REVIEWS AND DISCUSSION PAPERS

Considerations when examining the psychometric properties of measurement instruments used in health

AUTHORS

DUNCAN McKECHNIE RN, BN(Hons), DipPublicSafety, GradCertRehabNurs, PhD[1]

MURRAY J FISHER RN, DipAppSc, BHSc, MHPEd, ITU Cert, PhD[2,3]

1. Royal Rehab, NSW, Australia
2. Susan Wakil School of Nursing and Midwifery, The University of Sydney, Australia
3. Nursing Scholar in residence, Royal Rehab, NSW, Australia

CORRESPONDING AUTHOR

DUNCAN McKECHNIE Royal Rehab, PO Box 6, Ryde, Sydney 1680 NSW Australia.
Email: duncan.mckechnie@royalrehab.com.au

ABSTRACT

Objective: To discuss and provide insights on how to critique the psychometric properties of measurement instruments used by nurses and nurse researchers.

Design and data sources: Methodological discussion paper that is based on our own experiences and research and is supported by literature.

Primary argument: Nurses routinely use a variety of measurement instruments during their everyday practice. They do so with the assumption that the instrument has been thoroughly validated and has been shown to be reliable. There is the real possibility that frontline nurses are using measurement instruments that have not been validated in their patient population or context. Critiquing the psychometric properties of measurement instruments is, however, particularly complex. Complicating matters, there are conflicting standards regarding the quality criteria needed for validating the psychometric properties of measurement instruments.

Conclusion: Nurses need to be aware of the limitations of the measurement instruments they use. Consequently, nurses need to have some understanding about the psychometric domains of measurement instruments and their associated measurement properties, as well as the quality criteria used for evaluating these measurement properties. Through discussing these aspects this paper aids frontline nurses and nurse researchers in critiquing the research literature regarding an instrument's psychometric properties. This paper equips nurses for making informed decisions when evaluating whether an instrument is suitable for use.

What is already known about the topic?
- There is the possibility that nurses are using measurement instruments that have not been validated in their patient population
- A measurement instrument's psychometric properties should be re-established when used in a different patient population or context to the index study

What this paper adds:
- This paper provides nurses and nurse researchers with the necessary information for critiquing the research literature regarding an instrument's psychometric properties
- This paper equips nurses to make informed decisions for effecting change in deciding whether an instrument is suitable for use

Key words: Measure, instrument, utility, validity, reliability, psychometric

INTRODUCTION

Measurement in research and health can be characterised as assigning a score to observations. In health, this is done in order to quantify a concept ([construct] e.g., falls or pressure injury risk) and in doing so, operationalise the construct.[1] Regardless of an instrument's purpose, however, it can be difficult for end users to quantify a measurement instrument's efficacy and in doing so, ascertain if they should use a particular instrument. This is because there are a variety of psychometric domains and associated measurement properties needing to be considered when determining the efficacy of an instrument. Complicating matters, inconsistent and ambiguous information can be found in the literature regarding the terminology and definitions surrounding the measurement properties of instruments. Inconsistent information can also be found in the literature regarding the various statistical tests and associated quality criteria used when examining these properties.[2] These inconstancies are a contemporary issue as 'every time a scale is used in a new context, or with a different group of people, it is necessary to re-establish its psychometric properties.'[3(p.161)] This is because an instrument's efficacy may be affected when used in a different patient context to the index study.[4] Instrument efficacy is an important consideration for nurse managers and clinical leaders who often implement instruments into clinical practice. There is the real possibility that frontline nurses are using measurement instruments that have not been validated in their patient population or context; which in part prompted this discussion paper.

BACKGROUND

This is the eighth paper in a series of articles surrounding methodological aspects of health research. The overarching aim of this series is to assist nurses in critiquing the research literature to support evidence-based practice. Previous papers in the series have focused on considerations for falls risk screening tool selection versus development,[5] research paradigms,[6] the research process,[7] quantitative research methods,[8] considerations when choosing a statistical method for data analysis,[9] conducting a critical review of the research literature,[10] and most recently, quality appraisal of the research literature.[11] In this paper we begin by providing some background context to psychometric properties of measurement instruments.

CLASSIFYING MEASUREMENT INSTRUMENTS USED IN HEALTH

When determining whether a measurement instrument is suitable for use, end users should first consider how the instrument is to be used, the concept to be measured and ultimately, how the instrument is classified.[12] This is because how an instrument is classified can impact on which psychometric properties should be considered and what quality appraisal guidelines should be used to assist with critically examining an instrument's measurement properties.[13-15] There are various ways to classify measurement instruments used in health. Some classification criteria include the purpose or function of the measure, scope of the measure (descriptive) and methodological (technical) aspects.[16] Having three broad categories, discriminative, predictive (diagnostic) and evaluative (assessment),[15,16] perhaps a classification system based on a measure's function is the most clinically meaningful for nurses.

