PMKT
BRAZILIAN JOURNAL OF MARKETING, OPINION, AND MEDIA RESEARCH

Chief Editor: Fauze Najib Mattar | Associate Editor: Claudio Alberto de Moraes
Evaluation System: Triple Blind Review
Published by: ABEP - Brazilian Association of Research Companies
Languages: Portuguese and English | ISSN: 2317-0123 (online)

Use of item response theory to measure the scales reliability

Uso da teoria de resposta ao item para determinar a confiabilidade de escalas

Rafael Lucian*
Faculdade Boa Viagem – DeVry | FBV, Recife, PE, Brazil

ABSTRACT

This essay is dedicated to theoretical study via what procedures it is possible to consider valid ranges and suitable for use as legitimate scientific instrument. The scales are measurement tools that make up the middle of the science instruments, to build knowledge. The interest of this article is about the statistical treatment to determine the reliability of the scales. Therefore, the proposal here is to discuss in depth if there is sustainable argument for which the Academy adopts the techniques of Item response theory (IRT) as safe for calculating the statistical reliability of scales at the expense of classical techniques such as Cronbach's alpha Coefficient. To this end, inventoried the State of the art concerning the subject and reveal the properties of IRT to the point of being possible to conclude that this is a promising technique that already has immediate application conditions in studies as unique and safe tool for calculating the reliability of scales.

KEYWORDS: Item Response Theory; Reliability of scales; Cronbach's Alpha Coefficient.

RESUMO

Este ensaio teórico dedica-se a estudar por meio de quais procedimentos é possível considerar escalas válidas e aptas para o uso como instrumento científico legítimo. As escalas são ferramentas de mensuração que compõem o instrumentário meio da ciência, a fim de construir conhecimento. O interesse deste artigo, em particular, é sobre o tratamento estatístico para determinar a confiabilidade das escalas. Sendo assim, a proposta aqui é discutir em profundidade se há argumentação sustentável para que a academia adote as técnicas da Teoria de Resposta ao Item (TRI) como estatística segura para o cálculo da confiabilidade de escalas em detrimento de técnicas clássicas como o Coeficiente Alfa de Cronbach. Para tanto, inventariou-se o estado da arte relativo ao tema e revelaram-se as propriedades da TRI ao ponto de ser possível concluir que esta é uma técnica promissora que já possui condições de aplicação imediata nos estudos como ferramenta única e segura para o cálculo da confiabilidade das escalas.

PALAVRAS-CHAVE: Teoria de Resposta ao Item; Confiabilidade das escalas; Coeficiente Alfa de Cronbach.

Submission: 22 August 2016
Approval: 22 March 2017

***Rafael Lucian**
Doctor in business administration from the Federal University of Pernambuco. Coordinator of the Center for research support on a good trip – DeVry College | FBV. (CEP 51200-060, Recife, PE, Brazil).
Email: rlucian@fbv.edu.br
Address: Rua Jean Emile Favre, FBV, Ipsep, 51200-060, Recife, PE, Brazil.

1 INTRODUCTION

Measurement, according to Crowther (1995), is a technique that makes use of precision instruments to measure desired qualities with numerical basis. Therefore, in principle, any observable thing can be measurable, if it is a suitable instrument for this purpose.

However, the process of measurement is broader than the assignment of numbers to objects representing quantitatively any attribute you want to measure; Instead, your goal is to provide a mechanism of review that manages information and serve as a promotion for intelligent decision-making (Pooja & Sagar, 2012).

The scales, from this point of view, serve to decision makers and ordinarily in this vision of quality measurement is, to some degree, the quality of the decision (Pooja & Sagar, 2012). So, like Sanchez, Manning and Sordi (2011) argue, much as market Academy seek measurements that provide better data quality and promote improvements in the process of measurement is a legitimate, necessary and current contribution to the Academy. More objectively, Robertson (2012) States that provide improvements over the processes of measurement becomes a necessity to the Academy. This theoretical test endorses this assumption and cooperate in that direction.

