Protocol for the Evaluation of Tests, Scales and Questionnaires (PETEYC_E)

Users' Manual

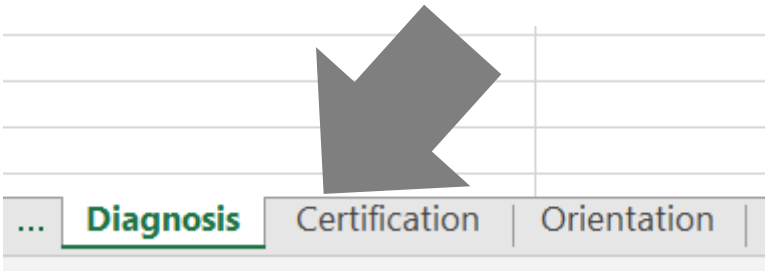
The manual of the Protocol for the Evaluation of Tests, Scales and Questionnaires (PETEYC_E) describes how to administer the tool and how to process the information collected with it.

The steps that must be followed to use PETEYC_E are:

1. Gather all relevant information on the Assessment tool (document 1), including the instrument itself, any available user manual and any empirical data collected for the purpose of assessing the properties of the instrument (qualitative data provided by experts, quantitative data from a pilot study, interviews, etc,).
2. Administer PETEYC_E to the target assessment instrument, completing each of the sections with the available information.
3. Download the "Assessment Processing Tool" (PROC_PETEYC_E) from the IAPA website and follow the instructions of this document to complete the information requested therein. Edit only blank cells.

The following information will guide you on transferring data gathered through PETEYC_E to PROC_PETEYC_E. To do this, the numbers into brackets located in *PETEYC_E- Assessment Tool* will be used.

[1] The first step is to identify the intended use of the test. The use indicated in PETEYC_E will determine the PROC_PETEYC_E tab in which the evaluation will be carried out. For example, if the intended use of the test is "Diagnosis", the corresponding tab will be selected at the bottom of PROC_PETEYC_E, as shown in the image.



[2] The sections marked with codes with the first digit being 2 will be used to assess the suitability of the **characteristics and the size of the sample** used in the pilot study. The information on the sample included in section **[2a]** of PETEYC_E will be used to establish the score for each of the cells in the row dedicated to analyzing "Sample size and suitability", as follows:

- **Size:** The sample size will be evaluated based on the total number of people and items answered. The sample size will be considered adequate when the number of observations is equal to or greater than 200 (Ferrando & Anguiano-Carrasco, 2010) or when, being less than 200, five or ten responses per item are collected (Muñiz & Fonseca-Pedrero, 2019). In both cases, a score of 1 will be assigned, otherwise a 0 will be assigned.
- **Representativeness**: The representativeness of the sample shall be assessed by comparing the characteristics of the sample with the characteristics of the target population specified in paragraph **[2b]**. A score of 1 will be assigned if the characteristics of the sample match those of the target population, and 0 if not.
- **Sample selection:** A score of 1 will be assigned if the sample has been selected randomly or through non-random sampling that has allowed the characteristics of the population to be adequately represented, and 0 in any other situation.

[3] The sections marked with codes which first digit is 3 will be used to assess the **reliability** of the instrument.

- **Inclusion:** A score of 1 will be assigned if data on the reliability of the instrument is provided, and 0 otherwise. If the evaluation has not been included, please fill in only the next cell on "adequacy of the decision" and continue to the next block.
- **Adequacy of the decision:** A score of 1 will be assigned if the reliability assessment is considered relevant, and 0 if not. The assessment of reliability is considered relevant when the test evaluates a single construct and provides a total score or, when evaluating several dimensions presents a reliability analysis for each of them. The assessment of reliability will not be relevant when the objective instrument aims to assess more than one construct through independent items that are not part of a single scale.
- **Adequacy of the procedure:** A score of 1 will be assigned if the procedure used to assess reliability is considered adequate, and 0 if not. The procedure will be appropriate if it adjusts to the characteristics of the assessment instrument. For example, it will be inappropriate if reliability is assessed by test-retesting instruments that assess constructs that may change over time.
- **Support to the intended use:** A score of 1 will be assigned if the results reflect that the instrument is reliable, and 0 if not. For example, the test will be reliable if values above .7 are obtained in indicators such as Cronbach's Alpha or McDonald's Omega. Values above .9 could indicate the presence of redundant content (Panayides, 2013). In the case of procedures such as test-retest or two halves, correlation values greater than .3 shall be considered appropriate. A score of .5 will be assigned when adequate results have been obtained on some subscales

and inadequate on others, or when the values of Alpha or Omega are close to the criterion of .7.

[4] The sections marked with codes whose first digit is 4 will be used to evaluate the extent to which the results reflect that the psychometric properties of the items are appropriate.

