Special Article

Clinical Instruments Validation: Key Aspects

Validación de instrumentos clínicos: aspectos esenciales

Haydee Alejandra Martini-Blanquel[*]

Summary

This study explores some key aspects of the validation of instruments used in clinical data collection, which is part of a critical process for diagnosis and treatment in the medical field. Clinimetrics underscores the importance of validity, and reliability in the use of instruments for the accurate identification, and measurement of signs and symptoms. Validity ensures that the instrument effectively measures the variable of interest, while reliability indicates that repeated use of the same instrument will produce consistent results. The article addresses aspects related to instrument construction and validation, including theoretical and empirical foundations, validation by expert judges, pre-sampling tests, as well as construct and criterion evaluations. In addition, the application of exploratory and confirmatory factor analysis to validate the internal structure of the instruments is addressed. This work highlights the importance of accuracy in clinical practice and provides a detailed framework for ensuring the efficacy and relevance of instruments in different population contexts.

Key Words: Validation Studies; Reliability; Validity.

Received: 11/06/2023
Accepted: 04/18/2024

[*]Family Medicine Unit No. 33, Mexican Institute of Social Security.

Correspondence:
Haydee Alejandra Martini-Blanquel
haydee.martini@imss.gob.mx

Suggestion of quotation: Martini-Blanquel HA. Clinical Instruments Validation: Key Aspects. Aten Fam. 2024;31(3): 177-184. http://dx.doi.org/10.22201/fm.14058871p.2024.388838

Resumen

Este estudio explora algunos aspectos clave sobre la validación de instrumentos utilizados en la recolección de datos clínicos, lo cual es parte de un proceso crítico para el diagnóstico y tratamiento en el campo médico. La clinimetría subraya la importancia de la validez y confiabilidad en el uso de instrumentos para la identificación y medición precisa de signos y síntomas. La validez asegura que el instrumento mida efectivamente la variable de interés, mientras que la confiabilidad indica que el uso repetido del mismo instrumento producirá resultados consistentes. El artículo aborda aspectos relacionados con la construcción y validación de instrumentos, incluyendo la fundamentación teórica y empírica, la validación por jueces expertos, pruebas de premuestreo, así como evaluaciones de constructo y criterio. Además, se aborda la aplicación del análisis factorial exploratorio y confirmatorio para validar la estructura interna de los instrumentos. Este trabajo resalta la importancia de la precisión en la práctica clínica y ofrece un marco detallado para garantizar la eficacia y relevancia de los instrumentos en diferentes contextos poblacionales.

Palabras clave: estudios de validación, fiabilidad, validez.

Introduction

The collection of clinical data is a process of great value for the physician, since it allows the establishment of timely diagnoses, treatments, and prognoses in the patients being treated. However, in the case of many nosological entities that are the object of an investigation -additional to the collection of data in a clinical history or a medical note-, it is necessary to explain and quantify signs and symptoms more accurately. To achieve this, the methodology used must be correct, as this will influence decision making, whether to make a diagnosis, prescribe medication, or establish the prognosis of a disease.[1] Hence, abstract concepts (signs and symptoms) are converted into scientific data that can be measured. This is known as "clinimetrics".[2]

Thus, the term "measurement" is always present in medicine, as in other areas of health. For example, when a person is suspected of being overweight, the physician will calculate the body mass index (BMI) and then use the World Health Organization scale to confirm his or her diagnostic suspicion.[3] Likewise, when the patient wishes to corroborate the diagnosis of depressive disorder, it is likely that physician will apply an accepted and recognized scale; in this case, it could be the Beck inventory[4] or the Hamilton rating scale.[5] However, despite the daily use of these tools, their importance, the process for developing them and how they help the physician's daily work are often unknown.

