, & Koops, 1996). Each item is a transitive reasoning task about physical properties of objects, with two items used as pseudo-items (items 11 and 12), four items about length relationships (items 01, 02, 07, and 09), five items about width relationships (items 03, 04, 05, 08, and 10), and one item related to area relations (item 06). The code used to conduct the analyzes can be accessed at: https://github.com/vthorrf/TutorialMokken.

First step: Dimensionality analysis. From the MSA perspective, the dimensionality analysis is performed through the Automated Item Selection Procedure (AISP; Mokken, 1971; Sijtsma & Molenaar, 2002). The AISP uses the scalability coefficient H_i to select the most representative item of the dimension and then, uses the scalability coefficient of pairs of items to select the largest subset of items that measure the same attribute (Mokken, 1971). After selecting the best items for the first dimension, unselected items are tested as an attempt to compose a second subscale, and so on, until it is no longer possible to allocate any item to any subscale.

A simulation study showed that among three traditional AISP implementations, the one that uses a genetic algorithm has the best performance in recovering the correct dimensionality of scales (Straat, van der Ark, & Sijtsma, 2013). It has also been identified in this study that the scalability coefficient of item pairs, using the best item as a reference, should not be less than 0.30. Sijtsma and Molenaar (2002) also suggest that it is necessary to use several possible limits for the relationship with the best item, starting from the value of 0.30, in order to ensure greater richness of the analysis. Using these recommendations, Table 1 was generated in which all items are represented in the rows and the minimum values of the scalability coefficient (H_j) of the best item represented in the columns.

It was expected that the pseudo-items would not be aggregated to any subscale, and this was exactly the result obtained. It can also be observed that the higher the minimum value of the scalability coefficient with the best item, the fewer the items kept in the scales. In general, the AISP identified that, at most, two scales can be generated, represented by the numbers 1 and 2. The empty spaces, per column, indicate that, using that limit, the respective item does not form a scale with any other item.

Table 1
Dimensionality analysis of the transitive reasoning test

		Scalability Index, H_j										
Item	Content	0.30	0.35	0.40	0.45	0.50	0.55	0.60	0.65	0.70	0.75	0.80
09	Length	1	1	1	1	1	1	1	2	2	2	1
12	Pseudoitem											
10	Width	1	1	1	1	1	2	2	1			
11	Pseudoitem											
04	Width	2	2									
05	Width											
02	Length	2	2									
07	Length	1	1	1	1	1	1			1	1	
03	Width	1	1	1	1	1	2					
01	Length	1	1	1	1							
08	Width	1	1	1	1	1	1	1	2	2	2	1
06	Area	1	1	1	1	1	1	2	1	1	1	

The transitive reasoning test was designed to be unidimensional. Therefore, using as a reference the scale numbered as 1 (because it is the most frequent at all levels), we can observe that this subscale is constant up to the limit of 0.45. This means that a very robust scale may probably be created using items 01, 03, 06, 07, 08, 09, and 10. Thus, as expected, pseudo-items 11 and 12 would be discarded, in addition to items 02, 04, and 05, which probably have more Guttman errors than would be expected for unidimensional items. On the other hand, the second scale does not show consistency when varying the limits of the scalability coefficient, which indicates that it is probably a spurious scale.

Second step: Latent monotonicity analysis. Junker and Sijtsma (2000) showed that, for dichotomous items, latent monotonicity implies observed monotonicity. Although for polytomous items this is not always true, tests of observed monotonicity also generate good estimates for polytomous items, although more conservative ones. The observed monotonicity test proposed by the authors involves a regression between the scores of individual items and the rest scores, which are obtained by omitting the selected item from the total test score. In other words, considering the scale found from the AISP, to test the observed monotonicity of item 09, for example, this item is regressed on the Likert score from items 01, 03, 06, 07, 08, and 10. If significant non-monotonic increments are detected between both variables, we infer that there is probably no observed and no latent monotonicity in item 09.

One problem with using rest scores to test latent monotonicity is that the number of respondents at different score levels can be very small (Sijtsma & Molenaar, 2002). This problem can be overcome by grouping respondents with adjacent rest scores until a minimum proportion of individuals per score is greater than a pre-defined criterion. However, using n/10 as a default for such criterion, if the sample (n) is greater than or equal to 500; n/5 if the sample is between 250 and 500; and max (n/30,50) if the sample is less than 250, robust results will generally be obtained.