- **Inclusion:** A score of 1 will be assigned if data on the psychometric properties of the items is provided, and 0 if not. If the evaluation has not been included, please fill in only the next cell on "adequacy of the decision" and continue to the next block.
- **Adequacy of the decision:** A score of 1 will be assigned if the psychometric data provided is considered relevant, and 0 otherwise. The data will be relevant when they include the distribution of responses in the different alternatives, discrimination indices, difficulty indices (only for tests of optimal performance) and reliability of the instrument when eliminating each item. A partial score (.5) will be assigned when some of the data is provided.
- **Adequacy of the procedure:** A score of 1 will be assigned if the procedure used to evaluate each of the psychometric properties is considered adequate according to the approach followed (Classical Test Theory or Item Response Theory), and 0 if not. The procedure will be appropriate if it adjusts to the characteristics of the assessment instrument.
- **Support to the intended use:** The score assigned in this section will reflect the proportion of items that show adequate properties. So, a score of 1 will be assigned when all items work properly. If 80% of the items reflect adequate properties, the score will be .8. Values greater than .3 will be appropriate

when calculating the discrimination index using item-total correlation (Bichi, 2016).

[5] The sections marked with codes whose first digit is 5 will be used to assess the extent to which the evidence based on the test content supports the use and intended scores' interpretations.

- **Inclusion:** A score of 1 will be assigned if data derived from analyzing the overlap between the theoretical model and the content of the test is provided, and 0 otherwise. If the evaluation has not been included, please fill in only the next cell on "adequacy of the decision" and continue to the next block.
- **Adequacy of the decision:** A score of 1 will be assigned if, given the characteristics of the instrument, the validity evidence based on the test content provides or could provide relevant information. This source of evidence will be relevant when it is necessary to show that the instrument has adequately collected the indicators of the target construct.
- **Adequacy of the procedure:** A score of 1 will be assigned if the procedure used to obtain validity evidence based on the test content is considered adequate, and 0 if not. For example, the procedure will be appropriate if it shows results on the representativeness and relevance of the items created to evaluate the construct, and content validity indices are presented. A score of 1 will be assigned if qualitative data are presented to identify potentially problematic items and suggestions or changes aimed at improving the quality of the items are incorporated.
- **Support to the intended use:** A score of 1 will be assigned if the results support the intended use of the test, i.e., if evidence is provided that confirms that the instrument evaluates the pursued construct**.** For example, when the

CVR values according to Aiken's V Index are equal to or greater than .8 (Penfield & Giacobbi, 2004). For additional details, see Sireci and Faulkner-Bond (2014). A score of 1 will also be assigned when qualitative data is provided to support the intended use of the instrument. A score of 0 will be assigned otherwise.

[6] The sections marked with codes which first digit is 6 will be used to assess the extent to which the evidence based on the internal structure supports the use and the scores' interpretation.

- **Inclusion:** A score of 1 will be assigned if data derived from analyzing the dimensionality of the instrument, and/or the invariance of the measure, and 0 otherwise. If the evaluation has not been included, please fill in only the next cell on "adequacy of the decision" and continue to the next block.
- **Adequacy of the decision:** A score of 1 will be assigned if, given the characteristics of the instrument, the validity evidence based on the internal structure provides or could provide relevant information. This source of evidence will be relevant when it is necessary to show that the structure of the instrument adequately reflects the theoretical dimensions of the construct and when it is relevant to show the invariance of the measure between different groups.
- **Adequacy of the procedure:** A score of 1 will be assigned if the procedure used to obtain validity evidence based on the internal structure of the test is considered adequate, and 0 if not. For example, the procedure will be appropriate if it reflects the overlap between the construct and instrument configurations.

- **Support to the intended use:** A score of 1 will be assigned if the results support the intended use of the test, that is, if evidence provided confirms that the instrument includes the theoretical dimensions of the construct. For example, when adequate values are provided on the fit of the model obtained through exploratory or confirmatory analyses (CFI >.95; TLI>.95, SRSM<.08; RMSEA<.06). A score of 0 will be assigned otherwise. Details of the criteria commonly used can be found in the following sources: Bentler (1990); Hu and Bentler (1999); Van Laar and Braeken (2021).

[7] The sections marked with codes whose first digit is 7 will be used to assess the extent to which evidence based on relationships with other variables supports the intended use and scores' interpretation.

- **Inclusion:** A score of 1 will be assigned if data derived from analyzing relationships between the scores of the target instrument and scores of other instruments that measure theoretically related variables are provided, and 0 otherwise. If the evaluation has not been included, please fill in only the following cell on "adequacy of decision" and continue to the next section.
- **Adequacy of the decision:** A score of 1 will be assigned if, given the characteristics of the instrument, the validity evidence based on relationships with other variables provides or could provide relevant information. This source of evidence will be relevant when it is necessary to show that the scores of the instrument are consistent with the scores in other instruments that assess theoretically related variables.