As described above, it is important that all instruments used in the clinic adequately represent the concepts or variables that the physician wants to measure as accurately as possible; this concept is known as validity. Thus, when the variables of interest are measured in this way, it is easier to give answers that are closer to reality.[1]

Reliability and Validity of an Instrument

Validity refers to the degree to which an instrument actually measures the variable it is intended to measure.[6] Validity is directly related to reliability, which means that a phenomenon, when measured many times with the same instrument, yields the same results.[1]

This definition of validity is the one we usually find in the literature; however, for decades, there have been other proposals, which state that an instrument is valid as long as its results help to make inferences and interpretations and therefore, there are social and ethical consequences of its application.[7] For example, when a test allows us to choose which patients are candidates to receive a pharmacological treatment and therefore, improve their state of health, or when a scale identifies people at high risk of suffering from a certain disease and with this, the physician implements preventive measures.

However, the fact that an instrument is reliable does not guarantee that it is accurate in its measurements. This applies because any instrument, when developed in a certain context, may not be useful in another population; even reliability tends to change when the instrument is adapted to a certain type of person or adjusted to a different language.

The validation process is therefore a continuous one, which means that we constantly carry out verifications that show us that an instrument is adequate to an acceptable degree, always considering the objectives for which it was created, and the population at which it is aimed.

Process for Constructing and Validating Research Instruments

Considering the above, the basic steps to carry out the construction and validation of an instrument are four: 1. Theoretical and empirical foundations of the instrument (considering the objectives for which it will be created), 2.

Aten Fam. 2024;31(3): 177-184. http://dx.doi.org/10.22201/fm.14058871p.2024.388838

Validation of the instrument by judges, 3. Pre-sampling test, 4. Construct and criterion validation, and 5. Calculation of the reliability (internal consistency) of the instrument.

1. Theoretical and Empirical basis of the instrument

Measurement instruments in clinical research, from a quantitative approach, should be developed considering several concepts and criteria; concept is synonymous with construct, which is developed with the aim of achieving a measurement with scientific rigor. Thus, each instrument must have a theoretical, and an empirical basis.[8]

a. Theoretical basis

This refers to concepts that are not directly observable. For example, if we want to construct an instrument that measures the variable "health-related quality of life in patients with rheumatoid arthritis", we will have to review the literature and determine whether there are instruments that measure this concept. After a detailed review, the researcher will decide whether or not, the existing information is close to what is intended to be measured (health-related quality of life in patients with rheumatoid arthritis). At this point, it is possible that the concept is partially defined, or that there is no precedent for it in the literature; if this is the case, it will be necessary to complete or fully develop a theoretical proposal, selecting the concepts, and then the indicators, which are the external manifestations that make it possible to measure a construct or concept.[9]

Here, the researcher becomes the number one expert on the chosen topic and therefore has the capacity to generate new concepts.[10] However, other options for obtaining information and developing the theoretical proposal are: 1) requesting support from judges or experts on the topic, or 2) conducting interviews with a population similar to the one at which the instrument will be aimed (especially in the case of those aimed at patients).

At the time of the search, it is possible that the construct we are trying to measure has only one attribute, in which case it will be unidimensional. However, many of the instruments used in clinical practice come from complex constructs, as is the case of the example mentioned above (health-related quality of life), since if literature is reviewed, it will be found that various aspects such as mobility, emotional well-being, personal care, etc., are considered for their evaluation.[11] For this reason, these instruments are called multidimensional, since they consider various indicators to measure a single characteristic. Thus, the principal investigator will select the concepts and indicators to be included in the instrument and will summarize them as much as possible, avoiding repetition.[10]

Finally, it is important to emphasize that at this point it is not necessary to question in depth the relevance of the concepts included, as specific procedures will be carried out later to verify this.

b. Empirical Foundation

The empirical part refers to the adequacy of the instrument based on theory, i.e., on concepts already defined; if this theory is well founded and congruent, it will make it easier to write the items or questions.

Following the previous example on "health-related quality of life in patients with rheumatoid arthritis", after developing or complementing our theoretical proposal, it must be ensured that there is congruence between the objective of the instrument, the concepts, and the items to be included. So, if "health-related quality of life in patients with rheumatoid arthritis" will be measured, it would not be logical to elaborate items on "quality of life" only, since conceptually they are different terms.

