ISSNe: 2182.2883 | ISSNp: 0874.0283
Available: http://dx.doi.org/10.12707/RIV15003

THEORETICAL PAPER/ESSAY

Contributions to the Discussion on the Assessment of the Reliability of a Measurement Instrument

Contributo para a Discussão da Avaliação da Fiabilidade de um Instrumento de Medição
Contribución a la Discusión sobre la Evaluación de la Fiabilidad de un Instrumento de Medición

Fernanda Daniel*; Alexandre Gomes da Silva**; Pedro Lopes Ferreira***

Abstract

Theoretical framework: The reliability of an instrument refers to the consistency of the results obtained during its administration. Nowadays, reliability assessment is almost imperative in the presentation of empirical data. Depending on the purpose of the presentation and the concept measured, the study of reliability can include multiple procedures.
Objectives: To map the main techniques for measuring reliability and their algorithms.
Main topics of analysis: The test-retest and equivalent forms are used to measure temporal stability. Cronbach's alpha, Kuder-Richardson and Spearman-Brown split-half models are used for measuring internal consistency.
Conclusion: Since the study of reliability of a scale is directly related to what we intend to measure or compare, it is impossible to postulate standards for the use of the above-mentioned estimators. For that reason, we propose a careful consideration in choosing the most suitable estimator for the study to be developed.

Keywords: Reproducibility of the tests; questionnaires; psychometrics

Resumo

Enquadramento: A fiabilidade de um determinado instrumento reporta-se à consistência dos resultados obtidos aquando da sua administração. A inspeção da fiabilidade de um instrumento assume, nos dias de hoje, caráter quase imperativo na apresentação dos dados empíricos. Dependendo do propósito da apresentação e do conceito medido, o estudo da fiabilidade pode incluir vários procedimentos.
Objetivos: Mapear as principais técnicas de medição da fiabilidade e os seus algoritmos.
Principais tópicos em análise: Para a medição da estabilidade temporal apresentamos o *teste-reteste* e as formas equivalentes. Na medição da consistência interna são apresentados os modelos alfa de Cronbach, Kuder-Richardson e da bipartição de Spearman-Brown.
Conclusão: Dado que o estudo de fiabilidade de uma escala está diretamente relacionado com o que se pretende medir ou comparar é impossível postular receitas padrão para o uso dos referidos estimadores. Por esse facto propomos uma ponderação cuidada na escolha do mais adequado ao estudo em causa.

Palavras-chave: Reprodutibilidade dos testes; questionários; psicometria

Resumen

Marco contextual: La fiabilidad de un instrumento se refiere a la consistencia de los resultados obtenidos durante su administración. La inspección de la fiabilidad de un instrumento conlleva, en estos días, un carácter casi imperativo en la presentación de los datos empíricos. Dependiendo de la finalidad de la presentación y del concepto medido, el estudio de la fiabilidad puede incluir varios procedimientos.
Objetivos: Mapear las principales técnicas de medición de la fiabilidad y sus algoritmos.
Principales temas de análisis: Para la medición de la estabilidad temporal presentamos el test-retest y las formas equivalentes. En la medición de la consistencia interna se presentan los modelos alfa de Cronbach, Kuder-Richardson y la bipartición de Spearman-Brown.
Conclusión: Dado que el estudio de la fiabilidad de una escala está directamente relacionado con lo que se pretende medir o comparar, es imposible postular recetas estándar para el uso de los estimadores. Por eso, proponemos una ponderación cuidada en la elección del más adecuado para el estudio en cuestión.

Palabras clave: Reproducibilidad de resultados; cuestionarios; psicometría

* Ph.D., Auxiliary Professor, Miguel Torga Institute of Higher Education and University of Coimbra Centre for Health Studies and Research, 3000-370, Coimbra [fernanda-daniel@ismt.pt]. Contribution to the article: Design, elaboration and approval of the final version.
** Ph.D., Coordinating Professor, Coimbra Business School - ISCAC/IPC, 3040-316, Coimbra, Portugal [alexmfgs@gmail.com]. Contribution to the article: Writing and revision of the algorithms.
*** Ph.D., Associate Professor, University of Coimbra Centre for Health Studies and Research, 3000-370, Coimbra [pedrof@fe.uc.pt]. Contribution to the article: Revision of the final version.

