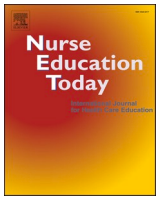
Nurse Education Today 119 (2022) 105603

Contents lists available at ScienceDirect

Nurse Education Today

journal homepage: www.elsevier.com/locate/nedt



Research article

Instrument development and psychometric testing in nursing education research

Kristina Mikkonen [a,*], Marco Tomietto [b], Roger Watson [c,d]

[a] Research Unit of Health Sciences and Technology, University of Oulu, Finland
[b] Department of Nursing, Midwifery and Health, Faculty of Health and Life Sciences, Northumbria University, Newcastle Upon Tyne, UK
[c] Nurse Education in Practice, Netherlands
[d] Southwest Medical University, China

ARTICLE INFO

Keywords:
Instrument development
Instrument testing
Psychometric testing
Nursing education research

1. Introduction

In health care, researchers have increasingly focused on instrument development and psychometric testing. For that reason, researchers, managers, and educators now have at their disposal many validated instruments to assess and measure health care problems in a valid and reliable way.

The availability of valid instruments allows researchers to advance the use of quantitative research methods and empirically test complex models. Testing models has the potential to build new theoretical perspectives (Im and Meleis, 2021) and to inform the nursing community on effective ways to implement interventions in different fields, from clinical practice (Vellone et al., 2021), to nurse management (Tomietto et al., 2019) and health care education (Mikkonen et al., 2020).

Many constructs used in health care are aligned with psychological and behavioural constructs. The caring construct, to cite a fundamental construct, is theoretically based, and it is a latent (conceptual) construct. It is only possible to measure caring by measuring nurses' and patients' perceptions of the behaviours which make caring visible. It is assumed that caring exists because it is possible to observe or perceive some specific behaviours. The same applies to health care education, e.g., the constructs of mentoring, clinical learning environments, and self-directed learning. The instrument development process includes measurements of perceptions and behaviours representing a specific construct; the construct validity testing is explained by confirming the measure's validity.

Psychometric testing is the evaluation of the quality of the instrument, including the reliability and validity. It is widely used in behavioural or social sciences to measure psychological and social phenomena while including variables as part of a broader theoretical framework (DeVellis, 2016). Health care education embraces complex cognitive, relational, emotional, and behavioural aspects, which benefit from psychometrically tested instruments to enhance a broader understanding of learning dynamics. Ensuring the psychometric testing of an instrument is an essential pathway to collecting valid and reliable results. In this article, we present the instrument development process phases, emphasizing the importance of a theoretical framework and the operationalization of a concept, validity types and phases, and reliability testing.

2. Instrument development process

2.1. The theoretical framework of an instrument

A common mistake observed in educational research in nursing and health care is a lack of theoretical framework development prior to developing an instrument. Instrument development is a long process and should not be taken lightly before clearly defining the research gap, evaluating already existing validated instruments and their functionality to address the research gap. When enough evidence has been collected on existing instruments and the theories behind those, instrument development needs to encompass a careful definition of concepts

* Corresponding author.
E-mail addresses: kristina.mikkonen@oulu.fi (K. Mikkonen), marco.tomietto@northumbria.ac.uk (M. Tomietto).

https://doi.org/10.1016/j.nedt.2022.105603
Received 21 June 2022; Received in revised form 28 September 2022; Accepted 7 October 2022
Available online 13 October 2022

describing the studied phenomenon. Instrument development can start from an inductive or deductive approach. In the inductive approach, the starting point is unstructured reality; concepts need to be further explored, described, and operationalized (Kyngäs, 2020; Mikkonen and Kyngäs, 2020). It can be done by conducting several qualitative studies, content analysis, systematic literature reviews and/or concept analyses. The process takes time and needs to be conducted carefully. When concepts are clearly defined, they can be operationalized into items measuring phenomena at a simple and clearly understandable level. In that way, the theoretical concepts can be operationalized and empirically tested. Instruments can also be developed deductively, taking the starting point of structured and clearly defined concepts taken from a theory or empirically tested models. There are developed instruments that are used in the exploration of theories and models, which have been used empirically to test the structure and connections of the concepts. Once researchers have defined their concepts and developed a theoretical stand explaining their studied phenomenon, items can be operationalized and further validated.