To promote such a contribution, the measurement should confer reliability and validity to the data collected in the field, but several authors (Thompson, 2002; Ten Berge & Socan, 2004; Maroco & Garcia-Marques, 2006; Vieira & Dalmoro, 2008; Sijtsma, 2009) are challenging the ability of existing statistical tests generate scales with such characteristics. The concern of such authors focuses on basic aspects of the classical theory of the tests (CTT), as the sample, variability in this theoretical test, will be primarily evidenced by means of the knowledge of the mathematical properties of Cronbach's alpha Coefficient used dominantly to establish the reliability of scales.

Alternatively, influenced by Psychology and education theorists, researchers (for example, Matteucci, Mignani, & Veldkamp, 2012; Schultz, Salomo & Talke, 2013; Lucian & de Silva, 2015) admit consider using Item response theory (IRT) as a natural substitute current techniques for calculating the reliability, however, there is a need for reflection on such a practice, after all it is an appropriation.

In coach, the central objective of this essay is to discuss in depth theoretical if there is sustainable, argument for the Academy adopts the techniques of IRT as statistics for calculating the reliability of scales at the expense of the traditional coefficient of Cronbach's alpha.

2 CALCULATION OF THE RELIABILITY

The use of scales of measurement requires certain methodological care, because their results become important indicators for decision making. Some performance assessment tools or Psychometry (e.g. satisfaction, loyalty or attitude) are used daily by global corporations who place faith in your ability to read correctly the truth. Contemporary researchers have been developing and testing models and schematic theories from empirical observations based on data collection by survey and, ultimately, in measurement through scales.

However, to align to the assumptions of the dominant scientific paradigm such instruments must be consistent enough to measurement the construct understood accurately, or at least, steadily. Thus, the ability of a scale to measure equally the same level of interaction by means of two or more distinct field interventions, is called reliability. Although very particular, reliability is based on some assumptions of the classical theory of the tests (CTT) who want revealed.

In the classical theory of the tests, the total score (T) is composed of two parts, your true score (V) plus the measurement error (E). Such a mistake is a variable that can assume positive or negative values, allowing the T is greater than or less than the V.

For the reliability of scales, so what if demand is the ratio of the variance of V and T, however such indicator is simply calculated by the fact of not being able to determine the value of V and by theorizing that the real score should have very little or no variation between the tests, resulting in zero variance.

Apart from that, the development of new scales typically involves an intense process of approximation and error. Given the rule, the researchers suggest the new scales, capable of measuring certain items and construct check such ability in the field by applying statistical tests of reliability. Sometimes these pre-test procedures may indicate that the instrument of measurement is not reliable enough and should be modified.

Another scenario is the elaboration of the scale reliability tests to confirm that there were no serious problems of sampling (connected to the permanently non-probability sampling, or worse, for convenience). Such application becomes a rule of thumb, minor but recurring, as a simple concern contradictory between adopt reliable scales and submit them again to the reliability test. Say that a declaration of non-reliability post-test suggests that the pre-test was improper validation (dropping, by the way, sampling errors that would harm any measurement by itself).

To this end, probably the most known and used to estimate the reliability of a scale is calculating Cronbach's alpha coefficient (1951). Such a calculation is based on a process, now in disuse, known as test-retest, which had the goal of testing the same scale two or (preferably) more times with the same sample over time.

A previous technique to Alpha Test, which certainly inspired and can still be found in the main statistical *software,* is to cut it to half or *Split-Half* (CAM), which consists of two equivalent scales in the same elaborate questionnaire. Theoretically, the person should give the same reply when answering the same thing twice by different methods, since the scales had internal consistency. This consistency would validate them together.

The scales validated by the cutting method-to-half should have all your items at least doubled (written differently, but equivalent) and that limited the ability of exploitation, since the limit to the size of the questionnaire was always the respondent's tolerance. If unsown against this technique, Cronbach (1951) stated that, to get a better interpretation, the scale should not be divisible into small blocks minors: must be unique, no subscales and without duplicate scales; that, according to the author cited, would assist in the internal validation of the scale.

The contribution of Cronbach (1951), however, was to produce a simpler calculation, in which the items are correlated with each other internally and not in crosses between the two equivalent scales of CAM. Thus, researchers no longer need to create items, since the alpha considers the item by the average of the scale. The new process has made the search for simpler scales quickly and efficiently.