LATENT VARIABLES AND CONSTRUCTS

A latent variable is a variable that cannot be directly observed such as pain. The presence of latent variables can be estimated through measuring their relationship with variables that can be observed.[17] Most constructs in research and health are made up of one or more latent variables. A primary goal of a measurement instrument used in health is to measure some underlying construct.[18] Consequently, a first step in choosing or developing a measurement instrument is understanding the construct being measured.[19] Identifying a construct's theoretical and empirical underpinnings is imperative here. This is because a construct needs to be systematically defined before it can be operationalised through the examination of measurable variables that when combined, quantify the construct.[1,17,20] For example, the construct or concept of falls risk cannot be directly observed but can be quantified by examining known patient characteristics that contribute to an increased risk for falls. In their prospective cohort study McKechnie, Fisher defined falls as an event which results in a person coming to rest inadvertently on the ground or floor or other lower level.[21] Using this definition, they collected data on observable variables (i.e., patient characteristics) at time of admission and first fall. The five resultant significant contributors to falls were then used to develop the Sydney Falls Risk Screening Tool through which the concept (construct) of falls risk is operationalised.

MEASUREMENT ERROR

Measurement error is the discrepancy between a measured value and the true or known value.[22] No instrument can be completely accurate; all instruments have some form of measurement error.[23] Measurement error can either be in the form of a systematic (constant, bias) error or random (chance) error.[23] Systematic error occurs when a measure has consistent scores but they are inaccurate.[19,20] For example, if you know your true weight is 75 kilograms but when repeatedly tested using the same instrument it is 76 kilograms, the one kilogram difference is a systematic measurement error. This can occur when an instrument is incorrectly calibrated. Random error occurs when: (i) the observer misreads or misinterprets the findings of an instrument;[20] (ii) there is a transient change in the

participant affecting the findings; or (iii) when there are instrument variations (intra-observer and inter-observer error).[23]

Classical measurement theory describes observed scores consisting of two components: a true fixed score and an unknown error variance being measurement error.[19,23] Consequently, the overarching aim of ensuring an instrument is valid and reliable is to minimise measurement error.[1] Measurement error can have an impact on a clinician's actions that were based on their interpretation of an instrument's finding which in turn, affects the instrument's clinical utility.[17]

AIM

To discuss and provide insights on how to critique the psychometric properties of measurement instruments used by nurses and nurse researchers.

DESIGN AND DATA SOURCES

Methodological discussion paper that is based on our own experiences and research and is supported by literature.

DISCUSSION

Determining whether a measurement instrument is suitable for use is not straightforward; there are several important considerations. Some include: (i) has the instrument been validated in the target population or context of intended use;[24] (ii) what is the instrument to be used for and how is it administered (i.e., screen, assessment or test); (iii) is the instrument in the public domain; (iv) how well has the concept (construct) been defined; (v) do the instrument's items reflect elements of the construct; and finally (vi) is there sufficient evidence for the instrument's psychometric properties.[1,23] Sample size of the index study and subsequent validation studies should also be considered. It has been recommended that a sample size of at least 400 subjects is required for precise estimates of reliability and validity coefficients,[25] however, this may depend on the number of items in the instrument.

PSYCHOMETRIC PROPERTIES OF MEASUREMENT INSTRUMENTS

There are three principal psychometric domains of measurement instruments needing consideration: reliability (i.e., consistency), validity (i.e., accuracy) and responsiveness (i.e., precision).[23,26] Within these three psychometric domains there are numerous measurement properties and associated statistical techniques needing consideration when examining the psychometric properties of an instrument. An instrument's interpretability and clinical utility (i.e., efficacy) are also a consideration.

The reliability and validity of an instrument are equally important.[1] By definition, however, an instrument can be deemed reliable but may not necessarily be valid and inversely, an instrument cannot be truly deemed valid unless it is reliable.[1,17,23,27] Nonetheless, validity and reliability are relative concepts; they are not all-or-nothing concepts being comprised of many measurement properties.[19,28] Figure 1 is a model that illustrates how the three principal psychometric domains of measurement instruments and their measurement properties relate, considering that an instrument used in health needs to have been shown to be reliable for it to be considered valid. In the model clinical utility is included as the outermost ring as it is an important overarching consideration when evaluating an instrument's efficacy; which is similar to the model for evaluating genetic tests in health.[29] This is because if a health-related instrument has no beneficial outcomes (such as, health, economic or clinical) the instrument's efficacy should be questioned, this is despite it being reliable and valid. Figure 2 shows the measurement properties represented in this model but as a general linear process illustrating when these measurement properties are a consideration during an instrument's development and validation.