2.2. Validity: types and phases

The validity of measurement is the degree to which an instrument measures what it claims to be measuring (Rattray and Jones, 2007). Validity is a broad concept, and it can be estimated by a range of methods each if which contributes to our understanding of the validity of a measurement. This overview aims to present the most adopted pathways to ensure and test validity. Other types of validity exist, such as criterion validity. This type of validity tests a new scale or measure against a gold standard. A gold standard is supposed to measure precisely the same theoretical construct the new scale aims to measure (Pett et al., 2003). While criterion validity is well adopted in other fields (e.g., biological sciences), its adoption in nursing and behavioural sciences is limited due to the lack of gold standards (Polit, 2015; De Vet et al., 2011). Testing criterion validity also implies to increase the participants' burden in filling more scales and items, with a negative impact on response rate and the quality of data. In a sample of 105 nursing studies, an inappropriate claim of a criterion validity measurement has been previously identified (Polit, 2015). This paper focuses on instrument development and psychometric testing of new instruments in absence of a gold standard.

2.2.1. Face validity

Face validity is explained by the appearance of the attribute or construct found in the instrument that it is claiming to measure (DeVon et al., 2007). The aim when examining face validity in the instrument development is to investigate the cultural appropriateness, understanding of meanings, logical flow, grammar, and syntax of the newly developed items (DeVon et al., 2007). This can be accomplished by inviting 10–20 participants into a focus group (maximum 10 participants per group) to examine the face validity of the new items. All items need to be further modified according to the outcomes of the face validity evaluation. Face validity has been criticized as one of the subjective and least valid aspects used in empirical studies (DeVon et al., 2007). However, it is an important first phase of the validity and is recommended to be included together when conducting content validity.

2.2.2. Content validity

The consideration of validity needs to be additionally strengthened by content validity, which represents the relevance of sampling adequacy relating to the content of the instrument (Cook and Beckman, 2006; DeVon et al., 2007; Kimberlin and Winterstein, 2008). The aim of the consideration of content validity is to evaluate newly developed items for the appropriateness and relevance of the items representing sampling adequacy. This can be accomplished by inviting 20 experts to participate in the evaluation individually via email, using Lynn's (1986) Content Validity Index method (CVI). The outcomes of the evaluation process depend on the invited experts having previous or present experiences with studied phenomena (Grant and Davis, 1997; Kimberlin and Winterstein, 2008). In the invitation for expert evaluation, the following information needs to be provided for the purpose of understanding the meaning of the instrument modification: the purpose of the study and the instrument, conceptual basis for the instrument, conceptual definitions, characteristics of participants, the meaning of expert input in the evaluation, and instructions of the evaluation form (Grant and Davis, 1997).

The CVI (Lynn, 1986) can be conducted by providing a four-point rating scale (Davis, 1992) of each item developed in the instrument and counting the main score of the whole newly developed instrument. Every invited expert in the study will rate each individual item (I-CVI). The result of I-CVI rating will be counted by summing up the same scores of each item and dividing the sum by the number of experts giving that score. The score required to retain item is recommended to be $\geq$0.78. (DeVon et al., 2007; Polit et al., 2007). The quantification process of the whole instrument is based on the S-CVI/Ave averaging approach, which will be determined by computing I-CVI of each item of the instrument and taking the average score of I-CVI from all items. The recommended score of S-CVI to have excellent content validity should reach $\geq$0.90; for good content validity 0.70 to 0.80 (Grant and Davis, 1997; Polit et al., 2007). Experts can be offered a possibility to leave their open comments on each item, in case of possible anticipated suggestions on modifications. In case of the interrater agreement/S-CVI/Ave scoring lower than 0.70, and the need to modify certain items, another round of expert evaluation will be required.

After the completion of Content Validity Index quantification (Lynn, 1986) with experts, piloting of the self-assessment instrument in an online survey can be conducted with participants. The aim of piloting the instrument is to evaluate the practicality, understandability, and interpretations of the items; and to receive feedback from students about the technical functioning of the questions and the survey (Sue and Ritter, 2007). Participants can also assess the readability, length, wording, and clarity and how time-consuming it is to answer the survey. After the pilot study and final adjustments on the instrument need to be conducted prior to the main data collection.