3 THE TRADITION OF CRONBACH'S ALPHA

The purpose of reliability testing is simple: measure how much the individual is consistent in their answers; However, such measurement has always demanded double data collection, which are in various times or duplicating the scales in the same questionnaire.

Cronbach (1951) was then the revolutionary your time by proposing a form of calculation of reliability in that collection of data, saving you time, effort and ensuring greater agility to empirical research, which is why your Alpha Coefficient is, to the present day, the most widely used method to purify a scale.

The Alpha Coefficient is the indicator of internal consistency and varies of course between 0 and 1, because it is a ratio of the true score and the total, however it can take negative values in some unconventional situations in which there is a negative correlation between the items of the scale.

The assignment of scores below zero for the coefficient can be the result of a reversal in part of the items during the data collection that was not properly corrected final tab, or if the scale is in fact measuring different constructs among its items (Henson, 2001). In practice, however, does not admit negative values, i.e. If the scores are near or below zero the only desirable interpretation is that there is no reliability.

As a rule of thumb, it is assumed that values above 0.6 (Malhotra, 2013) or 0.7 (Gouveia, Saints, & Martin, 2013) are minimum battens to check the scale reliability. However, Streiner (2003) emphasizes that high values on Cronbach's alpha should be read as failed on the measuring scale variability. It would be like if all items were one, so to say reliable, the maximum acceptable score

must be 0.9. Finally, in terms of results, reliable scales if these have Cronbach's alpha values between 0.6 and 0.9.

Although the Alpha for reliability estimate is certainly the most used, is not immune to criticism. For Sijtsma (2009), the Alpha calculation for internal reliability is more a tradition than a technical choice. It is revealed that author when it observes the vast literature that severely criticizes the use of this technique for the purposes of estimation of reliability, for example, Thompson (1994), Vacha-Haase (1998), Wilkinson (1999) and Maroco and Garcia-Marques (2006).

One of the main arguments of these authors is that the natural variability of despises Alpha sample. Maroco and Garcia-Marques (2006) pointed the same instrument features significantly different values if applied to different samples. Thompson (2002) States that the same extent when administered to a sample of homogeneous or heterogeneous more subjects, produces different reliability scores. In such situations, the Alpha Coefficient is not able to measure clearly the reliability of the instrument, what was measured was the homogeneity.

In this same perspective, for theorists like Caruso (2000), Yin and Fan (2000) and Streiner (2003) the Cronbach's alpha is not able to measure the reliability of the scale, because your result is very dependent on the sampling characteristics, as previously commented. Caruso (2000) and Henson, Kogan and Vacha-Haase (2001) emphasize that the more heterogeneous are the sample, the greater the variation in the total score and, consequently, the greater the value of the coefficient of reliability.

In this perspective, the scores obtained by Cronbach's alpha can only be extended if the measurement scale homogeneity is also revealed.

In addition, denotes that there is a high sensitivity Alpha test to the number of cases. Duhacheck and Iacobucci (2004) state that, the smaller the sample size, the greater the value of the estimate of reliability. Thus, Maroco and Garcia-Marques (2006) state that the values for the Alpha should always be interpreted in the light of the characteristics of the extent to which associates and of the population in which this measure is designed, a fact that already recognized previously by the Cronbach (1951) in your seminal publication.

For these reasons, also Ten Berge and Socan (2004) state that the calculation of the coefficient Alpha is not a measure of internal consistency, either a measure of unidimensionalidade. The Alpha Coefficient is strongly dependent on the length of the scale. Curtain (1993) has tested and confirmed that measurements based on ranges with many points raise the values of Cronbach's alpha, even if such variation does not influence anything in the internal consistency and the real score.

Sijtsma (2009) States that, while there is a collective understanding of the calculation of the coefficient Alpha can measure how much all items are measuring the same size, the test can submit high scores when applied both in one-dimensional as multidimensional scales, i.e. does not contribute effectively if the goal is to ensure that only a construct was the object of measurement.

Finally, Pasquali and Primi (2003) state that, in theoretical calculations of the classical tests as the Alpha Coefficient, there is a logical inconsistency, because the score of each item is tested against a total score that consists of all items of the test, including what is being analyzed. This suggests that the other items are suitable or otherwise it wouldn't make sense to be included in the calculations. But if you know, at first, reliability of the items, there would be no sense in testing them.