Using only the items that were kept after the AISP, we used three criteria for the minimum score union value: the default criterion; the number of possible scores; and the ratio between sample size and scale size. That is, with binary items on a seven-item scale, the lowest possible residual score is 0 and the highest possible rest score is 6, which represents 7 possible score categories. Thus, the criteria were equal, respectively, to 425/5 = 85; 7; and 425/7 ≈ 61. Table 2 presents the items, the scalability indices of each item (H_j), the number of active pairs (AP)—which represents the maximum possible amount of monotonicity tests for each item—, the number of violations (Vi) of monotonicity that were identified for each item, the magnitude of the largest violation (MaxVi), the z-value of this largest violation (Zmax) for inferential testing, and the number of violations that were significant in each item (Zsig).

The first thing to note is that, using different criteria, no monotonicity violation was ever observed. In part, this probably occurred since the scalability coefficient, used in the AISP, tends to keep items that are monotonic in relation to their dimension (Sijtsma & Molenaar, 2002). However, the AISP will not always select only monotonic items, which justifies this analysis. Moreover, the number of APs, in some cases, was equal to 1 or 0. When this occurs, it is not possible to adequately test the monotonicity in that item. Very high values should also be avoided since response categories that were not very expressive can generate spurious confirmations of monotonicity violation.

Table 2
Analysis of observed monotonicity of items on the transitive reasoning scale

Items	H_j	Criterion	AP	Vi	MaxVi	Zmax	Zsig
		85	1	0	-	-	-
09	0.50	61	3	0	-	-	-
		7	10	0	-	-	-
		85	3	0	-	-	-
10	0.52	61	3	0	-	-	-
		7	10	0	-	-	-
		85	3	0	-	-	-
07	0.51	61	3	0	-	-	-
		7	15	0	-	-	-
		85	3	0	-	-	-
03	0.53	61	3	0	-	-	-
		7	10	0	-	-	-
		85	3	0	-	-	-
01	0.46	61	6	0	-	-	-
		7	15	0	-	-	-
		85	1	0	-	-	-
08	0.55	61	3	0	-	-	-
		7	10	0	-	-	-
		85	0	0	-	-	-
06	0.59	61	1	0	-	-	-
		7	6	0	-	-	-

Note. H_j = scalability of items; AP = number of active pairs of rest scores; Vi = number of violations of monotonicity; MaxVi = largest violation of monotonicity; Zmax = z-score of the largest violation; Zsig = number of statistically significant violations.

Thus, from Table 2, we can see that using increasingly lenient criteria, they were not enough to identify monotonicity violations in any of the items. We can also observe that the standard criterion (85) caused many items to present APs equal to 1, which means that it would not be possible to adequately test the monotonicity in these items. Finally, item 06 can only be tested on the most lenient criterion of all, which means that the item probably does not have adequate variability in scores and, therefore, offers little information about the actual score of the respondents. In sum, we can choose to also exclude item 06 from our scale.

Third step: Non-intersection analysis. This third step is optional and depends on which MSA model one intends to use. If it is assumed that the items can intersect, then the monotone homogeneity model (MHM) will be used, and this step is not necessary (although there is inferential value in executing it anyway). However, if it is assumed that the items should not intersect, the double monotonicity model (DMM) will be used, and it is necessary to test if the non-intersection assumption is really upheld in the data.

Sijtsma and Molenaar (2002) describe three methods to test for non-intersection: p-matrix method; rest score method; and residual division method (*restsplit*). Although, to the best of our knowledge, there are no studies that compare the performance of each method, the residual division method has not yet been implemented in the mokken package and the rest score method is affected by the same limitation that was found in the monotonicity analysis: a minimum size must be established a priori for the size of the score clusters. Thus, the p-matrix method is preferred, which can be known in more detail in Mokken (1971). In short, the method creates matrices of partial associations among items, also using the amount of Guttman errors among items as a basis, controlling for the presence of other items. Thus, two matrices are generated: the P(++) matrix, which evaluates the positive associations; and the P(--) matrix, which evaluates negative associations. Using only the items that were kept by the AISP and by the monotonicity analysis, the analysis presented in Table 3 was performed.