Received for publication: 07.01.15
Accepted for publication: 16.10.15

Introduction

Measuring means to establish an agreement between the method used to measure and what is being measured. This process is relatively easy to perform with directly observable variables. However, with latent variables (not directly observed or measured), this process becomes more difficult, as it is the case of measuring anxiety, satisfaction or quality of life using scales with multiple items.

There is currently a wide range of measurement instruments that can be called tests, scales or inventories "according to the presumed relationship between the items, or questions that make up the questionnaire" (Ribeiro, 1999, p. 76). Every year, the number of measurement instruments built in different countries increases, which demands their validation and the performance of reliability tests. In this article, we will focus on reliability.

Until very recently, the analysis of reliability was a lengthy process, hard to perform if we take as reference the short deadlines established for contemporary academic research studies. With the advances in computer technology and the easier access to statistical software, these processes have become increasingly easier to perform because it is now possible to perform complex calculations almost immediately, which was something unthinkable a few decades ago.

A measurement instrument must be unequivocally reliable. The reliability of a measurement instrument ultimately relies on the quality of its items. To obtain a reliable instrument, one may have to consider deleting, replacing or reviewing the items.

The term *reliability* intends to express the quality of what or who is reliable. In the context of analysis, the term has specificities that should be discussed. Reliability refers to the consistency of the results obtained by the same individuals, when questioned in different moments or at a given moment, and, in the latter, reliability is determined based on equivalent items (Anastasi, 1977; Freeman, 1980). Scientific language has a wide range of terms for designating what we here refer to as reliability. The Portuguese translation of some terms used in reference works within this field includes *garantia* (Freeman, 1980), *precisão* (Anastasi, 1977), or even *fidelidade* (Bryman & Cramer, 1993). Authors writing in Portuguese use the terms *fiabilidade* (Hill & Hill, 2012; Maroco & Garcia-Marques, 2006) and *fidedignidade* (Vaz-Serra, Ponciano, & Freitas, 1980).

The estimation of the reliability of a test is a key procedure when we use measurement instruments with scales composed of several items. In the 1990s and in the beginning of the 21st century, the estimation of the reliability of a test, scale or inventory became a quick procedure, as a result of the increase in the number of statistical packages which, in association with increasingly more interactive and user-friendly platforms, allowed for a massive use of these procedures.

Reliability also presupposes the reproducibility of results. Anne Anastasi, a scientist whose brilliance is expressed in a vast work where psychometrics is deconstructed, argues that accuracy refers to "the consistency of scores obtained by the same individual on different occasions or with different sets of equivalent items" (1977, p. 84). In the same line, Freeman also says that reliability is reflected in the consistency of the results performed in several evaluations, meaning that "the degree in which the results obtained are exempt from those internal defects susceptible of causing measurement errors inherent to own items and the standardisation" (1980, pp.73-74).

Development

There are various estimates of the reliability of a test, and "there could, of course, be as many varieties of test reliability as there are conditions affecting test scores, because any of these conditions may have no meaning for a particular purpose and, thus, be considered as error variance" (Anastasi, 1977, p. 85). Depending on the method used to estimate reliability, various procedures are used. The assumption underlying the study of reliability is to calculate the error size. Thus, if there was a lack of variability in the results, there would be no error and reliability would be equal to 1. It should be noted that the definition of error is complex because it is associated with the concept being measured. If the test measures depression, the difference found in the results obtained between both administrations (error variance) may be attributed both to random fluctuations (and in this case we talk about a random error), and to a different emotional status (and, in this case, we cannot refer to it as a random error). If

we intend to measure the temporal stability of a test, we must ensure that there are no differences in the emotional status between both administrations. This procedure can be controlled through the inclusion of questions on significant intermediate experiences between administrations.

Temporal stability

An instrument is considered to have temporal stability if its results remain constant over time, i.e., "performing the same test twice, the correlation between the results gives us an indication about the stability of the results over time" (Laveault & Grégoire, 2002, p. 150). The procedures to assess temporal stability or external reliability imply the administration of two versions of the same test (test-retest) or two versions of equivalent tests.

Test-retest

The assessment of the temporal stability of a set of questions or equivalent items requires the use of a test-retest. To estimate this temporal stability, it is necessary to administer the set of questions or equivalent items to a group and to correlate the administrations of the obtained scores with those obtained in the other administration performed at another moment. Despite the time elapsed between both administrations, if the test is reliable and if no significant changes have occurred in the meantime, the answers obtained for each individual should be the same or almost the same.