2.2.3. Construct validity: testing the assumptions

Content validity is not recommended to be used alone in the validation process (Rattray and Jones, 2007), for which reason additional methods of internal consistency (reliability) and construct validity need to be employed. These approaches lead to statistical testing, and, in detail, they bring us into the psychometric field. Before proceeding with this type of validity, it is essential to be aware that the psychometric testing of an instruments' construct validity is based on solid assumptions. These assumptions are crucial to provide unbiased outcomes (Tabachnick and Fidell, 2001; Kline, 2015; Byrne, 2016). The main assumption is the multivariate normality of the distribution of the sample, which is examined by carefully identifying missing data and data outliers. Some authors reported that only 24.8 % of the papers report the assumptions to perform the test adopted, such as univariate or multivariate normality tests (Sajobi et al., 2018). Other authors highlighted a lack of information in reporting the structural equation modelling methods and results in research in nursing (Sharif et al., 2018). Authors from other disciplines also highlighted a failure to report missing data treatment as well as multivariate normality checks prior to performing multivariate statistics (Crede and Harms, 2019).

2.2.3.1. The first step to multivariate normality: missing data management. Missing data testing is essential to ensure unbiased outcomes when performing multivariate statistics. However, this preliminary test is rarely reported, despite the increasing use of multivariate statistics. Missingness in data distribution is typically checked with a test, referred to as Little's MCAR test, to discover if missing data are Missing

Completely at Random (MCAR) (Graham, 2009). If the test is non-significant, data are missing completely at random (MCAR). Therefore, when missing data are over 5–7 % in each record, the record can be deleted listwise, or imputation and estimation algorithms could be considered (Little et al., 2014). The main point in handling missing data is to detect the mechanism behind missingness. In detail, three situations are possible: when data are Missing At Random (MAR), missingness depends on observed data, not on unobserved data; when data are Missing Not At Random (MNAR), missingness is affected by unobserved data; while with Missing Completely At Random (MCAR) data's missingness does not depend on both observed and unobserved data (Graham, 2009). MAR data are considered missing "conditionally" at random as their missingness can be explained by other data and variables in the dataset. Data MCAR, on the other hand, are not affected at all by other observed or unobserved data, so everything that can be understood about the data depends only on the variables in the dataset, without any bias. Both MAR and MCAR missingness do not bias parameters' estimation in data analyses. MNAR missingness leads to biased estimations, and it invalidates data analyses' reliability and validity. Instead, the only consequence of MAR and MCAR missingness is the loss of statistical power. Listwise deletion should be considered with caution, even if it does not bias the distribution (Newman, 2014; Graham, 2009). A good practice is to perform a sensitivity analysis and to test the multivariate statistics with both missing values and without them: if the missing values do not affect the parameters' estimation and the fit indexes of the model, they could be retained to use all available data (Newman, 2014).

2.2.3.2. The second step to multivariate normality: outliers management. Outlier detection and management is also relevant to ensure unbiased outcomes. Outliers can be univariate or multivariate. Univariate outliers are usually identified by z values above 3 or below −3: z values stand for the standardized values of the distribution for each variable, and they are calculated as follows:

$$z = (x - \bar{x})/s$$

where x = measured value; $\bar{x}$ = mean value for the variable; s = standard deviation for the variable.

The detection of univariate outliers is a useful exercise, but it does not have a significant impact on the decision-making process when approaching multivariate analyses. Instead, it is useful to test the univariate normality by performing the Kolmogorov-Smirnov and/or the Shapiro-Wilk tests. If statistically significant ($p < 0.05$), these tests state the univariate normality. This is important, because univariate normality also leads, by definition, to multivariate normality and it is not necessary to test further the assumptions for multivariate normality. On the other hand, a distribution can still verify multivariate normality even if it is not normally distributed in the univariate tests (Enomoto et al., 2020; Tabachnick and Fidell, 2001).

The next step, if the univariate normality is not verified, is to check multivariate outliers. A multivariate outlier identifies an unusual combination of values between variables. They are detected by calculating the Mahalanobis distances and the probability for outliers in the chi-square distribution, where $\alpha = 0.001$ and the degrees of freedom equal the number of variables of the dataset. Therefore, the multivariate outliers are identified by p-values below 0.001 (Enomoto et al., 2020).