For all the limitations mentioned above, there is the alternative of Item response theory, which is not just another way of measuring the reliability, but an alternative to CTT.

4 ITEM RESPONSE THEORY

A promising evolution of the use of the Alpha Coefficient is Item response theory which was developed by Psychometrics to assess psychological tests based on a one-dimensional dichotomous latent variable, as in Lord (1952). Due to the complexity of the calculations based on warhead and integral, IRT remained underutilized for decades; However, with the advent of specialized *software* and with the suggestion of Birnbaum (1968) to replace the warhead for the logistic function, the technique became accessible and gained more space at the Academy.

Its most famous application in Brazil is in education. Tests on a large scale, as the national high school Exam (ENEM), are built by several teachers who probably will not be able to work out issues with the same degree of difficulty. However, it would also be unlikely to, intentionally, any test I could repeat the same difficulty in its various versions, or, more unlikely still, measure properly the weight of each correct question according to your difficulty.

By CTT, if a question is marked properly for 80% of the candidates can say that your difficulty is 0.8. Although the calculation is simple, it depends on the skill of the respondents, that is, if for example, the same test is applied to a group of students with less skill, the percentage of hits of the same question you can download to 50% and so the difficulty would be 0.8 for the first group and 0.5 for the second group. In this way, the IRT comes as the solution to measure the difficulty of the question regardless of the sample.

Item response theory allows individuals who have the same number of hits have different scores If the ability of the candidates is also different, after all, your goal is not to count the number of correct answers and Yes measure the ability of the respondent. In fact, the only way to match the scores is in case of coincidence of reply on all issues (Drasgow, Levine, Tsien, Williams, & Mead, 1995).

The IRT, then, second Lord and Novick (1968), calculates the probability of response to the item considering the characteristics of the item (the item parameters) and the ability in relation to latent variable (construct). This probabilistic relationship is defined by the characteristic curve of the item (CCI), which according to Chernyshenko, Stark, Chan, Drasgow, & Williams (2001) is a logistic function of the probability of a response be ticked. At the extremes of the JRC verifies that an individual with ability equal to 3.0 will have approximately 100% probability of hit while another score-3.0 virtually has no chance to correctly answer the question.

Although it is used with success in education, for example, in the calculation of the national high school examination (Andrade & Klein, 2005), your use in academic research is still very restricted. There are however applications of IRT for reliability, as in Lucian and de Silva (2015).

To understand the IRT one must understand initially that all estimates are on the item and not about the sample and that concepts such as probability sampling are secondary. The important thing in this technique is the behavior of the item, the independent group that is being tested. Therefore, item analyses are made based on the latent variable that is independent of the sample-specific behavior (Matteucci, Mignani, & Veldkamp, 2012).

It is called Item response theory to a set of mathematical models that seek to represent the probability of an individual have a right answer to an item, as a function of the parameters of the item and the ability of the respondents.

Selectable models differ by number of test parameters and the type of the variable. As for the number of parameters, they can be a parameter (only the difficulty of the item), two parameters (the difficulty and discrimination) or three parameters (difficulty, discrimination and chance to hit at random). Already in relation to the type of variable, or nominal reason (Lucian & de Silva, 2015).

The difficulty of the item is represented by the letter *b* and can vary between-4.0 to + 4.0 and approaching zero results indicate difficulty average. At this point, it is important to emphasize that any one of the three models can measure the skill (Θ) that is expressed on the same scale of difficulty of the item, i.e. it is represented by the *x* axis on the graph of CCI.

The second parameter, present only in the two or three parameters is the breakdown of the item represented by the letter *a* and is the most important score for measuring reliability of scales. This score has your usual variability between 0 and + 3.0 and higher, more reactive, IE has greater ability to detect small variations in the ability of the respondents. There is a possibility *to* return negative values if the probability of hit and Θ are inversely proportional, if that indicates reversal of scale items or problems of formulation on the issue. In CCI *the* parameter is calculated at the point of inflection (Bacci, 2012).