The concept underlying the test-retest reliability may be extrapolated to our daily life and to physical measures. Let us imagine then that the same person was weighted twice, one time at 12 a.m. and another time at 1 p.m. What should we think if the weight increased by one kilogram between both measurements? One of the reasons for this increase may be that the scale is not well calibrated. Let us place ourselves now in the role of a researcher who has administered two versions of a test and let us imagine that his/her score doubled between both administrations. What should we think? In the case of the scale, after the initial surprise, one will probably raise the hypothesis of excessive food intake. The same is true for the result obtained in our set of questions or equivalent items: The test is either not reliable or something happened between both administrations that can explain the difference in the results.

The differences found in the tests can thus be attributed to various factors: (i) administrator, (ii) administered, or (iii) context. In relation to the *administrator* factor, let us imagine that, in one of the test administrations, the respondent perceived one of his/her behaviours as inappropriate. Will this behaviour negatively affect the results? May it contribute to an inaccurate, unprofessional or even dishonest completion of the test, scale or inventory, thus interfering with the results? On the other hand, let us imagine that the administrator has shown excessive sympathy alongside a long and inappropriate narrative about the measuring instrument. May these behaviours influence the respondent and trigger responses in conformity with the social standards? It is recognised that "the desire for social acceptance can bias the parameters assessed in scientific studies, constituting a threat to their validity, and therefore must be controlled" (Poínhos et al., 2008, p. 223).

As regards the *administered* factor, let us imagine one of multiple possibilities: In the period of time between both administrations, some of the respondents became unemployed. This new situation of forced inactivity is experienced by some of the respondents with high levels of anxiety. If the concept to be measured is state-anxiety (emotional state or transient condition of the human organism characterised by unpleasant feelings of tension and apprehension consciously perceived, and by an increase in the activity of the autonomic nervous system; Spielberger, 1983), the identified differences should not be assigned to a low test-retest reliability, but rather to the changes in the state-anxiety levels as a result of the new situation. In this case, we would have a test that is reliable, despite presenting apparent low test-retest reliability. It is relevant to mention that the reliability of the measurement instruments that assess mood states can show a significant variation, i.e. significant differences among the scores are likely to occur. In contrast, the scores of measurement instruments assessing personality characteristics should remain stable over time.

Finally, in relation to the *scenario* factor, let us imagine that the room where the set of questions or equivalent items is being administered is poorly insulated and cold, or even that the outside noise causes discomfort. In this situation, the results are expected to reflect the environmental setting being experienced.

Regarding this, we should note that, when using a measurement instrument, we should always use

its manual or original article or, if it does not exist, contact the original authors to obtain information on the necessary period of time between both administrations. This information is valuable because the period of time between administrations can have different effects depending on the concept to be measured. Short intervals may cause memory-related effects. On the contrary, long intervals may allow for the acquisition of new knowledge. Regarding this issue, Anastasi mentions that:

> Examples of tests with high accuracy in periods of a few days or a week may easily be cited, but their scores reveal an almost complete lack of correspondence when the interval is extended to as long as ten or fifteen years. For example, many intelligence tests for preschool-aged children provide moderately stable measures for the pre-school period, but are virtually useless to predict the IQ in school age or adulthood. (1977, p. 32)

The subjects' age is equally important. Anastasi suggests that the test-retest interval should be shorter in children, because the developmental changes in children are visible in short periods (one month or less) and that, regardless of age, test-retest intervals should not exceed six months (1977). The same can apply to reliability tests concerning health conditions that may progress in a very different way.

Furthermore, it should also be mentioned that the interpretation of the test-retest reliability score should take into account the fact that a test is followed by another test. The experience of the first administration may affect the performance of the subject in the second administration.

We should not expect, even in a test measuring a stable personality trait, a test to show perfect test-retest reliability with a correlation equal to 1, since there are several factors that influence the scores. For example, fatigue, different levels of concentration and motivation, different environmental conditions (temperature, noises, and environmental distractions), the effect of practice and learning, the interval between administrations, unexpected personal events in this time period, and the errors inherent to the administration are some of the possible variations that interfere with error variance.