Once multivariate outliers are identified, they can be managed. The next decision-making node is to test multivariate normality.

2.2.3.3. Testing multivariate normality. Multivariate normality is tested by comparing Mardia's kurtosis coefficient with a threshold value. Mardia's kurtosis coefficient is defined by the mean of the squared Mahalanobis distances that have been previously calculated to detect multivariate outliers. The threshold value is v * (v + 2), where v is the number of variables of the model or the degrees of freedom. If Mardia's coefficient is below the threshold value, the multivariate normality is verified (Tabachnick and Fidell, 2001). The rate of multivariate outliers affects the achievement of multivariate normality. If the latter is not verified, consideration should be given to deleting multivariate outliers listwise. The deletion of multivariate outliers, even if decreasing the statistical power by affecting the sample size, is often a crucial step to achieving multivariate normality and providing the basic assumptions to perform unbiased multivariate statistics properly (Leys et al., 2019). Multivariate outliers compromise the linearity of data distribution and jeopardize the fit indexes of a Confirmatory Factor Analysis or a Structural Equation Model (Kline, 2015). After deleting the multivariate outliers, a new multivariate normality check needs to be performed. As for the missing data, it is recommended to keep track of the deleted records and report a sensitivity analysis, including outliers. It is also useful to explore the characteristics of the outliers by checking the differences with the normal distribution.

2.2.4. Construct validity: Exploratory Factor Analysis (EFA), Confirmatory Factor Analysis (CFA), and parameter estimation

Testing multivariate normality addresses the choice of parameters' estimation approach in CFA. Multivariate normality enables the adoption of the Full Implementation Maximum Likelihood (FIML) estimation that reduces the biases in fit indexes and parameters' estimation. The other option, if multivariate normality is not verified, is to adopt the Asymptotically Distribution Free (ADF) approach, but this estimation approach leads to a biased parameters' estimation, which affects the fit indexes (Byrne, 2016; Benson and Fleishman, 1994). The ADF's biases are preferred to those of the FIML approach if multivariate normality is not verified (Curran et al., 1996).

The quality of statistical testing is deeply rooted in the instrument's development phase, in designing the items, in ensuring content validity and in pilot testing the instrument (O'Leary-Kelly and Vokurka, 1998). Over the years, the focus of testing construct validity shifted to CFA: while Exploratory Factor Analysis (EFA) is useful to check the structure of an instrument without assuming its factorial structure, CFA assumes a factorial structure and it tests the empirical data against the hypothetical model (Kline, 2015). This approach is the most adopted and reported in the recent literature and it is based on the classical test theory developed to check the construct validity of a psychological test (Furr, 2021).

EFA is a helpful approach to detect the items' aggregation into factors, to measure the variance of the instrument in explaining a given phenomenon and to identify the cross-loadings across the items. In the instruments' development stage, EFA provides preliminary knowledge on factors. When developing a new instrument, specific items should be created that lead to a latent factor and, by definition, to a hypothetical factorial structure. When performing an EFA in educational research in nursing, factors are, most commonly, inter-correlated (factor correlation >0.20) and it is recommended to adopt an oblique rotation to properly calculate the items' loadings (e.g. Promax or Direct Oblimin). Furthermore, we recommend adopting the Principal Axis Factoring instead of the Principal Component Analysis: the former assumes a probabilistic approach and is oriented to identify the latent factors linked to the items, while the latter aims to deterministically elicit the maximum variance of the instrument, given a set of known factors (Pett et al., 2003).

In CFA, a model based on our theoretical knowledge of a construct is built. Then the observed variables (behaviours and perceptions as measured by the designed items) are linked to latent variables (the factors which constitute the construct), and this hypothetical structure is empirically tested against the data collected. The estimation of the model leads us to the parameters' estimation approaches previously described. This is why the preliminary analyses to check multivariate normality are so important. All our efforts are about ensuring an unbiased construct validity. After the model's parameters estimation is performed, the main step is to test the hypothetical model's fit with the empirical data. To do this, fit indices of the model are calculated (Byrne, 2016; Kline, 2015). Many fit indexes have been developed over the