Table 3
Non-intersecting analysis of response functions to the transitive reasoning scale item

Items	H_j	AP	Vi	MaxVi	Zmax	Zsig
09	0.49	20	0	-	-	-
10	0.51	20	0	-	-	-
07	0.49	20	0	-	-	-
03	0.53	20	0	-	-	-
01	0.47	20	0	-	-	-
08	0.55	20	0	-	-	-

Note. H_j = scalability of items; AP = number of active pairs of rest scores; Vi = number of violations of non-intersection; MaxVi = largest violation of non-intersection; Zmax = z-score of the largest violation; Zsig = number of statistically significant violations.

To identify the items that do not intersect, just check the Zsig column. If this value is equal to or greater than one, then the item in question has at least one intersection that is statistically significant and, therefore, may present problems for the construction of the scale. Again, although after the removal of the items by the AISP and the analysis of manifest monotonicity, only items that present non-intersecting remained, this will not always be true. For this reason, if the DMM model is to be used, it is a necessary condition to carry out a non-intersection analysis and to remove items that have violations. Unlike the monotonicity analysis, in the case of non-intersection, a single violation can already be considered as critical to reject the DMM, since this violation demonstrates that the items had at least one point of intersection.

Fourth step: Analysis of local independence. The last step involves testing the assumption of local independence. Straat et al. (2016) proposed the use of conditional associations tests, which, if they demonstrate positive covariances between items, indicate the existence of local independence for the items. Conditional associations generate three local independence deviation indices: $W^{(1)}$; $W^{(2)}$; and $W^{(3)}$. These indices are used to identify different types of local independence violations. We present in Table 4 the test only for the items that were kept after the previous analyses.

Table 4
Local independence analysis of the transitive reasoning scale

Index	Item	Item 09	Item 10	Item 07	Item 03	Item 01	Item 08
$W^{(1)}$	09		1.847	1.476	1.329	0.794	0.310
	10	0.725		0.968	1.151	0.657	0.114
	07	0.005	0.936		1.166	1.014	0.023
	03	0.506	1.247	0.353		0.169	0.077
	01	0.310	1.494	1.997	1.933		0.426
	08	0.052	2.087	1.227	2.455	0.400	
$W^{(2)}$		7.811	5.753	5.449	4.691	7.617	7.150
$W^{(3)}$	09						
	10	0.698					
	07	1.967	0.543				
	03	1.623	1.555	0.078			
	01	**2.375**	1.489	1.133	0.626		
	08	1.148	1.468	1.728	0.809	1.996	

To know which item has local dependence, it is necessary to carry out a procedure to find out how extreme is each of the values presented. It is necessary to test each value in the table for the following relationship: $w_{ij} > Q_{i3} + (3 \times [Q_{i3} - M_i])$. In this test, w_{ij} represents each possible value of the index $W^{(i)}$, Q_{i3} is the third quartile of the set of values of the index $W^{(i)}$ and M_i is the median of the set of values of the index $W^{(i)}$. Using such a procedure, it has been verified that only item 1 is considered as extreme in the $W^{(3)}$ index. From this result, the researcher can choose one of two actions. The first is to keep the item, given that it was identified as a local dependent in only one of the three indices. The second, which may be the most appropriate, is to remove the item and redo the analysis, as the results are dependent on the dataset used. Following the second action, Table 5 was generated.

Table 5
Local independence analysis of the transitive reasoning scale after removing item 01

Index	Item	Item 09	Item 10	Item 07	Item 03	Item 08
$W^{(1)}$	09		0.967	0.564	0.464	0.037
	10	0.489		0.553	0.836	0.011
	07	0.005	0.936		1.114	0.018
	03	0.432	1.247	0.353		0.032
	08	0.052	1.473	0.557	1.754	
$W^{(2)}$		3.432	2.446	2.956	2.760	4.097
$W^{(3)}$	09					
	10	0.205				
	07	1.283	0.131			
	03	1.117	0.923	0.089		
	08	0.826	1.187	1.452	0.631	

In Table 5, none of the items was considered as an outlier in the distribution of any of the indices. Therefore, we have thus discovered the best scale that can be generated from the original transitive reasoning scale, having as a criterion the fulfillment of all the necessary assumptions to have a good measure in terms of unidimensionality, latent monotonicity, non-intersection, and local independence.