Subsequently, we need to specify the characteristics of the items, i.e., whether they will be statements in the form of an assertion or multiple-choice questions, as well as the type of answers, etc. In the construction of the items, the list of topics or concepts chosen in the previous phase (theoretical basis) must be taken into account, as well as the characteristics of the population to whom the instrument is addressed (medical personnel, children, elderly, women, people with disabilities, etc.). Once the items and the list of possible response options are available, the measurement scale and the type of instrument are established, which will depend on the nature of the concept, and the indicators (attributes) to be evaluated. For example, if we wish to evaluate the consumption of certain foods, a questionnaire with multiple-choice answers may be the most convenient; on the other hand, if we wish to measure the level of patient satisfaction with the care provided, then statements can be made in the form of an assertion, with Likert-type responses, in which the respondent must indicate his/her agreement or disagreement with each assertion, with the minimum score being given when there is greater disagreement, and the maximum when there is greater agreement. Generally, the numerical response options range from 1 to 5.[12]

At this point, there is not a structured tool; there is still the possibility to formulate more questions, and to change the form and order of them, in addition to the concepts already obtained. This means that this is a flexible phase in which we can add elements not previously considered, without losing sight of the concept and the indicators to evaluate.

In summary, if there is a strong relationship between the indicators, which are the observed responses, and the unobservable concepts, it can be said that there is empirical applicability of the theoretical propositions, and therefore an adequate measurement will be made, the results of which will lead to a better understanding of the phenomenon studied.[7]

2. Validation of the Instrument by Judges

The principal investigator selects the judges to whom he/she will present the initial version of the instrument. The judges should be familiar with the research and, especially, with the instrument validation process, but are not necessarily related to the object of study in question. Therefore, a judge will not provide input on the content of the instrument; his or her main task will be to evaluate the items that were constructed, taking into account three aspects: sufficiency (that the number of items is sufficient to evaluate the expected concept), relevance (that the items are precise, and according to the topic of interest), and clarity in the wording (that the use of language and terminology are appropriate to the type of population to whom the instrument is directed).[10]

Thus, the difference between a judge and an expert is that the latter is in direct contact with the research topic, and has sufficient experience in the area, although not necessarily in the validation of instruments. For example, if the knowledge that family physicians have in the diagnosis of benign breast pathologies is evaluated, the experts to validate the instrument would be first contact physicians, specialists in gynecology or any other health care personnel who are directly involved, and in constant contact with this type of patient. Having said that, it is important to mention that a judge can be an expert at the same time; however, the fact that a judge is not an expert is not a condition to exclude him/her from the validation process.[10]

Before the judges begin their work, it is essential that they are clear about the theoretical basis of the research, and the objectives of the instrument, i.e., the construct. To this end, it is desirable for the principal investigator to provide them with written information about the instrument, and the manner in which they are to conduct the observations in various formats: letters, instructions, task lists, etc. Likewise, the judges will make the corrections and give feedback to the responsible researcher, also in written form, considering the three elements previously mentioned (sufficiency, pertinence, and wording). This activity is called "rounds" and it should be noted that there is no minimum or maximum number of rounds; it depends on the complexity of the instrument, the number of items, and the level of experience of all those involved. After these revisions, the most common thing is to reduce the number of items, and indicators, since the judges make different contributions to the instrument, ranging from modification to elimination of items, as well as hierarchization and elimination of indicators. This is known as the Delphi technique, in which the judges make a blind and independent evaluation of the stems, and items according to the theoretical view of the instrument, as well as the responses to the items they consider appropriate.[13]

Another task that can be carried out with the help of the evaluators is to ensure that the number of questions per indicator is balanced; for example, if an instrument has 100 items, and 5 indicators, it is desirable to include 20 items per indicator to be evaluated, as this will avoid the questions being more focused on one issue than another. This is not a rule for all instruments, but it is an aspect that can be considered, taking into account the objectives of the instrument in question. Similarly, some instruments will also require a certain homogeneity in the responses of the items; an example is those with "false" or "true" response options, where it is desirable that half be constructed to be answered as false and the other half as true.