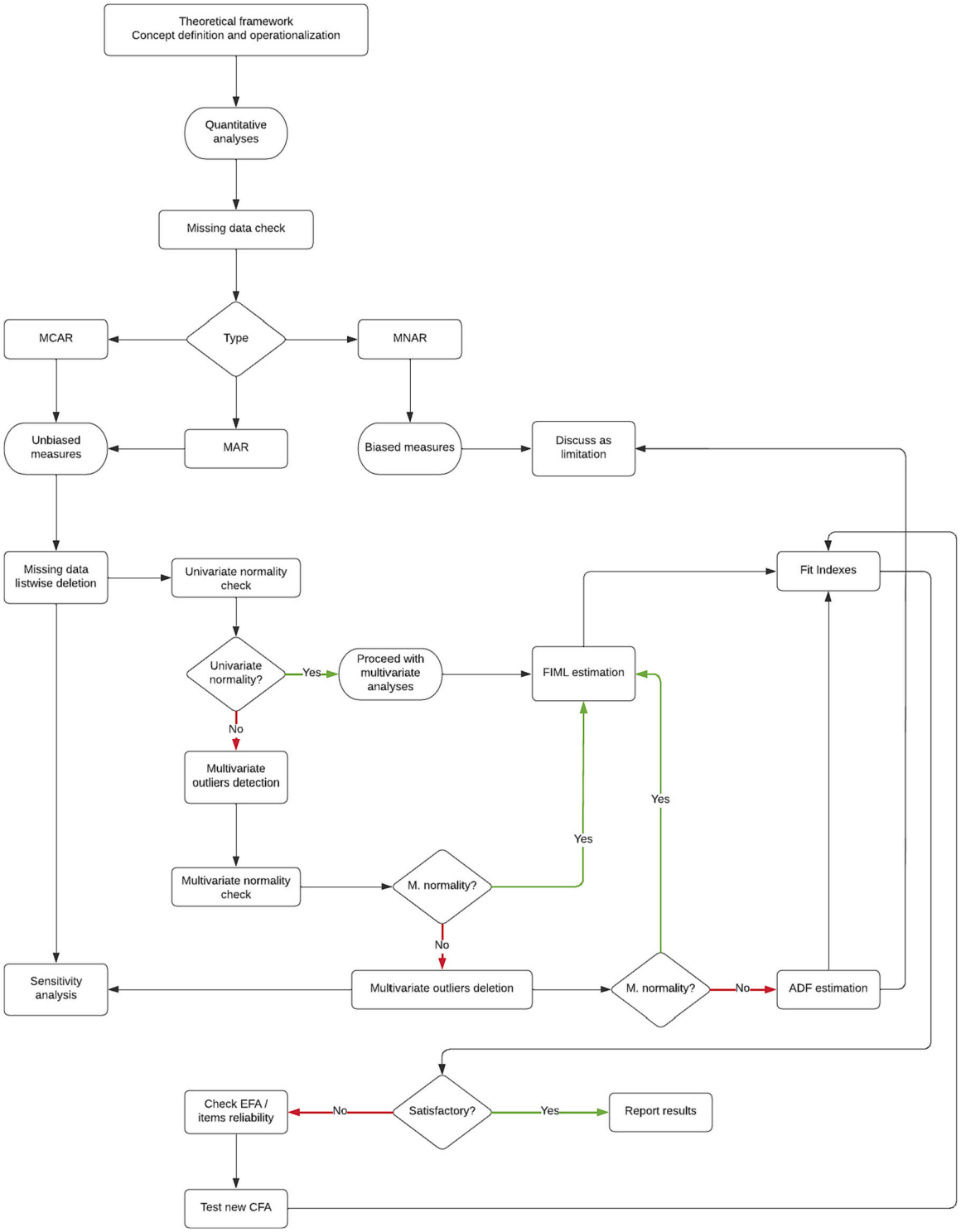


Fig. 1. The following flow chart represents the decision-making process described in this manuscript.

years; however, four fit indices are the most adopted and reported: Root Mean Square Error of Approximation (RMSEA), Standardized Root Mean Residual (SRMR), Comparative Fit Index (CFI), and Tucker-Lewis Index (TLI). A model is generally considered acceptable if the fit the between model and the dataset is described by RMSEA and SRMR <0.08, along with CFI and TLI >0.90 (Kline, 2015).