There continues to be inconsistent and conflicting information in the literature regarding the terminology, definitions, various statistical tests and quality criteria surrounding the measurement properties of instruments. This is despite Mokkink, Terwee who used the Delphi technique to identify and define a taxonomy of measurement properties to be considered when evaluating health-related instruments.[26] For example, construct validity is described

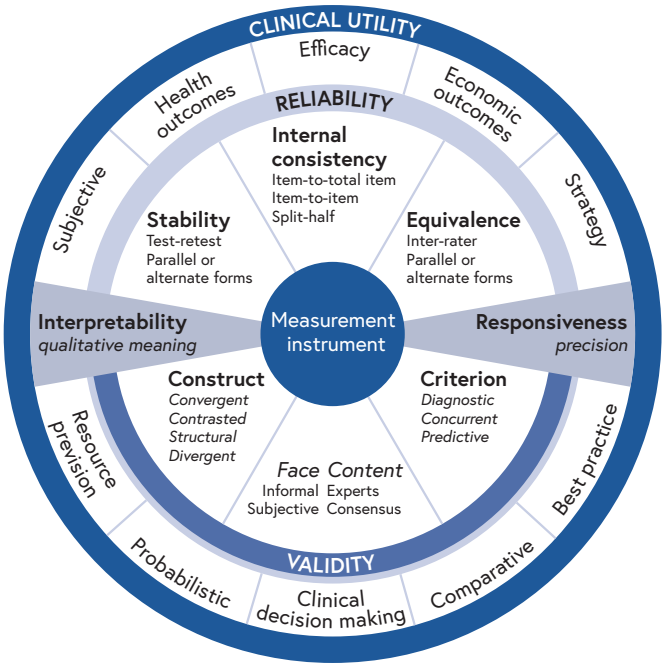

FIGURE 1: MODEL ILLUSTRATING DOMAINS OF MEASUREMENT INSTRUMENTS AND THEIR ASSOCIATED MEASUREMENT PROPERTIES

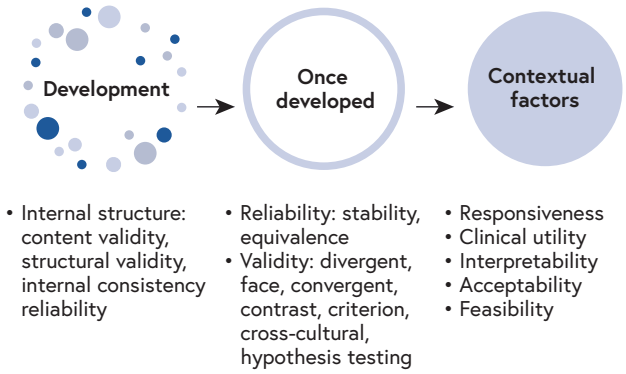

FIGURE 2: A GENERAL LINEAR PROCESS ILLUSTRATING WHEN AN INSTRUMENT'S MEASUREMENT PROPERTIES ARE A CONSIDERATION DURING AN INSTRUMENT'S DEVELOPMENT AND VALIDATION

by some authors as essential to the overall validity of any instrument and as such, overarches the other measurement properties of validity being translation (how accurately a construct has been operationalised) and criterion (the extent to which an instrument's score correlates with an external criterion) validity.[1,16,20,30] Other authors,[23,28] however, describe the measurement properties of validity as a tripartite model with construct, criterion and translation (face and content) validity being equally important. While some authors use equivalence to describe a defined measurement property of reliability,[23,31] other authors use measurement error,[26] which serves as another inconsistency. In this paper we have used the most widely referred to terminology in the research literature to describe the measurement properties of an instrument. When there is more than one term often used in the literature, we have made reference to both.

RELIABILITY

Reliability broadly refers to the consistency or stability of the measurement process across time, patients or observers (e.g., nurses).[16,31] Reliability estimates change when an instrument is administered by different users or when used in a variety of contexts or situations.[20,22] Being a function of measurement error, in classical test theory an instrument is said to be reliable when it measures the same thing twice and the same results are obtained and therefore, is free of random measurement error.[17,26] However, no measure can be perfectly reliable.[19] This is because reliability is not a unitary concept but rather an approximated concept that is determined through the evaluation of three underlying instrument properties: internal consistency, stability and equivalence.[17,28,31] The relative importance of these measurement properties, and therefore whether they are used to approximate reliability, depends on how the instrument is administered, who is to use the instrument, and in what context and patient population is it administered.[19,28] Reliability correlation coefficients are often used to approximate an instrument's reliability. In this instance, the reliability coefficient statistic indicates the extent to which the individual items of a scale are related and show the ratio between true score variance and observed score variance.[3] For instance, a reliability coefficient of 0.85 indicates that 15% of the observed variance is due to random measurement error.[16]

Internal reliability

Internal reliability is concerned with the extent to which an instrument is consistent within itself, evidenced by providing consistent results. Internal consistency or homogeneity reflects the extent to which the individual items within an instrument are interrelated and unidimensional in measuring the one domain or construct.[27,28] Unidimensionality of multidimensional scales is also a consideration.[18] For example, in an internally consistent unidimensional instrument each item equally contributes to the total score of the instrument and in the case of multidimensional scales, each subscale should measure different but related constructs.[23] Internal consistency can be measured by comparing the correlations amongst items in an instrument using techniques such as item-to-item and item-to-total item correlations, and the Split-half technique. While Cronbach's alpha is the most widely reported coefficient for internal consistency, it is known for underestimating true reliability.[32] McDonald's omega and Revelle's beta are alternatives statistical technique worth considering here.[32-34] Nonetheless, Zinbarg, Revelle advises that 'choosing among these [three] indices should be based on one's research question and considerations of dimensionality and equality of general factor loadings rather than which index is largest.'[35(p.132)] Quality criteria for Cronbach's alpha and other statistical techniques commonly used are provided in table 1.