The third parameter is represented by the letter *c* and indicates the chance to hit at random, i.e., *c* the number of search hits measured by people of very low skill. No absolute parameters for the chance to hit at random, must observe your value in relation to the calculation 1/*n*, where *n* is the number of

alternatives. If the scores of *c* may increase a lot of expected this indicates problems in the construction of the question.

For the calculation of reliability, interest is by the model of two parameters, because only *the* score is taken into consideration. The IRT, however, is a recent technique and with the advancement of the interest of researchers for their benefits, other parameters may have its benefits revealed.

There is yet another model item response theory that can be adopted in the study, originally proposed by Bock (1972), it makes use of dichotomous variables and is associated with the treatment of nominal scales. Thus, adapts perfectly to the reliability analysis of Likert-type scales itself, in which there are two options: positive and negative attitude.

The nominal IRT is based on logistics and distribution function of the normal curve (*sigma*) and, due to your character of item orientation, does not require any effort on sampling (Pasquali & Primi, 2003). The important thing is to measure the amount of item information, as noted in Bernardi, Bussab and Camargo (2009).

In practice, this value is not given directly by the *software*, because the IRT is not specific for calculation of reliability. Instead, calculate the amount of information for each point of the distribution of the latent variable and displays the results in graphic form. The calculation of the chart area estimates if there's enough information or not to consider the item reliable, i.e. an estimated reliability based on the value of *the* parameter.

The defaults for *the*, as already mentioned, ranging from 0 to + 3.0 (and may take negative values in specified situations previously), the null value for when there is no discrimination and 3 for perfect discrimination. The higher the value of *a* means that the greater the likelihood of success of the respondents that have greater Ɵ which in this case represents the presence of latent variable.

When the goal is the calculation of reliability, what is sought is the estimation of item discrimination and in relation to the desired values for *a* parameter, if your value is less than 0.85 there will be enough information to consider the reliable item. There is also a second track of trusted values when *the* is greater than 1.7; Therefore, we can say that the item is reliable if the discrimination parameter does not have values between 0.85 and 1.70 (Lucian & de Silva, 2015).

The composition of such values in two tracks is unusual in CTT and, at first, can confuse the user of the IRT, however explains that a scale is said to be confident when your items can discriminate the presence of latent variable in the sample, thus, the intermediate values have less representation than the extremes.

Clarifying the concept exemplifies that, in a hypothetical scale for measuring user interaction with intelligent systems, it is expected that the scale discriminates individuals who have very little interaction or that have really a lot of interaction, so if you know that the cut was rigorous and the results are reliable.

The main advantage of the IRT in determining the reliability of a scale is that it assumes the heterogeneity of the contribution of each information item to measure the scale, because it assumes a role of information for each item (Lord, 1980). Thus, the calculation need IRT cancels traditional reliability, as the estimate Cronbach's alpha (Zagorsek, Stough, & Jaklic, 2006).

The requirements for use of the IRT in broad sense are that the items being one-dimensional arrays (each scale measured just a construct), have features of independent events (low internal correlation between items) and monotones (probability proportional hit between items and in one sense).

Due to the overlapping feature of probability in Likert-type scales (not monotonic because the answers can float freely between the extremes), the calculation of the *c* parameter is not suitable (Gutierrez, 2005), but the breakdown of the item indicated by *the* coefficient is independent to the probability distribution and meets the purposes of the calculation of the reliability (Bernardi, Bussab, & Camargo, 2009), so it is suggested to be adopted the model of two parameters for such a purpose.

5 FINAL CONSIDERATIONS

On the benefits provided by Item response theory on the classical theory of the tests submitted, on your use for the calculation of the reliability of scales to the detriment of Cronbach's alpha Coefficient,

we recommend the adoption of the IRT as a statistical tool for purification of scales and reliability test.

Responding finally to the objective of the study, says that the Item response theory is a promising technique that already has policies of immediate application as safe and unique tool for calculating the reliability of scales. Therefore, it is believed that your adoption for such a purpose is a natural evolution in relation to Cronbach's alpha Coefficient like this was in relation to CAM technique.

The main resistance, however, the popularization of IRT is certainly the complexity of the calculations, including the lack of intuitiveness of specific *software*. There are a few options that can be purchased for Windows operating systems as the ConQuest 3, Facets, RUMM2030, WINMIRA, Winsteps and Xcalibre besides the option for Mac Quest.