- **Adequacy of the procedure:** A score of 1 will be assigned if the procedure used to obtain validity evidence based on relationships with other variables is considered adequate, and 0 if not. For example, the procedure will be appropriate if it reflects relationships between the scores of the instrument and scores of other instruments that assess theoretically related variables.
- **Support to the intended use**: A score of 1 will be assigned if the results support the intended use of the test, i.e. if evidence is provided confirming that the intended relationship exists, for example, when values of correlations are greater than .3 (positive or negative, as appropriate) or values reflect an area under the curve (AUC) close to 1 (in ROC curve analysis; Muñiz, 2018). Partial scores (.5) will be assigned when the expected relationships are found with some variables, but not with all of them; or when the values obtained are close to the expected values. A score of 0 will be assigned in case no relationship is found with any variables.

[8] Sections marked with codes whose first digit is 8 will be used to assess the extent to which evidence based on response processes supports the intended use and scores' interpretation.

- **Inclusion:** A score of 1 will be assigned if data derived from analyzing response processes occurred when responding to the instrument is provided, and 0 if not. If the evaluation has not been included, please fill in only the next cell on "adequacy of the decision" and continue to the next block.
- **Adequacy of the decision:** A score of 1 will be assigned if, given the characteristics of the instrument, the validity evidence based on the

response processes provides or could provide relevant information. This source of evidence will be relevant when it is necessary to collect information that reflects that people develop response processes aligned with the target construct.

- **Adequacy of the procedure:** A score of 1 will be assigned if the procedure used to obtain validity evidence based on the response processes is considered adequate, and 0 if not. For example, the procedure will be appropriate if it provides information about the process developed to respond to the instrument.
- **Support to the intended use:** A score of 1 will be assigned if the results support the intended use of the test, in other words, if evidence is provided that confirms that people think about the intended indicators. For example, when they provide arguments that reflect that the answers refer to the intended construct. A score of 0 will be assigned otherwise. Partial scores (.5) will be assigned when evidence is provided only in some cases.

[9] Sections marked with codes with the first digit number 9 will be used to assess the extent to which the validity evidence based on testing consequences supports the intended use and scores' interpretation.

- **Inclusion:** A score of 1 will be assigned if data derived from analyzing the consequences of the evaluation are provided, and 0 if not. If the evaluation has not been included, please fill in only the following cell on "adequacy of decision" and continue to the next section.
- **Adequacy of the decision:** A score of 1 will be assigned if, given the characteristics of the instrument, the validity evidence based on testing consequences provides or could provide relevant information. This source

of evidence will be relevant when it is necessary to collect information that reflects that the evaluation has not had unintended consequences for the people evaluated.

- **Adequacy of the procedure:** A score of 1 will be assigned if the procedure used to obtain validity evidence based on testing consequences is considered adequate, and 0 if not. For example, the procedure will be appropriate if it provides information on the consequences of the evaluation for the people being evaluated.
- **Support to the intended use:** A score of 1 will be assigned if the results support the intended use of the test, in other words, if evidence is provided confirming that people have not suffered unintended consequences. For example, when they provide arguments that reflect that the evaluation has not generated a situation of injustice. A score of 0 will be assigned otherwise**.**

Once PROC_PETEYC_E is completed, a total score between 0 and 100 will be obtained. Scores close to 100 will indicate that the instrument has sufficient evidence to be used in the intended context. Scores close to 0 indicate the need for a thorough review of the instrument. PROC_PETEYC_E also facilitates the identification of weaker areas within the instrument. Those rows whose total score (column G) is lower than the value included in the weight (column F) will be the areas that could be worked on to improve the quality of the instrument. To work on them, the red indications incorporated in PETEYC_E can be followed.

Referencias

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological bulletin*, *107*(2), 238-246.

Bichi, A. A. (2016). Classical Test Theory: An introduction to linear modeling approach to test and item analysis. *International Journal for Social Studies*, *2*(9), 27-33.

Ferrando, P. J., & Anguiano-Carrasco, C. (2010). El análisis factorial como técnica de investigación en psicología. *Papeles del psicólogo*, *31*(1), 18-33.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural equation modeling: a multidisciplinary journal*, *6*(1), 1-55.

Muñiz, J. (2018). *Introducción a la Psicometría: Teoría clásica y TRI*. Pirámide.

Muñiz, J., & Fonseca-Pedrero, E. (2019). Diez pasos para la construcción de un test. *Psicothema*, *31*(1).

Panayides, P. (2013). Coefficient Alpha: Interpret With Caution. *Europe's Journal of Psychology*, *9*(4), 687-696. https://doi.org/10.5964/ejop.v9i4.653

Penfield, R. D., & Giacobbi, J. (2004). Applying a score confidence interval to Aiken's item content-relevance index. *Measurement in physical education and exercise science*, *8*(4), 213-225.

Sireci, S., & Faulkner-Bond, M. (2014). Validity evidence based on test content. *Psicothema*, *26.1*, 100-107.

Van Laar, S., & Braeken, J. (2021). Understanding the Comparative Fit Index: It's all about the base! *Practical Assessment, Research & Evaluation*, *26*(1).

Funded by Program 39 of the Granada's University Research and Transfer Plan, 2021