2.3. Reliability

The reliability involves accuracy, consistency, and reproducibility of the scores in the instrument measuring the attributes it is supposed to measure (Cook and Beckman, 2006; DeVon et al., 2007). The reliability testing of the instrument includes the internal consistency (DeVellis, 2016). Internal consistency reliability involves the computation of Cronbach's alpha coefficient by looking at what items relate to the same attribute/concept. This is conducted by looking at one set of items at one

time (Munro, 2005). Cronbach's alpha aims to measure a unidimensional construct (Heo et al., 2015; Waltz et al., 2010). It is important to remember that Cronbach's alpha represents the internal consistency of results across the instruments' items, but it is not independent of the number of items in the instrument and tends to be larger the greater the number of items in the dimension or scale being assessed (Sijtsma, 2009). For the purpose of measuring the external reliability, other measures should be adopted, such as the inter-rater reliability (e.g. Cohen's Kappa) or the test-retest reliability to assess the instruments' reliability over time (DeVellis, 2016). Another important point about Cronbach's alpha is that it is a unidimensional measure. This means that alpha values must be calculated for each factor of an instrument instead of on the entire instrument.

The result of high internal consistency of a single attribute demonstrates that items are measuring the same attribute, whereas the result of low internal consistency indicates a possibility that items are measuring more than one attribute (Cook and Beckman, 2006). The possible ranking of the reliability coefficient is from 0.00 to 1.00 (Kimberlin and Winterstein, 2008). The outcomes of Cronbach's alpha measurement are in accepted standard, if newly designed research tool scores $\geq$0.70, well-established instrument scores $\geq$0.80 and clinically reliable tool scores $\geq$0.90 (DeVon et al., 2007; Rattray and Jones, 2007). The new items of an instrument need to be possibly considered to be deleted/modified/corrected in case Cronbach's alpha score indicates low internal consistency (Rattray and Jones, 2007).

While excellent values are desirable, they also suggest that a shorter version of the instrument could also ensure excellent or good reliability. In this case, a short version has the benefit of decreasing the participants' burden and, potentially, increasing the response rate while keeping good reliability.

For this purpose, it could be useful to identify the contribution of each item to the overall internal consistency of the instrument. Alpha values can be calculated following the one-by-one deletion of items from each factor; an item should be removed from the instrument if the instrument's reliability increases over 0.10 (Ferketich, 1991). Corrected item-to-total correlation is also useful to identify the contribution of each item to the instrument or factor: it is considered acceptable if over 0.30 (DeVellis, 2016). When a "weak" item is identified, consideration should be given to deleting it and to further assess the instrument's validity and reliability.

These item-focused analyses provide useful information to further test the EFA and/or CFA by deleting specific items. In case of unsatisfactory fit indexes in the CFA, both these analyses and EFA could support the test of a different hypothetical model (see Fig. 1).

While Cronbach's alpha is a consolidated standard for testing reliability, its adoption has been widely debated as a single measure of reliability. Other measures are available but scarcely implemented in the most common statistical software, until recent years. There is a growing interest in the adoption of McDonald's omega as a reliability test (Hayes and Coutts, 2020) and new macros are available in SPSS or other statistical packages (e.g. R, MPlus or SAS) to calculate this measure. McDonald's omega is supported by a robust number of methodological reasons, which mainly rely on the consideration of item loadings from a CFA within the computation of reliability. A threshold for McDonald's omega is recommended to be >0.80. Recent research also demonstrated that, when applied to empirical data, McDonald's omega does not largely differ from Cronbach's alpha (Hayes and Coutts, 2020). An example in nursing seems to support this and it demonstrates the adoption of this measure in a scales' development and validation (Vélez-Morón et al., 2022). It is recommended to consider this reliability test along with Cronbach's alpha.

3. Conclusion

Instrument development and testing requires a rigorous and careful process of building a theoretical framework, defining the concepts, and operationalizing those concepts into simply understandable items. In nursing science, many instruments have been developed, commonly just for the purpose of one study. The sample size for an instruments' validation process needs to be ensured and preferably repeated by testing instruments' validity in different contexts with different sets of participants. We highly recommend that researchers consider the instrument development process carefully by preferably integrating instrument development guidelines from the first phase of instrument development (for example COSMIN by Mokkink et al., 2010).