Some offer versions for study or are completely free of charge as the Bigsteps, ConstructMap, Facets, Ganz Rasch, ICL, jMetrik, Minifac, MULTIRA and WinLLTM. Such programs, however, are not intuitive and some even require the user to enter lines of code to perform the calculations. As far as can be seen, none of them is able to import files from other spreadsheets, being mandatory conversion to TXT or CSV.

Considering the usability, the import and export tools and the diversity of included templates (one-dimensional, multidimensional, scalar, nominal, dichotomic or politômicos), the best program for calculation of Item response theory seems to be the IRTPRO that has paid versions and free versions for study with some limitations.

It is recommended, so that future interested in dealing with data and purify measurement scales use the IRT as a tool for calculating the reliability of the instruments. The benefit of this practice will transcend research practices and collaborate with more accurate results that ultimately will benefit the professionals and executives in their managerial practices that depend on the advancement of research in the area.

REFERENCES

Andrade, D. F., & Klein, R. (2005). Aspectos quantitativos da análise dos itens da prova do Enem. In Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. *Exame Nacional do Ensino Médio (Enem): Fundamentação teórico-metodológica*, Brasília: O Instituto.

Bacci, S. (2012). Longitudinal data: Different approaches in the context of item-response theory models. *Journal of Applied Statistics*, *39*(9), 2047-2065. doi:10.1186/2196-0739-2-1

Bernardi, P. Júnior, Bussab, W. de O., & Camargo, R. A. (2009). Análise da Confiabilidade do Índice de Predisposição para a Tecnologia na Estrutura da Teoria de Resposta ao Item. *Anais do XXXIII Encontro da ANPAD.* São Paulo: ANPAD.

Bimbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord, & M. R. Novick (Eds.). *Statistical theories of mental test scores* (pp. 397-472), Reading, MA: Addison-Wesley.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, *37*(1), 29-51. doi: 10.1007/BF02291411

Caruso, J. C. (2000). Reliability generalization of the NEO personality scales. *Educational and Psychological Measurement*, *60*(2), 236-254. doi: 10.1177/00131640021970484

Chernyshenko, O. S., Stark, S., Chan, K. Y., Drasgow, F., & Williams, B. (2001). Fitting item response theory models to two personality inventories: Issues and insights. *Multivariate Behavioral Research, 36*(4), 523-562. doi: 10.1207/S15327906MBR3604_03

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology*, *78*(1), 98-104.

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

Crowther, J. (1995) *Oxford advanced learner's dictionary*. Oxford: Oxford University Press.

Drasgow, F., Levine, M. V., Tsien, S., Williams, B., & Mead, A. D. (1995). Fitting polytomous item response theory models to multiple choice tests. *Applied Psychological Measurement, 19*(2), 143-165. doi: 10.1177/014662169501900203

Duhachek, A., & Iacobucci, D. (2004). Alpha's standard error (ASE): An accurate and precise confidence interval estimate. *Journal of Applied Psychology, 89*(5), 792-808.

Dutra, L. H. de A. (2010). *Introdução à epistemologia.* São Paulo: Editora UNESP.

Gouveia, V. V., Santos, W. S., & Milfont, T. L. (2013). O uso da estatística na avaliação psicológica: Comentários e considerações práticas. In C. S. Hurtz (Org). *Avanços e polêmicas em avaliação psicológica*. São Paulo: Casapsi Livraria e Editora Ltda.

Gutierrez, G. C. (2005). *Estimação das escalas dos construtos capital social, capital cultural e capital econômico e análise do efeito escola nos dados do Peru-PISA 2000* (Dissertação de Mestrado) - Pontifícia Universidade Católica do Rio de Janeiro, 200p. Rio de Janeiro.

Henson, R. K. (2001). Undestanding internal consistency reliability estimates: A conceptual primer on coeficiente alpha. *Measurement and Evaluation in Counseling and Development*, *34*(3), 177-189.