Thus, after each round, a new version of the instrument is produced and each judge should be informed about the coincidence of his observations with other judges; if not, he should be asked to explain the reasons for his opinion. These arguments for and against the content serve the responsible researcher to make decisions on what to include or not, since he has the last word on the content of the instrument. The review rounds end when the level of consensus among the judges is as expected (minimum four of the five judges). At this point, a test can be applied to assess the intra-judge reliability or inter-judge reliability, which aims to determine

the percentage of agreement between them, i.e., to what extent they agreed on the classification in relation to the total number of items examined. This is called the inter-rater concordance index, the most commonly used formula being the Kappa index:[14]

$$k = \frac{p_o - p_e}{1 - p_e}$$

Where:
Po= proportion of observed agreement (sum of agreements achieved in each category divided by the number of records).
Pe= proportion of agreement expected at random (sum of the probability of agreement at random for each category).

The result ranges between 0 and 1 (0.1, 0.2, 0.3, etc.), so that if it is closer to one there will be greater agreement.

Once the instrument has been balanced, with all the items and the concordance index calculated, the next stage will proceed.

3. Pre-sampling Test

At this point, the instrument should already have undergone content validation (preliminary version of the instrument given by the judges). However, it is necessary to know the characteristics it has to measure what it is supposed to measure. Then we will need a population to use it, and then perform appropriate statistical tests. This is known as pre-sampling or pilot testing.

The first thing to do is to select a group of people that is as close as possible to the population at which our instrument is aimed. For example, if the objective of the instrument is to determine the presence of unfavorable eating habits in older adults with diabetes, the ideal would be to conduct a pre-sampling test on patients living with this disease who are also adults over 60 years. It should be emphasized that the size of the group selected for a pilot test is not the most important thing, since in research it is often pointed out that large samples significantly reduce the possibility of error; however, for studies whose purpose is to validate an instrument, this is not entirely true, since what is considered is not the number of participants, but the number of questions or items that form part of the instrument.[15]

Normally, for validation, between 5 and 10 participants per item should participate, with a minimum of 300, since this gives us greater guarantees of the validity of the instrument; others point out that between 2 and 3 participants per item is sufficient, as long as the total number is not less than 200. However, smaller samples may be allowed if the intention is to replicate the measurement with different groups, where the number of subjects is at least double the number of items, with a total of not less than 100 participants per group.[16]

It is important to mention that one of the main purposes of the pilot test is to assess the clarity of the instrument, and therefore it is a requirement that the responsible researcher be present so that he/she can clarify any doubts directly with the respondents, on the understanding that the instrument has not yet been fully validated.[10]

4. Construct and Criterion Validation

Multivariate statistical techniques include factor analysis (FA), which is frequently used in the instrument validation process. In general, two basic types of factor analysis are known: exploratory factor analysis (EFA), and confirmatory factor analysis (CFA).

The first type: EFA, aims to try to establish an underlying structure between the variables of the analysis, based on correlation structures between them; that is, items (better known as factors) that are highly correlated with each other are grouped, and assigned a concept.

Before performing an EFA, the assumption of correlation between variables should be evaluated to determine whether or not its use is justified. One of the most commonly used strategies to evaluate this assumption is to examine the correlation matrix. If the variables generally have low correlation values between them (values less than 0.30), it is necessary to question whether this analysis makes sense. Another alternative to evaluate these correlations is to use Bartlett's sphericity test, which has as its null hypothesis that there is no correlation between the variables; if this hypothesis is rejected, it shows that there is in fact some degree of statistically significant correlation. A third method is to assess the strength of the relationship between two variables, or items using the Kaiser-Meyer-Olkin Index (KMO), which takes values between 0 and 1, with values below 0.5 considered unacceptable; from 0.5 to 0.59, poor; from 0.6 to 0.79, fair; and from 0.8 to 1, acceptable.[17]