When a measurement instrument is operationalised on a composite scale of multiple items (e.g., through the sum or means of its items), Pearson's linear correlation coefficient can be used to estimate inter-form reliability. One of the formulas used to calculate this correlation coefficient *r* is shown below:

$$r = \frac{n \times \sum XiYi - \sum Xi \times \sum Yi}{\sqrt{\left[n \times \sum Xi^2 - \left(\sum Xi\right)^2\right] \times \left[n \times \sum Yi^2 - \left(\sum Yi\right)^2\right]}}$$

where x and y represent the variables.

For example: Le us imagine that we have administered a measurement instrument composed of four items that assesses the work routine of the workers of a woollen factory. After a while, we applied the same instrument again to the same sample. Table 1 describes the procedure that should be performed to calculate temporal stability through Pearson's linear correlation coefficient.

Table 1
Procedure to calculate Pearson's correlation coefficient

No.	Routine (1st Moment)					Routine (2nd Moment)							
	It. 1	It. 2	It. 3	It. 4	X_i	It. 1	It. 2	It. 3	It. 4	Y_i	X_i^2	Y_i^2	X_i*Y_i
1	4	2	4	5	15	4	3	4	5	16	225	256	240
2	2	1	2	2	7	2	2	2	2	8	49	64	56
3	3	2	4	5	14	3	2	4	5	14	196	196	196
4	2	1	3	3	9	2	1	3	4	10	81	100	90
5	2	2	2	3	9	2	2	2	3	9	81	81	81
6	2	1	2	1	6	2	2	2	2	8	36	64	48
7	3	4	2	4	13	3	4	2	4	13	169	169	169
8	2	1	3	1	7	2	2	3	1	8	49	64	56
9	2	1	3	2	8	2	2	3	1	8	64	64	64
10	2	1	2	1	6	1	2	2	1	6	36	36	36
∑	24	16	27	27	94	23	22	27	28	100	986	1094	1036

Note. No. = Number assigned to the respondent; it. = item; ∑= sum; X_i and Y_i = composite variables (X_i = moment 1 and Y_i = moment 2); X_i^2 and Y_i^2 = squared composite variables; X_i*Y_i = products of the composite variables

A Pearson's correlation score of 0.978

$$r = \frac{10 \times 1036 - 94 \times 100}{\sqrt{\left[10 \times 986 - (94)^2\right] \times \left[10 \times 1094 - (100)^2\right]}} = 0{,}978$$

is considered significant at the level of 0.01, which means that there is no higher probability than one in 100 that the correlation in the population is zero. It should be noted that the sample size and its variability can affect all the measures presented, i.e. a correlation score can be considered significant in a large sample and not in a smaller sample. Finally, we can state that, in studies where data are collected through tests already validated for the Portuguese population, this type of estimate is not, from our perspective, an essential procedure.

Equivalent forms

According to Anastasi, the reliability coefficient is "a measure of both temporal stability and consistency of response to different item samples (or test forms)" (1977, p. 95).

The so-called method of equivalent, alternate or parallel forms is similar to the test-retest method in so far as the same test is applied twice. However, there are considerable differences between both methods. While in the test-retest method the same versions are applied in both sessions, the method of equivalent forms uses versions which are equivalent, i.e. the versions must be equal in terms of instructions, form and in all the other characteristics, and similar in content. This method aims to eliminate two types of biases found in the test-retest method: The fact that the individuals can recall the previous test and the possible effects of the practice. This method thus seeks to eliminate the effect of the practice and memory by testing individuals through the use of comparable but not equal versions in both sessions.

High reliability in equivalent forms suggests that the items of both versions of the test are representative of the same population of items that hypothetically represent the concept that is to be measured. Low reliability in equivalent forms suggests that both formats of the test are not measuring the same thing. It should be noted that reliability through a test of equivalent format includes many of the same limitations of the test-retest reliability.

When measuring the test-retest reliability and the equivalent forms, the length of the interval between both administrations must always be mentioned. If both forms are administered one after the other, in immediate succession, the correlation only shows reliability across forms (content) and not across occasions (temporal stability). Based on the characteristics presented above, we believe that this form is not frequently used in research.