CRediT authorship contribution statement

All authors contributed to study design, writing and revising of the manuscript.

Declaration of competing interest

Given their role as Editor on this journal, Professors Kristina Mikkonen and Marco Tomietto had no involvement in the peer-review of this article and has no access to information regarding its peer-review. Full responsibility for the editorial process for this article was delegated to an independent.

References

Benson, J., Fleishman, J.A., 1994. The robustness of maximum likelihood and distribution-free estimators to non-normality in confirmatory factor analysis. Qual. Quant. 28 (2), 117–136.

Byrne, B.M., 2016. Structural Equation Modeling With AMOS: Basic Concepts, Applications, and Programming, Third Edition, 3rd ed. Routledge. https://doi.org/10.4324/9781315757421.

Cook, D.A., Beckman, T.J., 2006. Current concepts in validity and reliability for psychometric instruments: theory and application. Am. J. Med. 119, 166.

Crede, M., Harms, P., 2019. Questionable research practices when using confirmatory factor analysis. J. Manag. Psychol. 34 (1), 18–30. https://doi.org/10.1108/JMP-06-2018-0272.

Curran, P.J., West, S.G., Finch, J.F., 1996. The robustness of test statistics to nonnormality and specification error in confirmatory factor analysis. Psychol. Methods 1 (1), 16.

Davis, L.L., 1992. Instrument review: getting the most form a panel of experts. Appl. Nurs. Res. 5, 194–197.

DeVellis, R., 2016. Scale Development. Theory and Applications, 3rd edition. Sage, Newbury Park, NT.

De Vet, H.C., Terwee, C.B., Mokkink, L.B., Knol, D.L., 2011. Measurement in Medicine: A Practical Guide. Cambridge University Press.

DeVon, H.A., Block, M.E., Moyle-Wright, P., Ernst, D.M., Hayden, S.J., Lazzara, D.J., Savoy, S., Kostas-Polston, E., 2007. A psychometric toolbox for testing validity and reliability. J. Nurs. Scholarsh. 39, 155–164.

Enomoto, R., Hanusz, Z., Hara, A., Seo, T., 2020. Multivariate normality test using normalizing transformation for Mardia's multivariate kurtosis. Commun.Stat.-Simul. Comput. 49 (3), 684–698.

Ferketich, S., 1991. Focus on psychometrics. Aspects of item analysis. Res.Nurs.Health 14 (2), 165–168.

Furr, R.M., 2021. Psychometrics: An Introduction. SAGE Publications.

Graham, J.W., 2009. Missing data analysis: making it work in the real world. Annu. Rev. Psychol. 60, 549–576.

Grant, J.S., Davis, L.L., 1997. Selection and use of content experts for instrument development. Res.Nurs.Health 20, 269–274.

Hayes, A.F., Coutts, J.J., 2020. Use omega rather than Cronbach's alpha for estimating reliability. But…. Commun. Methods Meas. 14 (1), 1–24.

Heo, M., Kim, N., Faith, M.S., 2015. Statistical power as a function of Cronbach alpha of instrument questionnaire items. BMC Med. Res. Methodol. 15 (1), 1–9.

Im, E.O., Meleis, A.I. (Eds.), 2021. Situation Specific Theories: Development, Utilization, and Evaluation in Nursing. Springer International Publishing.

Kimberlin, C.L., Winterstein, A.G., 2008. Validity and reliability of measurement instruments used in research. Am. J. Health Syst. Pharm. 65, 2276–2284.

Kline, R.B., 2015. Principles and Practice of Structural Equation Modeling. Guilford Publications.

Kyngäs, H., 2020. Theory development from the results of content analysis. In: Kyngäs, H., Mikkonen, K., Kääriäinen, M. (Eds.), The Application of Content Analysis in Nursing Science. Springer Publisher. https://doi.org/10.1007/978-3-030-30199-6.

Leys, C., Delacre, M., Mora, Y.L., Lakens, D., Ley, C., 2019. How to classify, detect, and manage univariate and multivariate outliers, with emphasis on pre-registration. Int. Rev.Soc.Psychol. 32 (1).