Henson, R. K., Kogan, L. R., & Vacha-Haase, T. A. (2001). Reliability generalization study of the Teacher Efficacy Scale and related instruments. *Educational and Psychological Measurement*, *61*(3), 404-420. doi: 10.1177/00131640121971284

Lord, F. M. (1952). *A theory of test scores* (Psychometric Monograph, 7). Iowa City, IA: Psychometric Society.

Lord, F. M. (1980). *Application of item response theory to practical testing problems*. Hillsdale: Lawrence Erlbaum Associates.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading: Addison-Wesley.

Lucian, R., & Dornelas, J. S. (2015), Mensuração de atitude: Proposição de um protocolo de elaboração de escalas. *RAC. Revista de Administração Contemporânea (Online), 19*(1), 157-177.

Malhotra, N. (2013). *Review of marketing research*. Emerald Group Publishing Limited. Bingley.

Maroco, J., & Garcia-Marques, T. (2006). Qual a fiabilidade do Alfa de Cronbach? Questões antigas e soluções modernas? *Laboratório de Psicologia*, *4*(1), 65-90.

Matteucci, M., Mignani, S., & Veldkamp, B. P. (2012). The use of predicted values for item parameters in item response theory models: An application in intelligence tests. *Journal of Applied Statistics*, *39*(12), 2665-2683. doi: 10.6092/unibo/amsacta/3241

Pasquali, L., & Primi, R. (2003). *Fundamentos da Teoria de Resposta ao Item – TRI*. Avaliação Psicológica, 2(2), 99-110.

Pooja, S., & Sagar, M. (2012). High impact scales in marketing: A mathematical equation for evaluating the impact of popular scales. *Advances in Management*, *5*(4), 31-48.

Robertson, J. (2012). Likert-type scales, statistical methods, and effect sizes. *Communications of the ACM*, *55*(5), 6-7. doi: 10.1145/2160718.2160721

Sanches, C., Meireles, M., & Sordi, J. O. de. (2011, agosto). Análise qualitativa por meio da lógica paraconsciente: Método de interpretação e síntese de informação obtida por escalas Likert. *Anais do Encontro de Ensino e Pesquisa em Administração e Contabilidade*, João Pessoa, PB, Brasil.

Schultz, C., Salomo, S., & Talke, K. (2013). Measuring new product portfolio innovativeness: How differences in scale width and evaluator perspectives affect its relationship with performance. *Journal of Product Innovation Management*. *30*(2), 93-109. doi: 10.1111/jpim.12073

Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika*, *74*(1), 107-120. doi: 10.1007/s11336-008-9101-0

Streiner, D. L. (2003). Starting at the beginning: An introduction to Coefficient Alpha and internal consistency. *Journal of Personality Assessment*, *80*(1), 99-103. doi: 10.1207/S15327752JPA8001_18

Ten Berge, J. M. F., & Socan, G. (2004). The greatest lower bound to the reliability of a test and hypothesis of unidimensionality. *Psychometrika*, *69*(4), 613-625. doi: 10.1007/BF02289858

Thompson, B. (2002). *Contemporary thinking on reliability issues*. Newbury Park: Sage.

Thompson, B. (1994). Guidelines for authors. *Educational and Psychological Measurement*, *54*(1), 837-847.

Vacha-Haase, T. (1998). Reliability generalization: Exploring variance in measurement error affecting score reliability across studies. *Educational and Psychological Measurement*, *58*(1), 6-20. doi: 10.1177/0013164498058001002

Vieira, K. M., & Dalmoro, M. (2008). Dilemas na construção de escalas tipo Likert: O número de itens e a disposição influenciam nos resultados? *Anais do XXXII Enanpad*. Rio de Janeiro.

Wilkinson, L. (1999). Task force on statistical inference, APA Board of Scientific Affairs. Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, *54*(8), 594-604.

Yin, P., & Fan, X. (2000). Assessing the reliability of beck depression inventory scores: Reliability generalization across studies. *Educational and Psychological Measurement*, *60*(2), 201-223. doi: 10.1177/00131640021970466

Zagorsek, H., Stough, S., & Jaklic, M. (2006). Analysis of the reliability of the leadership practices inventory in the item response theory framework. *International Journal of Selection and Assessment*, *14*(2), 180–191. doi: 10.1111/j.1468-2389.2006.00343.x