Internal consistency

As regards internal consistency, which is also called internal reliability by Bryman and Cramer (1993), only one of the versions is applied and only on one occasion. Let us imagine that we want to assess a latent variable based on multiple items. As expected, the group of items must operationalise the latent variable and not another variable. Do the items have internal consistency? It is only possible to state that a measurement instrument has internal consistency if all its items contribute for the measurement of the same characteristic. The procedure of internal consistency to estimate reliability is currently one of the most used methods in research, in so far as it is a method that, in addition to be cost-effective (Polit & Hungler, 1992) for requiring only a single test, it is also the best method to assess one of the most important sources of measurement errors that is the selection of test items. Cronbach's alpha coefficient is commonly used to estimate the reliability of instruments in which the items have multiple answers. The basic rule here is also that the values should range between 0.8 and 1.0 (Bryman & Cramer, 1993). When a concept and its measurement comprise several dimensions, the reliability coefficients for each of the underlying dimensions is usually estimated, instead of estimating a single one for the measure as a whole.

The following formula is used to calculate the alpha coefficient:

$$\alpha = \frac{k}{k-1}\left[1 - \sum_{i=1}^{k} \frac{s_i^2}{s_t^2}\right]$$

where K is the number of items in the scale, s_i^2 is the variance of the results of the test on item I and s_t^2 is the test variance.

The process of measuring the internal consistency through Cronbach's alpha coefficient produces low estimates of test reliability, but overestimates the reliability of speed tests. Consequently, the procedures of internal consistency are considered inappropriate to determine the reliability of speed tests (Ary, Jacobs, & Razavieh, 1990).

Although there is consensus that a scale should be reliable, the values from which the reliability of a scale is inferred are not consensual, as shown in Table 2, adapted from Peterson (1994).

Table 2
Values proposed by several authors on the levels recommended for Cronbach's α

Author	Situation	Recommended levels
Davis (1994)	Prediction for individual	Above 0.75
	Predictor for group of 25-50	0.5
	Predictor for group over 50	Below 0.5
Kaplan and Sacuzzo (1982)	Basic research	0.7-0.8
	Applied research	0.95
Murphy and Davidsholder (1988)	Unacceptable level	Below 0.6
	Low level	0.7
	Moderate to high level	0.8-0.9
	High level	0.9
Nunnally (1978)	Preliminary research	0.7
	Basic research	0.8
	Applied research	0.9-0.95

Note. Adapted from "A meta-analysis of Cronbach's coefficient alpha", de R. A. Peterson, 1994, *Journal of Consumer Research*, 21(2), p. 382.

This coefficient has raised some criticism for having some weaknesses, which led authors such as Maroco and Garcia-Marques (2006) to present alternatives. The composite reliability defined by Fornell and Larcker (1981) for a factor j with *k* items is given by

$$fc = \frac{\left(\sum_{i=1}^{k} p_i\right)^2}{\left(\sum_{i=1}^{k} p_i\right)^2 + \sum_{i=1}^{k} e_i}$$

where *p* represents the factor loadings of each item and *e* represents the error.

Whereas the Cronbach's alpha coefficient is used in tests with items with multiple answers, the *KR20* formula is intended for tests in which items only offer two alternative hypotheses, as is the case of true/false, yes/no, or correct/wrong:

$$R_{kR-20} = \frac{K}{K-1}\left[1 - \frac{\sum_{i=1}^{K} p_i q_i}{\sigma_X^2}\right]$$

where *K* is the number of items, *p* is the proportion of correct answers, *q* is the proportion of incorrect answers, and is the variance.

Split-Half

Another procedure used to measure if a set of questions or equivalent items is reliable is designated as split-half, also known as reliability of the halves or of the bipartition. It is possible to arrive at a measure of reliability from a single administration. In this case, the two halves of a test are correlated. The split-half should be based on criteria which take into account both the concept and how it is operationalised. A variant of this process is known as the even-odd method, which is perhaps the oldest method to estimate internal consistency (Polit & Hungler, 1992). The split-half reliability is similar to the alternate way, splitting a single scale in two. This method assesses the level of inter-item consistency, determining the internal consistency of the scale. It does not measure temporal stability, but offers the advantage of making it possible to obtain a reliability measure from a single administration and assumes that all items equally contribute to the measurement of a key concept (Anastasi, 1977; Freeman, 1980).

Let us imagine a 10-item scale. One of the ways to divide the scale would be to decompose it according to the numbering which was initially allocated to it, i.e. one of the halves corresponds to the even items, whereas the other is composed of the odd items. Another way is to divide the items in two major halves, i.e. the first five items, from the first half, and the remaining five items, from the second half. It should be noted that the choice of division should

be related to the type of test. The division in two major halves, the first half and the second half, can have many disadvantages, particularly in tests with an increasing level of difficulty. The halves can also be created through a random selection.