External reliability

External reliability is concerned with measuring the extent to which an instrument varies when used at different times (stability) and by different observers (equivalence).[30]

Stability

Stability or repeatability reliability refers to the stability of a measure over time which is examined using the test-retest statistical technique.[20] The test-retest technique may not be suitable for phenomena that are likely to change over time, such as falls risk.[30] This technique examines whether a self-rated (intra-rater) or observer rated (intra-observer) measure produces constant results when used by/on the same patient under similar circumstances but taken/administered at two different points in time.[27] This measures the degree of random measurement error an instrument might have when the measure is repeated.[36] The test-retest intervals should be far enough apart to mitigate the effects of fatigue or patient learning, but close enough to avoid genuine changes in the construct being measured;[19] **two to 14 days is usually adequate.**[3,37]

Intra-class correlation coefficient (ICC) is the most commonly used indices to measure consistency of scores when continuous measures are used.[19,24,38] See table 1 for ICC criteria when examining stability. ICC is now preferred over Pearson and rank-order (Spearman) correlation coefficients as these statistical techniques can exaggerate the impression of reliability and measure the strength of a relation between two variables not the agreement between them.[3,16,39] Being a product–moment correlation, the tetrachoric correlation can also be used which describes the linear relation between two continuous variables that have both been measured on a dichotomous scale.[40] Parallel and alternate forms can also be used to approximate stability which are discussed next under equivalence.[23]

Equivalence

Equivalence, equal in value or worth, is concerned with the agreement or consistency among observers who use the same instrument and when altered forms of an instrument are used.[23]

Inter-rater reliability: Inter-rater or inter-observer reliability involves examining the strength of agreement or level of consensus between two or more observers (raters) when they are observing the same variable and rating their observations using the same instrument.[27] Percent agreement, tetrachoric correlation or Kappa (see table 1) are used to measure inter-rater reliability of categorical and nominal data.[23,38] Cohen's weighted kappa is commonly used for comparing agreement between two raters while Fleiss' kappa is used when there are more than two raters.[1,28] Modified Kappa, which is an 'index of agreement among experts that indicates beyond chance that the item is relevant', can also be used when there are two or more raters.[41(p.1276)] For measuring relationships between scores, Pearson's correlation coefficient can be used for continuous data and Spearman's rank correlation for ordinal data. Information regarding agreement and bias amongst raters can be evaluated by mean differences and confidence intervals.[28] Finally, ICC can be used to examine intra-class and inter-rater reliabilities (i.e., strength of agreement) when examining interval-level data.[23,42,43] In this instance, the ICC

TABLE 1: MEASUREMENT PROPERTIES FOR RELIABILITY AND COMMONLY USED STATISTICAL TESTS AND CRITERIA

Measurement properties & their aspects	Definition/purpose	Statistical test and criteria	Considerations
Reliability: consistent results and stability of an instrument			
Internal consistency	Extent to which items within an instrument are unidimensional and interrelated in measuring the same construct (aka homogeneity)	• Split-half & Cronbach's a (scale randomly split): <0.70 = low/ inadequate; 0.70-0.80 = adequate; ≥0.8 = desired/excellent[2] BUT >0.70 generally interpreted as sufficient[3,24,27,45,46] AND EFA or CFA performed with adequate sample size[24,33] • Item-to-total (omitting one item), Pearson correlation: >0.2[3] to >0.3[47] considered as adequate • inter-item correlation: coefficients ranging from 0.15 to 0.50 is considered as acceptable[18]	• A Cronbach's a statistic >0.90 may indicate redundancy of one or more items which could be removed without affecting the scale's reliability[3,48] BUT >0.95 has also been recommended for this.[23,24] However, Bland and Altman[49] recommends that for research purposes Cronbach's a statistic >0.80 is suitable and for clinical practice >0.90 is acceptable but >0.95 is desired. For psychological or achievement tests an a >0.80 is desirable[50] • The greater the number of items in a scale, the higher the Cronbach's a coefficient tends to be[1,16] • Interpretation: reliability coefficient <0.5 = high and >0.9 low measurement error[23] • Sample size: recommendations vary from 4 to 10 subjects per variable but >100 subjects is needed to ensure stability for factor analysis[24] and total number of subjects should exceed total number of variables by 50[30]
Stability test-retest	• The ability of a measure to produce the same results when used at two different points in time (aka repeatability reliability)	ICC, weighted Kappa & Tetrachoric correlation coefficient: generally 0.01-0.20 = slight agreement; 0.21-0.50 = fair/poor; 0.50-0.74 = moderate; 0.75-0.89 = strong; >0.90 = excellent agreement/reliability[19,38,51] BUT for most purposes <0.40 = poor; 0.40-0.59 = fair; 0.60-0.74 = good; ≥0.75 = excellent agreement[52,53]	• ICC criteria is arbitrary as acceptable reliability is a judgement call dependent on how the instrument and what test statistic is used[16,19,20,52] • Terwee, Bot[24] recommends 0.70 should be considered as minimum criteria for measures of external reliability in a sample size of 50 patients. For clinical measurements and ongoing progress measurement in treatment situations the ICC test-retest for reliability should exceed 0.90[54] • Must report what ICC is measuring (consistency or agreement) and classification of ICC (model and form) used[24,43,55] • The magnitude of both Pearson's correlations and ICCs can be influenced by the range of scores and presence of extreme values[56] • Salkind[27] comments that for inter-rater Kappa nothing less than 90% agreement should be accepted
Equivalence inter-rater	• Examining whether the same results are obtained when two or more observers use the same instrument		
Parallel/ alternate forms	• Examines results consistency when altered versions of the same instrument or alternate instruments are used to measure the same construct		