The interpretation of the results is one of the most important aspects of the AFE, since it depends largely on experience. One of the ways is the factor

rotation method, which, as the name implies, means that the axes of the factors are rotated to different degrees, but the origin is fixed, redistributing the variance of the original variables in the factors in order to obtain a better interpretation of the results. Two types of rotations are currently used in AFE, chosen by the researcher according to his knowledge of the problem. These rotations are orthogonal and oblique, of which the best known are Varimax, Quartimax, and Equamax (orthogonal), and Oblimin, and Promax (oblique).[17]

On the other hand, CFA confirms that the set of factors previously organized theoretically (by concepts) fits. This is where the researcher plays a very important role, since the greater the knowledge of the problem, the greater the ability to formulate, and test much more concrete and specific hypotheses.[17]

The two analyses are not mutually exclusive, but depending on the objectives of the tool, it is necessary to decide which is the most appropriate, although in some cases it is preferable to carry out both.

One of the recommendations when grouping concepts or dimensions is to seek a balance in the number of items each has, although this is not an absolute requirement. Thus, it is possible that some of the items may be grouped in a different dimension than the one to which they originally belonged.[10]

a. Perform Criterion Validation

A criterion is nothing more than the second way of evaluating the concept that it is intended to measure. Therefore, it is necessary to return to point 1, referring to the theoretical, and empirical foundations of the instrument.

Criterion validity means that the results obtained with the instrument developed are similar to those obtained from other instruments applied to the same population. To do this, first it is needed to know whether the concept to be measured with the instrument is already clearly defined in the literature.

If the severity of lower urinary tract symptoms associated with benign prostatic hyperplasia (BPH) is wanted to be measured, it is known that there is already a validated, reliable and widely used instrument, the IPSS or International Prostate Symptom Score. For this reason, this tool would be considered the gold standard.[18]

And why is an instrument called that? Because it usually takes the authors who develop it a lot of time, a lot of resources, and may even involve invasive testing of patients to construct it properly. Then it is necessary to compare our instrument with the gold standard by calculating the concordance or correlation between the two scales, which must be greater than 0.8. This can be done using the Cohen's Kappa concordance test, or Spearman's correlation test, especially if we want to analyze the categories of the instrument. On the other hand, if the numbers obtained from the variables of the instrument are compared, the Pearson's r correlation coefficient will be used.

Another option is that an instrument already exists for the concept of study, but it is not exactly the gold standard. This is very frequent in social, educational, and psychological research. For example, if an instrument is constructed to evaluate self-esteem, several scales already exist, such as the Rosenberg self-esteem scale,[19] and the Coopersmith scale.[20] Accordingly, the next step is to evaluate the same population with the two instruments: the one that already exists, and the one we construct. With the results obtained, the concordance between instruments will also be established, establishing a kind of consensus for the measurement of this concept.

However, there is a third possibility: that the concept is not previously defined in the literature. Therefore, it is not possible to carry out a criterion validation, since the line of research is completely new, and the results of the application of the instrument will tell us if it is indeed useful to evaluate the proposed concept.[5]

5. Calculation of the Reliability (Internal Consistency) of the Instrument

This step consists of calculating the internal consistency, which refers to the degree to which the items or items that are part of a scale, or instrument correlate with each other so that they measure the same construct. Reliability is a measure of homogeneity, and items are expected to be highly correlated; in addition, the questions of each indicator should be similar to each other.[21]

a. Define the Type of Scale Used

First, it is important to know what type of scale was used. Thus, for those instruments with a dichotomous response pattern (for example, true or false), we would use the Kuder-Richardson formula 20 (when the items have different difficulty indexes), or 21 (when the difficulty indexes are equal). For those with a polytomous response scale, Cronbach's Alpha test is used. These are two of the most common methods for calculating internal consistency and are mathematically equivalent.[15]

Aten Fam. 2024;31(3): 177-184. http://dx.doi.org/10.22201/fm.14058871p.2024.388838

Table 1.