Any correlation coefficient obtained through the split-half technique tends to generate a systematically lower estimation than the total scale, i.e. the correlation obtained refers to half of the test. The scales with more items generate higher reliability values (Polit & Hungler, 1992). With the aim of overcoming this difficulty, a formula was created to adjust the correlation coefficient to the total scale. Thus, a coefficient that can be interpreted in the same way as Pearson's correlation coefficient is obtained, to the extent that it varies from 0 to 1. Ideally, it should be higher than or equal to 0.8 (Bryman & Cramer, 1993) and is obtained through the following formula:

$$r = \frac{\frac{1}{2}\left(s_p^2 - s_{p1}^2 - s_{p2}^2\right)}{s_{P1}s_{P2}}$$

where s_p, s_{p1} and s_{p2} represent the values of the standard deviations for all items and for each one of the halves, respectively.

On the other hand, by determining the correlation value between both halves, it is possible to estimate the correlation coefficient for the total scale. The correction equation is called Spearman-Brown formula and is expressed in the following algorithm:

$$r_{cSB} = \frac{2r_{xy}}{1 + r_{xy}}$$

In this formula, represents the correction introduced by Spearman-Brown and represents the correlation between both halves. For example, if the correlation coefficient between both halves of the test was 0.65, the estimate for the total scale would be as follows:

$$r_{cSB} = \frac{2 \times 0{,}65}{1 + 0{,}65} = 0{,}79$$

In this context, we can also estimate the effects in the reliability of a certain scale when the number of items increases or decreases through the following formula:

$$\frac{nr}{1 + (n-1)r}$$

where n is the proportion of the number of items in each form and r is the correlation between both halves. Thus, if a scale has 60 items, if the r value obtained is equal to 0.5, and if we are interested in increasing the number of items to 150, then the proportion of the increase will be 2.5 (150/60). Through this formula, the reliability estimate will decrease from 0.71 to 0.50:

$$\frac{2{,}5 \times 0{,}5}{1 + (2{,}5 - 1) \times 0{,}5} = 0{,}71$$

If both parts are not balanced, the following algorithm is used:

$$ULD = \frac{-r^2 + \sqrt{r^4 + 4r^2(1 - Rr^2)k_1k_2/k^2}}{2(1 - R^2)k_1k_2/k^2}$$

where r is the correlation obtained above and k is the number of items of each part (UL, *Unequal Length*) (IBM® SPSS®, 2011).

The fact that there are many ways of dividing the items into two groups leads to the possibility of obtaining multiple reliability estimates. For this reason, Cronbach's alpha coefficient and the Kuder-Richardson formula are usually the only methods used because they express the mean of all possible bipartitions.

Analysis of the sensitivity of the scale items

Several procedures can be used to assess the sensitivity of the scale items, such as the analysis of the inter-item correlation matrix, Cronbach's alpha value if each one of the items was deleted, and the item-total correlation.

The matrix of correlations shows the correlation of each item with all the other items. In the diagonal of the matrix, we should find the value 1, since the correlation of an item with itself will always be equal to the unit. The values of the other correlations should be high and positive, indicating that the items measure the same latent variable. If there are negative values, we should check if the items in question were in the same conceptual direction as the other items and, if this is the case, we should recode them.

It is always possible to check, as a sensitivity analysis, the impact of each item on the global model by assessing the alpha without that item or the correlation of that item with the sum of the other items. If there is no impact on the scale, the simpler assessment is preferable. Two measures that show this are the item-total alpha and the item-total correlation, as can be seen in the following formulae.

The analysis of Cronbach's alpha value, when a specific item is deleted, allows us to analyse the impact of deleting a certain item. This is provided by

the formula

$$\bar{A}_i = \frac{k-1}{k-2}\left(1 - \sum_{\substack{l=1 \\ l\neq i}}^{k} S_l^2 / \tilde{S}_i^2\right)$$

where k is the number of items, s_l^2 is the variance of the results of the test in item I, and $\widetilde{s}_i^2$ is the variance of the test without the item.

If, from the analysis of the elimination of certain scale items, Cronbach's alpha values are higher than the alpha value of the total scale, we should consider the elimination of those items from the scale. It should be mentioned that Cronbach's alpha values lower than 0.7 are not favourable, hence the possibility of elimination of the item should be equated (Pallant, 2007). Additionally, scale items should only be deleted when the scale is being created or to reduce the number of scale items. Otherwise, the deletion procedure inhibits, strictly speaking, any comparison between our results and the original scale.