Abbreviations: ICC, intra-class correlation coefficient; EFA, exploratory factor analysis; CFA, confirmatory factor analysis.

measures the proportion of variance of an observation that is the result of between and within subject variance in the true scores.[28,44]

Parallel and **alternate forms** reliability: parallel forms of reliability are applicable when scale items from the same 'pool' are used to develop two differing instruments.[23,30] Alternate forms is applicable when two versions of the same measuring instrument are used to measure the same construct.[19,20,27] The equivalence between the two instruments is examined to determine which set of questions or instrument is best to use. Correlation coefficients and Student's *t*-test can be used for this.[19]

Summary of reliability evidence

- Internal consistency: same population, individual instrument items.
- Stability (test-retest): same observer (intra-rater) or same patient (intra-individual), same instrument, different times.
- Equivalence (inter-rater): different observers, same population, same instrument, same time.
- Equivalence (parallel/alternate forms): same observer, same population, different instruments or forms, same time.

VALIDITY

Validity broadly refers to the accuracy of an instrument and is subsequently concerned with systematic measurement error.[19] A measurement instrument's validity is the extent to which it is estimated to have correctly measured a construct (e.g., fall risk) it purports to measure within the context of which it was used.[22,26,28] This estimate is based on inferences made in determining whether the results of an instrument are accurate.[17] Therefore, validity is different to reliability in that, validity is not a property of the instrument itself (i.e., the item inventory) but rather inferences or interpretations of the test score that should be contextually relevant, meaningful and useful.[1,28] Consequently, an instrument's validity can only truly be evaluated within the context, circumstances and population of intended use.[3,19] Measurement properties of validity include face, content, construct and criterion validity. During instrument development the items used need to be based on some theoretical underpinnings (content validity) after which the instrument's construct (construct validity) can be examined and finally, the instrument can then be compared to an external criterion (criterion validity) (see figures 1 and 2).[16]

Face validity

Face validity is primarily concerned with whether an instrument superficially appears to, or 'looks like' it is going to, measure the concept (construct) it purports to measure.[19,26] That is, the operationalisation of a construct.[20]

While face and content validity are closely related, having face validity does not guarantee that the item inventory of an instrument represents the theoretical domain of the construct (i.e., content validity). Face validity is more concerned with the language/syntax of individual items and the flow/organisation of an instrument in its entirety.[30] Examining an instrument's face validity is a subjective assessment after an instrument is developed, usually by those who administer it. Consequently, face validity is the weakest measurement property of validity.[19,30]

Content validity

Content validity relates to how well an instrument's content domain (item inventory) truly represents the theoretical domain (e.g., what it means to be at risk of falls) of the latent construct, such as falls risk, it purports to measure in the context of its intended use.[22,31] More simply, how accurately a construct has been operationalised while still representing the construct that is being measured.[20,28] Content validity is more a qualitative judgement opposed to an absolute value.[17] The overall goal of examining content validity is to remove redundant items of an instrument so a minimal number of items remain while still defining and operationalising the construct, and retaining face validity.[23,28] Content validity generally evolves out of the process of planning and developing an instrument. For instance, a systemic review of the research literature can assist with initially generating an instrument's items. Following this, subject matter experts (such as, patients, researchers or clinicians) are often engaged through focus groups, questionnaires, interviews or the Delphi technique to generate new items and review existing ones.[1,19,28] The RAND/UCLA disagreement index can be used to quantify agreement amongst raters in Delphi studies and the Content Validity Index can be used when multi-item scales are being rated (see table 2).[19,41,57]

The series of studies used to develop the Sydney Falls Risk Screening Tool is a good example of research techniques that could be used to ensure the face and content validity of an instrument. In the first instance an integrative review that used a systematic search strategy to identify quantitative papers was undertaken.[58] This was followed by a retrospective nonequivalent case-control study that further described the characteristics of patients who fall in the target patient population and context.[59] The patent characteristics identified in the review and case-control study formed the basis for the items included in a modified Delphi study.[57] Participants had the ability to add additional patient characteristics that they believed contributed to falls in questionnaire round one. The resultant list of patient characteristics were then examined in a multisite prospective cohort study from which the five-item Sydney Falls Risk Screening Tool was developed.[21]