Discussion

MSA is a special type of nonparametric item modeling that allows checking the robustness of more flexible versions of the assumptions of parametric IRT models. The advantages of using MSA are summarized by Junker and Sijtsma (2001) in three main reasons. The first is to provide a deeper understanding of how IRT parametric models work by directly analyzing their assumptions. Second, it offers a more flexible framework for applications where parametric models fit the data poorly, which allows for more flexible scales than are possible with parametric models. Finally, MSA procedures make it possible to use a smaller number of items and samples (Wind, 2022) than those used in large-scale tests adjusted with parametric models, given that, instead of trying to estimate parameters, MSA tests the assumptions of the models.

In more practical terms, we believe that Stout (2001) offers an orientation that maximizes the quality of scale development. The author suggests that before using any parametric model of IRT, it is necessary to perform an MSA. By doing this, in addition to being able to verify the assumptions necessary to carry out a parametric IRT analysis, it is also possible that, if these assumptions are not met, a more flexible scale is developed, which meets nonparametric IRT assumptions. Among these analyses, one of the most important is the dimensionality analysis, usually assessed through exploratory or confirmatory factor analyses. MSA, being a type of IRT analysis, unlike Factor Analysis, allows the analysis of dimensionality to be better aligned with the theory and mathematical form of the IRT. This makes the MSA the center of a thorough analysis of evidence of structural validity, given its basically necessary condition of evaluating the assumptions that form a scale if its structure is not known a priori (Stout, 2001).

The present study prioritized the exposition of the fundamentals of MSA, as well as the basic analytical procedures to carry it out. Thus, many more advanced discussions ended up being left out. For example, Ligtvoet (2015) proposed a method to assess measure invariance in the context of MSA, a topic that we did not discuss. We also did not discuss how violations of assumptions can be accessed with the effect size named as Crit (Crișan, Tendeiro, & Meijer, 2019). Of course, we did not include many other MSA innovations. However, this happened for two reasons. First, many of these procedures are not yet implemented in any statistical package. Thus, we would not be able to implement the analyzes directly. Second, some of these procedures are new and, therefore,

their properties and validity are still poorly understood. Therefore, it is emphasized that future studies should delve deeper into these more modern procedures and in which scenarios they are valid and provide additional information to the basic procedure in a robust and reliable way.

Finally, it is important to emphasize the existence of other nonparametric IRT models that can be used in order to estimate item parameters and latent scores, as an alternative to parametric models. An example is Wiberg et al. (2018), who proposed the optimal scoring procedure, which apply nonparametric regression techniques, and which can estimate latent scores that do not follow a normal distribution, as long as the scores have a scale with limits (for example, scores ranging from 0 to 10). This model does not exhaust the list of all nonparametric IRT models, but it helps to demonstrate that it is possible to complement MSA to estimate latent item parameters and scores without the need to use parametric IRT models.

References

Andrade, J. M., Laros, J. A., & Lima, K. S. (2021). Teoria de resposta ao item paramétrica e não paramétrica [Parametric and Nonparametric item response theory]. In C. Faiad, M. N. Baptista, & R. Primi (Orgs.), *Tutoriais em análise de dados aplicados à psicometria* [*Tutorials in data analysis applied to psychometrics*] (pp. 183-204). Petrópolis, RJ: Vozes.

Bond, T., Yan, Z., & Heene, M. (2021). *Applying the Rasch model fundamental measurement in the human sciences* (4th ed.). New York, NY: Routledge.

Crişan, D. R., Tendeiro, J., & Meijer, R. (2019). *The Crit value as an effect size measure for violations of model assumptions in Mokken Scale Analysis for binary data*. PsyArXiv. doi:10.31234/osf.io/8ydmr

Engelhard, G., Jr. (2008). Historical perspectives on invariant measurement: Guttman, Rasch, and Mokken. *Measurement: Interdisciplinary Research and Perspectives, 6*(3), 155-189. doi:10.1080/15366360802197792

Gonçalves, F. B., & Dias, B. C. C. (2018). Um estudo da relação entre traço latente e variáveis contextuais no Saeb e Enem [A study relating latent traits to contextual variables for the Saeb and Enem exams]. *Examen: Política, Gestão e Avaliação Da Educação, 2*(2), 152-172. Retrieved from https://examen.emnuvens.com.br/rev/article/view/91

Gregory, R. J. (2014). *Psychological testing: History, principles, and applications* (7th ed.). Boston, MA: Allyn & Bacon.

Irwing, P., Booth, T., & Hughes, D. J. (Eds.). (2018). *The Wiley handbook of psychometric testing: A multidisciplinary reference on survey, scale and test development*. Hoboken, NJ: John Wiley & Sons.