KR-20= $k / k - 1 [1- \Sigma piqi / \sigma T^2]$	Cronbach's Alpha= $k / k - 1 [1- \Sigma \sigma i^2 / \sigma T^2]$
k: number of ítems	*k*: number of items
pi: % affirmative of the item	Σi: variance of the item
qi: complement of p	σT: total variance of the scale
σT: total variance of the scale	

Once the results of the pre-sampling test are available, the next step is the application of the corresponding tests: (see table 1).

The SPSS and STATA statistical packages can be used to perform these psychometric tests.

For the internal consistency of an instrument to be considered acceptable or high, it must be between 0.70, and 0.90. Now this coefficient means will be explained. For example, if an instrument has an internal consistency of 0.8, it means that 80 percent of the variability is true, and that the remaining 20 percent may be the result of measurement error, not the instrument itself. Finally, any value below 0.7 indicates that there is little correlation between the items of the instrument; on the other hand, if the coefficient is above 0.9, there is a risk of redundancy or duplication of items, so the instrument should be revised to eliminate those that are duplicated.[15]

It is common for instruments with more than 20 items to have an internal consistency greater than 90, because when the substitution in the corresponding formula is made, the number of items should be entered. Therefore, it is advisable to calculate the internal consistency also by groups of items, which will avoid overestimating the instrument.[10]

One of the methods that can be used is that of semi-partitioning (dividing the instrument into two halves), so that each half has the same number of items, and can be considered parallel. The total score is then calculated for each of these halves. An example of how the halves might be separated would be in an instrument with 200 items: one half would be the first 100 items, and the second the remaining 100, or one group of items would be the even items and the other the odd items. Thus, the Spearman-Brown correction formula can be used to calculate the consistency between groups of items.[22]

$$r_{xx'} = \frac{2r_{AB}}{1 + r_{AB}}$$

This formula expresses the relationship between the length and the reliability of the instrument, under the assumption that both parts of the instrument are parallel. Here, r is the reliability coefficient for the half of the test, rxx' is the reliability for the total test; also, it is obtained the score in form A, and in form B for each subject.[22]

To give an example, if the correlation of the total scores of the odd items with the total scores of the even items is 0.85, the estimated reliability of the whole test would be:

$$r_{xx'} = \frac{2r_{AB}}{1 + r_{AB}}$$

b. Calculation of the Correlation Index

Once the internal consistency of all the items has been obtained (both overall and by groups of items), they are ordered according to their correlation index, from those with the highest correlation to those with the lowest. Following the previous example, if we have a group of 200 items, those with lower correlation magnitude (consistency less than 0.8) will be eliminated. One possible explanation for why items may have a low correlation is that they are ambiguous. If you still have items with low correlation, the most appropriate thing to do is not to continue eliminating them, but to implement a method to increase the Cronbach's alpha value. To do this, it is necessary to arrange the items not according to their correlation index, but according to their variance.

The item with the highest degree of variability or the greatest magnitude of variance is placed in the first row, and those with the lowest degree of variability are placed downwards, in order. Once ordered, the first items with the lowest degree of variability are selected, and the wording is modified, so that the way in which the subjects answer is more dispersed; the Cronbach's alpha value is recalculated. If not modified, continue

with the following items, in ascending order. By making modifications to the items to avoid all subjects answering the same thing, Cronbach's alpha will benefit.[10] This same method can be used with the Kuder-Richardson test, following the steps previously mentioned.

Conclusion

This study highlights the critical importance of adequate validation of instruments used in clinical data collection to ensure accuracy and reliability in medical diagnosis and treatment. The rigorous application of methods such as clinimetrics, expert validation, and factor analysis, both exploratory and confirmatory, allows confirm that these instruments meet the necessary standards to effectively measure the variables of interest.