Construct validity

Construct validity refers to whether an instrument measures an underlying construct, as defined by theory, which it purports to measure based on the established variables that define the construct.[1,27] Exploratory and confirmatory factor analysis techniques are a consideration here as these statistical techniques can be used to examine dimensionality.[32] Exploratory factor analysis (EFA) can be used to examine linear relationships of underlying factors that explain a construct.[19] This is important for item pruning, when introducing new items and when evaluating revised instruments (structural validity).[32] With EFA there are no prior assumptions regarding the relationships between factors and as such, EFA can be thought of as theory-generating.[19] Conversely, confirmatory factor analysis (CFA) can be thought of as theory-testing as it is used to examine if sample data fit a previously determined factor structure of a construct.[19,24] Consequently, with CFA a model is proposed at the outset and the hypothesis, number of factors encountered and which variables should load onto each factor are all known.[19]

Exploratory Structural Equation Modeling (ESEM) is another consideration during instrument development and validation.[28] ESEM incorporates some advantages of the less restrictive EFA (i.e., allowing cross-loadings) and advanced aspects of CFA (i.e., goodness-of-fit or multi-group models) creating a synergy between the two.[60] Measurement properties for and associated quality criteria for construct validity, and associated statistical techniques for examining construct validity, are provided in table 2.

Criterion validity

Criterion validity of an instrument examines the extent to which an instrument's score correlates with some external criterion. A key factor in establishing criterion validity is the quality of the criterion.[27] Consequently, criterion validity, as well as convergent and divergent (discriminative) validity, can be difficult to evaluate when there is no 'gold standard' measure.[2] In this case, these aspects of validity can 'represent a form of construct validity in which the relationship to another measure is hypothesised' or a well-established reference standard may be used.[19, 36(p.194)]

The criterion can either be another instrument, a discrete but related variable or an outcome that will occur in the future. Concurrent validity is studied when two instruments, one new and one the criterion (accepted reference standard), examine the same concept at the same time while attempting to identify the same existing condition.[1,20,26] Predictive validity explores how well a measure of a construct predicts some future criterion being an outcome score, event or behaviour.[16,23,28] For example, the ability of a falls risk screening tool to predict patients who end up falling.

The primary criterion-related evidence in health is an instrument's classification or discriminatory accuracy in differentiating patients with and without a specific condition, that is, its diagnostic validity.[19] Consequently, measures of sensitivity and specificity are used as criterion-related evidence for validity (see table 2).[28] Correlation coefficients, regression modelling and the ICC can also be used to examine criterion validity.[19]

RESPONSIVENESS

An instrument's ability to detect or track clinically meaningful patient change over time (such as, between hospital admission and discharge) in a discrete patient condition is of interest in health.[2,24,26] This relates to an instrument's responsiveness, that is, its precision.[36] While an instrument's responsiveness is a relative concept being dependent on its reliability and validity, instrument responsiveness is a consideration when deciding whether it should be implemented into clinical practice.

Husted, Cook describe responsiveness as having two aspects.[72] Firstly, internal responsiveness which relates to the ability of an instrument to measure change over a pre-determined period of time, or change before and after a treatment with a known effect.[72] An instrument's internal responsiveness can be measured using student's paired *t*-test and measures of effect size (i.e., magnitude of difference/change/effects between treatments or relationship between variables) (see table 3). External responsiveness is the same as internal responsiveness but is used when the responsiveness of two instruments are compared, usually a new instrument against an external standard or reference instrument.[72(p.459)] Measures of sensitivity and specificity, Pearson's correlation coefficient and regression modelling can be used to examine external responsiveness as these statistical methods indicate how change in two measures vary together.[72]

Of consideration when examining an instrument's responsiveness is its floor and ceiling effects 'as they indicate limits to the range of detectable change beyond which no further improvement or deterioration can be noted.'[36(p.194)] Floor and ceiling effects of an instrument are measured by examining the proportion of patients that achieve the lowest and highest possible score.[24] Patient characteristics, such as age and acuity, can have confounding effects on the floor and ceiling attributes of an instrument and therefore need to be considered.[73]

INTERPRETABILITY

By providing useful and informative information, a measurement instrument can contribute to a clinician's decision making in identifying patients who will and will not benefit from particular actions.[79] This information needs to be contextually relevant to be of worth; that is, have clinical meaningfulness.[79,80] Interpretability of information provided