Junker, B. W., & Sijtsma, K. (2000). Latent and manifest monotonicity in item response models. *Applied Psychological Measurement, 24*(1), 65-81. doi:10.1177/01466216000241004

Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement, 25*(3), 258-272. doi:10.1177/01466210122032064

Ligtvoet, R. (2015). A test for using the sum score to obtain a stochastic ordering of subjects. *Journal of Multivariate Analysis, 133*, 136-139. doi:10.1016/j.jmva.2014.09.003

Loevinger, J. (1948). The technique of homogenous tests compared with some aspects of "scale analysis" and factor analysis. *Psychological Bulletin, 45*(6), 507-529. doi:10.1037/h0055827

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

Mair, P. (2018). *Modern psychometrics with R*. Cham, Switzerland: Springer.

Mokken, R. J. (1971). *A theory and procedure of scale analysis*: *With applications in political research.* Berlin, German: De Gruyter.

R Core Team. (2022). *R: A language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org/

Sijtsma, K., & Molenaar, I. W. (2002). *Introduction to nonparametric item response theory*. Thousand Oaks, CA: Sage.

Sijtsma, K., & van der Ark, L. A. (2017). A tutorial on how to do a Mokken scale analysis on your test and questionnaire data. *British Journal of Mathematical and Statistical Psychology, 70*(1), 137-158. doi:10.1111/bmsp.12078

Slaney, K. (2017). *Validating psychological constructs: Historical, philosophical, and practical dimensions*. London, United Kingdom, Mcmillan.

Stout, W. (2001). Nonparametric item response theory: A maturing and applicable measurement modeling approach. *Applied Psychological Measurement, 25*(3), 300-306. doi:10.1177/01466210122032109

Straat, J. H., van der Ark, L. A., & Sijtsma, K. (2013). Comparing optimization algorithms for item selection in Mokken scale analysis. *Journal of Classification, 30*(1), 72-99. doi:10.1007/s00357-013-9122-y

Straat, J. H., van der Ark, L. A., & Sijtsma, K. (2016). Using conditional association to identify locally independent item sets. *Methodology, 12*(4), 117-123. doi:10.1027/1614-2241/a000115

van der Ark, L. A., Koopman, L., Straat, J. H., & van den Bergh, D. (2021). Mokken: Conducts Mokken Scale Analysis. Retrieved from https://cran.r-project.org/web/packages/mokken/index.html

Verweij, A. C., Sijtsma, K., & Koops, W. (1996). A Mokken scale for transitive reasoning suited for longitudinal research. *International Journal of Behavioral Development, 19*(1), 219-238. doi:10.1177/016502549601900115

Wiberg, M., Ramsay, J. O., & Li, J. (2018). Optimal scores: An alternative to parametric item response theory and sum scores. *Psychometrika, 84*(1), 310-322. doi:10.1007/s11336-018-9639-4

Wind, S. A. (2022). Identifying problematic item characteristics with small samples using mokken scale analysis. *Educational and Psychological Measurement, 82*(4), 747-756. doi:10.1177/00131644211045347

Zhang, S., Chen, Y., & Liu, Y. (2020). An improved stochastic EM algorithm for large-scale full-information item factor analysis. *British Journal of Mathematical and Statistical Psychology, 73*(1), 44-71. doi:10.1111/bmsp.12153

Víthor Rosa Franco is a Professor of the Universidade São Francisco, Campinas - SP, Brazil

Jacob Arie Laros is a Professor of the Universidade de Brasília, Brasília-DF, Brazil.

Rafael Valdece Sousa Bastos is a master's candidate of the Graduate Program in Psychology at Universidade São Francisco, Campinas-SP, Brazil

Authors' Contribution:
The first author contributed with the conception of this study. The structure and writing were equally due to all authors, as well as the manuscript revision and approval of the final version. All the authors assume public responsibility for the content of the manuscript.

Associate editor:
Luciana Mourão Cerqueira e Silva

Received: Nov. 05, 2021
1st Revision: May. 04, 2022
2nd Revision: May. 25, 2022
Approved: May. 26, 2022

How to cite this article:
Franco, V. R., Laros, J. A., & Bastos, R. V. S. (2022). Theoretical and practical foundations of Mokken scale analysis in psychology. *Paidéia (Ribeirão Preto), 32,* e3223. doi: https://doi.org/10.1590/1982-4327e3223