References

1. Villasís KM, Márquez GH, Zurita CJ, Miranda NM, Escamilla NA. El protocolo de investigación VII. Validez y confiabilidad de las mediciones. Rev Alerg Mex. 2018;65(4):414-21.
2. Iglesias GA, Quintana G. Análisis histórico de la clinimetría y de la autoclinimetría. Estado del arte. Rev Colomb Reumatol. 2013;20(1):1–8.
3. Organización Mundial de la Salud [Internet]. [Citado 2023 Jul 5]. Disponible en: https://www.who.int/es/news-room/fact-sheets/detail/obesity-and-overweight
4. Padrós BF, Montoya PK, Bravo CM, Martínez MM. Propiedades psicométricas del Inventario de Ansiedad de Beck (BAI, Beck Anxiety Inventory) en población general de México. 2020; 26:181–187.
5. Dos Santos ERP, Coelho JCF, Ribeiro I, Sampaio F. Translation, cultural adaptation and evaluation of the psychometric properties of the Hamilton Anxiety Scale among a sample of Portuguese adult patients with mental health disorders. BMC Psychiatry. 2023;23(1):520.
6. Hernández SR, Fernández CC, Baptista LM. Metodología de la investigación. 6a ed. México: McGraw-Hill Interamericana; 2014.
7. Soriano AM. Diseño y validación de instrumentos de medición. Diálogos. 2014;14,19-40.
8. Kerlinger F. Investigación del Comportamiento. 4a ed. México: McGraw-Hill; 2002.
9. Bravo PT, Valenzuela GS. Desarrollo de instrumentos de evaluación: cuestionarios. Chile: Centro de Medición MIDE UC; 2019.
10. Supo J. Cómo validar un instrumento. La guía para validar un instrumento en 10 pasos. Perú: 2013 [Internet]. [Citado 2023 Ago 22]. Disponible en: https://dspace.uniandes.edu.ec/handle/123456789/16000
11. Urzúa MA. Calidad de vida relacionada con la salud: elementos conceptuales. Rev Med Chile. 2010;138(3):358-365.
12. Matas A. Diseño del formato de escalas tipo Likert: un estado de la cuestión. REDIE. 2018;20(1):38-47
13. Cobos AH. Cómo construir un instrumento para evaluar la lectura crítica de investigación de informes médicos. Inv Ed Med. 2021;10(39):96-105.
14. Dubé JE. Evaluación del acuerdo interjueces en investigación clínica. Breve introducción a la confiabilidad interjueces. Revista Argentina de Clínica Psicológica. 2008; XVII(1):75-80.
15. Campo AA, Oviedo HC. Propiedades psicométricas de una escala: la consistencia interna. Rev Salud Pública. 2008;10(8):831-839.
16. Roco VA, Hernández OM, Silva GO. ¿Cuál es el tamaño muestral adecuado para validar un cuestionario? Nutr. Hosp. 2021;38(4):877-878.
17. Méndez MC, Rondón SM. Introducción al análisis factorial exploratorio. Rev Colomb Psiquiatr. 2012;41(1):197-207.
18. Preciado ED, Kaplan SA, Iturriaga GE, Ramón TE, Mayorga GE, Auza BA, et al. Comparación del Índice Internacional de Síntomas Prostáticos versus Escala Visual Análoga Gea para la evaluación de los síntomas de la vía urinaria inferior. Rev Mex Urol. 2017;77(5):372-382.
19. Martínez RG, Alfaro UA. Validación de la escala de autoestima de Rosenberg en estudiantes paceños. Fides Et Ratio. 2019;17(17):83-100.
20. Díaz RA, Pérez MG, Puentes ML, Castillo MM. Fiabilidad y validez de constructo del Inventario de Autoestima de Coopersmith en estudiantes de medicina. Revista de Ciencias Médicas de Pinar del Río. 2022; 26 (3):e5371.
21. Rodríguez RJ, Reguant AM. Calcular la fiabilidad de un cuestionario o escala mediante el SPSS: el coeficiente alfa de Cronbach. REIRE Revista d'Innovació i Recerca en Educació. 2020;13(2):1–13.
22. Warrens MJ. Transforming intraclass correlation coefficients with the Spearman-Brown formula. J Clin Epidemiol. 2017;85:14-16.