TABLE 2: MEASUREMENT PROPERTIES FOR VALIDITY AND COMMONLY USED STATISTICAL TESTS AND CRITERIA

Measurement properties & their aspects	Definition/purpose	Statistical test and criteria	Considerations
Validity: accuracy of an instrument			
Face & content validity: how accurately a construct has been operationalised while still reflecting the construct being measured			
Face	• Assesses whether an instrument superficially appears to measure the concept it purports to measure	• Item-level content validity index (I-CVI): Lynn[61] average criteria = 1.00 with 3 to 5 raters; >0.78 for 6 to 10 raters[62] AND • Scale-level content validity index (S-CVI): >0.90 = acceptable agreement[63,64] but other authors[65] recommend ≥0.80 agreement is acceptable • All items are relevant to: (i) measurement aim; (ii) the construct being measured; (iii) the target population; (iv) the context of use; (v) together comprehensively reflect the construct to be measured; AND (vi) investigators or experts were involved in item selection[24,45]	• Lynn[61] suggests a minimum of 3 experts and more than 10 is probably unnecessary for I-CVI analysis; 2 raters are used in S-VCI analysis • Lynn[61] suggests a 4-point ordinal scale be used to prevent an ambivalent midpoint for CVI analysis where 1 = irrelevant and 4 = extremely relevant. The CVI of a scale item is the proportion of experts who rate the item as a 3 or 4.
Content	• Examines the relevance of individual items in an instrument		
Construct validity: a measure of the underlying construct based on the established variables that define the construct			
Structural	• The degree to which an instrument's items and scores reflect the dimensionality of the construct being measured so facilitating the removal of redundant items (pruning).	• Classical test theory, item response theory and Rasch modeling[66]: EFA/CFA[19,45], ESEM[28,60], CFI, TLI[45,67]	• EFA loadings: >0.32 = poor; 0.45 = fair; 0.55 = good; 0.63 = very good; >0.70 = excellent.[68] These criteria are considered a rough guideline for CFA studies[69] • Generally, factor loadings >0.40 are considered meaningful[30]; some researchers use >0.30[19] • Cutoff values for acceptable fit between the hypothesised model and observed continuous data: CFI/TLI = close to 0.95 or higher; SRMR = close to 0.08 or lower; RMSEA = close to 0.06 or lower[70] • Correlation statistics quantify the association between two measurement instruments and indicate how accurately one rating can be predicted from another, they do not indicate agreement[16] • Correlation criteria depends on the measurement property and statistic used[16] • Salkind 27 suggest general correlation criteria as: <0.20 = extremely poor; 0.21-0.40 = week; 0.41-0.60 = moderate; 0.61-0.80 = strong; >0.81 very strong
Contrasted groups	• To examine contrasting groups where one group is expected to score high and one low (aka extreme groups)	• ANOVA, Student's t-test and regression[19]	
Convergent	• The extent to which the similarities (sensitivity) between two instruments measure the same construct. High correlations are expected.	• Correlation statistics: ≤0.40 = poor; 0.41-0.75 = adequate; ≥75 = excellent[2]; DeVon, Block[30] recommends ≥0.45 analysis.	
Divergent (discriminate)	• To examine the differences (specificity) between two instruments that measure the same construct and that conforms to a priori hypotheses. Low correlations are expected.	• Correlation statistics: DeVon, Block[30] recommends ≤0.45.	
Hypothesis testing	• To examine if an instrument measures the construct being based on the theoretical underpinnings used to develop the instrument	• Formal and discrete hypothesis formulated AND at least 75% of the results are in accordance with the hypotheses[24]	
Cross-cultural	• The degree to which the performance of the items on a translated or culturally adapted instrument reflect the original version	• Differential item functioning AND multiple group factor analysis[45,67]	
Criterion validity: extent to which an instrument's score correlates with an external criterion			
Concurrent	• Examines the correlation between a new and a validated instrument at the same time	1. Convincing evidence that the gold standard instrument has been identified[24] AND 2. AUC: AUC generic standard: ≤0.5 = no discrimination; >0.5-0.7 = poor; ≥0.7-0.8 = acceptable; ≥0.8-0.9 excellent; ≥0.90 = outstanding discrimination[71] BUT for criterion validity: ≥0.70 is acceptable[67]; OR 3. Correlation with gold standard: ≥0.70 is acceptable[24]	• Sensitivity and specificity criteria are arbitrary as the chosen criteria depends on how the measure is to be used, what is being measured and in what context • Consider other sensitivity and specificity statistics: PPV, NPV, DOR and LR[16,28]
Predictive	• Examines the ability of an instrument to predict some future criterion		

Abbreviations: AUC, area under the curve; PPV, positive predictive value; NPV, negative predictive value; DOR, diagnostic odds ratio; LR, likelihood ratio; EFA, exploratory factor analysis; CFA, confirmatory factor analysis; ESEM, Exploratory Structural Equation Modeling; CFI, Comparative Fit Index; SRMR, standardised root mean squared residual; RMSEA, root mean squared error of approximation; TLI, Tucker-Lewis index.

TABLE 3: MEASUREMENT PROPERTIES FOR RESPONSIVENESS AND COMMONLY USED STATISTICAL TESTS AND CRITERIA

Domain & measurement properties	Definition/purpose	Statistical test and criteria	Considerations
Responsiveness (precision): an instrument's ability to detect clinically important change over time in the construct to be measured			
Internal	An instrument's ability to measure change over a pre-determined period	1. Effect size: standardised effect size, standardised response mean (SRM), Guyatt's Index benchmark values: 0.20 = small; 0.50 = moderate; 0.80 = large responsiveness[16,73] 2. Effect size: Pearson's correlation & Chi-square/ ANOVA/multiple regression: small = 0.10/0.10/0.20; medium = 0.30/0.25/0.15; large = 0.50/0.40/0.35, respectively[73,74] 3. Student's paired t-test[2,16]	• SRM an estimate of change in the measure relative to the between patient variability in change scores • Guyatt's index examines repeated observations of the measure in clinically stable subjects providing information on MIC change on the measure[71] • t-test used to test the hypothesis that there was no statistically significant change in the average response on the measure over the two time points
External	A comparison of two instrument's responsiveness	1. AUC: AUC generic standards as per table 2 BUT for measuring responsiveness ≥0.70 is acceptable regarding MIC and MDC[24,67] 2. Pearson correlation coefficient 3. Regression models	
Floor/ceiling effects	An instruments limits in detecting change beyond which no further improvement or deterioration can be noted	• ≤15% = adequate for number of respondents either achieving the minimum (floor) or maximum (ceiling) score[75] with a sample size >50 participants[24]	Other authors recommend[2,36] or have used ≤20% in their studies[76,77]