Another way to obtain information about the internal consistency of a scale is through the item-total correlation that indicates the level of correlation of each item with the total score. This correlation is given by $R_i = \frac{cov(X_i,P)-s_i^2}{s_i \tilde{s}_i}$. In case it is necessary to delete items, the item-total correlation allows us to obtain information about the items with lower correlations with the rest of the scale and, consequently, about the items to delete. Therefore, the scale items will have a higher internal consistency, despite the fact that, if many items were deleted, it would reduce the reliability of the total scale, measured through the internal consistency. According to Pallant (2007), low scores (below 0.3) indicate that the item is not measuring the scale as a whole, and that there may be an alternative to the scale presented. If the Cronbach's alpha for the total scale is low, for example below 0.7, there may be incorrect items which may need to be deleted according to low item-total correlations.

Conclusion

Every measurement instrument has advantages and disadvantages because each relies on several factors, such as the type of data (scale), how the construct was operationalised (unifactorial or multifactorial), and the objective of the instrument (assessment of a personality trait or knowledge) and the respective items. These factors will dictate the measures used for each case. We are facing a situation of relative measure, given that reliability or consistency are always directly related with what we intend to measure or compare.

All estimators will obtain different results for the same situation. The test-retest usually shows lower values because it depends on more than one assessment. On the other hand, some care must be taken in experimental design, given that the analysis is usually performed in quasi-experimental studies in which there is often no randomness and balance in measurements.

Thus, we cannot postulate a standard for the use of these estimators; instead, we propose a thoughtful choice of the most suitable one for the study in question.

References

Anastasi, A. (1977). *Testes psicológicos* (2.ª ed.). São Paulo, Brasil: Editora Pedagógica e Universitária.

Ary, D., Jacobs, L. C., & Razavieh, A. (1990). *Introduction to research in education* (4th ed.). Fort Worth, USA: Ted Buchholz.

Bryman, A., & Cramer, D. (1993). *Análise de dados em ciências sociais: Introdução às técnicas utilizando o SPSS* (2.ª ed.). Oeiras, Portugal: Celta.

Fornell, C., & Larcker, D. (1981). Evaluating structural equation models with unobservable variables and measurement error. *Journal of Marketing Research*, *1*, 39–50.

Freeman, F. (1980). *Teoria e prática dos testes psicológicos* (2.ª ed.). Lisboa, Portugal: Fundação Calouste Gulbenkien.

Hill, M., & Hill, A. (2012). *Investigação por questionário* (2.ª ed.). Lisboa, Portugal: Edições Sílabo.

IBM® SPSS®. (2011). *IBM SPSS statistics 20 algorithms*. Recuperado de ftp://public.dhe.ibm.com/software/analytics/spss/documentation/statistics/20.0/en/client/Manuals/IBM_SPSS_Statistics_Algorithms.pdf5

Laveault, D., & Grégoire, J. (2002). *Introdução às teorias dos testes em ciências humanas*. Porto: Porto Editora.

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.

Pallant, J. (2007). *SPSS survival manual: A step by step guide to data analysis using SPSS for windows* (3th ed.). New York, USA: Mc Graw Hill.

Peterson, R. (1994). A meta-analysis of Cronbach's coefficient alpha. *Journal of Consumer Research*, *21*(2), 381–391.

Poínhos, R., Correia, F., Faneca, M., Ferreira, J., Gonçalves, C., Pinhão, S., & Medina, J. L. (2008). Desejabilidade social e barreiras ao cumprimento da terapêutica dietética em

mulheres com excesso de peso. *Acta Médica Portuguesa*, *21*, 221–228.

Polit, D., & Hungler, B. (1992). *Investigacion cientifica en ciencias de la salud* (3rd ed.). México: Nueva Editorial Interamericana.

Ribeiro, J. L. P. (1999). *Investigação e avaliação em psicologia da saúde*. Lisboa, Portugal: Climepsi.

Vaz-Serra, A., Ponciano, E., & Freitas, F. (1980). Resultado da aplicação do Eysenk personality inventory a uma amostra da população portuguesa. *Psiquiatria Clínica*, *21*, 127–132.