Little, T.D., Jorgensen, T.D., Lang, K.M., Moore, E.W.G., 2014. On the joys of missing data. J. Pediatr. Psychol. 39 (2), 151–162.

Lynn, M.R., 1986. Determination and quantification of content validity. Nurs. Res. 35, 382–385.
Mikkonen, K., Tomietto, M., Cicolini, G., Kaucic, B.M., Filej, B., Riklikiene, O., Kääriäinen, M., 2020. Development and testing of an evidence-based model of mentoring nursing students in clinical practice. Nurse Educ. Today 85, 104272.
Mikkonen, K., Kyngäs, H., 2020. Content analysis in systematic review. In: Kyngäs, H., Mikkonen, K., Kääriäinen, M. (Eds.), The Application of Content Analysis in Nursing Science. Springer Publisher.
Mokkink, L.W., Terwee, C.B., Knol, D.L., Stratford, P.W., Alonso, J., Patrick, D.L., Bouter, L.M., de Vet, H.C.W., 2010. The COSMIN checklist for evaluating the methodological quality of studies on measurement properties: a clarification of its content. BMC Med. Res. Methodol. 10, 22. https://doi.org/10.1186/1471-2288-10-22.
Munro, B.H., 2005. Statistical Methods for Health Care Research, 5th edition. Lippincott Williams & Wilkins, Philadelphia.
Newman, D.A., 2014. Missing data: five practical guidelines. Organ. Res. Methods 17 (4), 372–411.
Pett, M.A., Lackey, N.R., Sullivan, J.J., 2003. Making Sense of Factor Analysis: The Use of Factor Analysis for Instrument Development in Health Care Research. Sage.
Polit, D., Beck, C., Owen, S., 2007. Is the CVI an acceptable indicator of content validity? Res.Nurs.Health 30 (4), 459–467.
O'Leary-Kelly, S.W., Vokurka, R.J., 1998. The empirical assessment of construct validity. J. Oper. Manag. 16 (4), 387–405.
Polit, D.F., 2015. Assessing measurement in health: beyond reliability and validity. Int. J. Nurs. Stud. 52 (11), 1746–1753.
Rattray, J., Jones, M.C., 2007. Essential elements of questionnaire design and development. J. Clin. Nurs. 16, 234–243.
Sajobi, T.T., Brahmbatt, R., Lix, L.M., Zumbo, B.D., Sawatzky, R., 2018. Scoping review of response shift methods: current reporting practices and recommendations. Qual. Life Res. 27 (5), 1133–1146.
Sharif, S.P., Mostafiz, I., Guptan, V., 2018. A systematic review of structural equation modelling in nursing research. Nurse Res. 26 (2), 28–31.
Sijtsma, K., 2009. On the use, the misuse, and the very limited usefulness of Cronbach's alpha. Psychometrika 74, 107–120.
Sue, V.M., Ritter, L.A., 2007. Conducting Online Surveys. SAGE Publications, Los Angeles.
Tabachnick, B.G., Fidell, L.S., 2001. In: Cleaning Up Your Act: Screening Data Prior to Analysis. Using Multivariate Statistics, 5, pp. 61–116.
Tomietto, M., Paro, E., Sartori, R., Maricchio, R., Clarizia, L., De Lucia, P., PN Nursing Group, 2019. Work engagement and perceived work ability: an evidence-based model to enhance nurses' well-being. J. Adv. Nurs. 75 (9), 1933–1942.
Vélez-Morón, A., Andújar-Barroso, R.T., Allande-Cussó, R., García-Iglesias, J.J., Aquino-Cárdenas, G., Gómez-Salgado, J., 2022. Measuring anxiety and fear of Covid-19 among older people: psychometric properties of anxiety and fear of Covid-19 scale (AMICO) in Spain. BMC Public Health 22 (1), 1–11.
Vellone, E., Riegel, B., Alvaro, R., 2021. The situation-specific theory of caregiver contributions to heart failure self-care. In: Situation Specific Theories: Development, Utilization, and Evaluation in Nursing. Springer, Cham, pp. 193–206.
Waltz, C.F., Strickland, O.L., Lenz, E.R., 2010. Measurement in Nursing and Health Research, 4th edition. Springer Publishing Company, New York.