Abbreviations: ANOVA, analysis of variance; MIC, minimal important change; MDC, minimal detectable change; AUC, area under the curve.

by an instrument is a consideration here. Mokkink, Terwee describe interpretability as, 'the degree to which one can assign qualitative meaning – that is, clinical or commonly understood connotations – to an instrument's quantitative scores or change in scores.'[26(p.743)] That is, how clinically meaningful is an instrument's resultant score and are there consistent definitions and classifications for interpreting the results.[36] Some authors suggest that interpretability should be one of the principal psychometric domains (i.e., validity, reliability, responsiveness) needing consideration when deciding whether an instrument should be used in health.[24,26,36] Information in the index study, such as mean, SD and minimal important change, can aid in interpreting the clinical meaningfulness of a measurement instrument's score.[24] Interpretability is an important foundation for an instrument's clinical utility.

CLINICAL UTILITY

Utility is generally described as a subjective outcome measure of satisfaction regarding how beneficial an action, intervention, product, or process is.[80] The clinical utility of a measurement instrument refers to what extent the instrument contributes to beneficial health outcomes relative to best practice alternatives.[79] These beneficial outcomes are not solely patient dependent but multidimensional that could include economic, clinical, administrative and subjective domains.[36,80] Acceptability (respondent burden [the patient]) and feasibility (administrative burden [effort, time, expense, disruption]) of a patient or staff completing an instrument are important considerations here.[2,36]

As clinical utility is not a measurement property of an instrument, instruments do not have clinical utility per se; it is best-practice outcomes that determines the clinical utility of an instrument.[81] The clinical utility of an instrument is, however, dependent on its reliability, validity and responsiveness. For instance, clinical utility of a falls risk screening tool is dependent on its predictive validity resulting in appropriate allocation of resources for preventing falls. Bossuyt, Reitsma describes clinical utility of health-related diagnostic tests as having four key elements:[79] (i) individual and societal health outcomes; (ii) probabilistic (reliability as measured by diversity of outcomes); (iii) strategy being management strategy for testing and subsequent clinical actions; and (iv) comparative being relative to some comparator strategy, best practice or clinical actions.

Clinical utility is rarely quantified as it is a subjective concept and is context dependent.[80] There are, however, some clinical utility indices that can be used to quantify the expected gain in the utility of a test.[82] These indices can be used for tests 'if the clinical situation can be described by pre-test probability of disease and the ratio of the cost of erroneously treating individuals without the disease to the net benefit of correctly treating individuals with the disease.'[82(p.564)]

IMPLICATIONS FOR RESEARCH, POLICY AND PRACTICE

Nurse researchers and frontline nurses use a variety of measurement instruments during their everyday practice. They do so with the assumption that the instrument has been rigorously validated and has been shown to be reliable. However, determining whether an instrument is valid and reliable is complex. Nonetheless, nurses need to question the efficacy of the instruments they use. This is because there is the real possibility that nurses are using measurement instruments that have poor clinical utility. The discussion in this paper can enable frontline nurses and nurse researchers to confidently critique the research literature regarding an instrument's psychometric properties and therefore, make informed decisions regarding whether it is suitable for use. Some recent studies that have examined the psychometric properties of health-related measurement instruments is provided in supplementary file 1. These papers provide some working examples describing what nurses should consider when critiquing the literature regarding the efficacy of a measurement instrument.

CONCLUSION

Fundamentally, nurses are caught between the clinical need to use an array of measurement instruments and the availability of contextually validated ones; falls risk screening tools are an example of this. Consequently, nurses need to be aware of the limitations of the measurement instruments they use. However, there continues to be inconsistent and conflicting information regarding the terminology and quality criteria surrounding the psychometric measurement properties of measurement instruments. These properties and their associated quality criteria have been discussed in this paper. This paper empowers nurses to confidently question the suitability of measurement instruments they may use.

Disclosure of funding: The authors report that there was no funding for all or any part of this study.

Declaration of conflict of interest: The authors report no conflicts of interest.

REFERENCES

1. Kimberlin CL, Winterstein AG. Validity and reliability of measurement instruments used in research. *Am J Health Syst Pharm.* 2008; 65: 2276-84. Available from: https://doi.org/10.2146/ajhp070364
2. Andresen EM. Criteria for assessing the tools of disability outcomes research. *Arch Phys Med Rehabil.* 2000;81:S15-S20. Available from: https://doi.org/10.1053/apmr.2000.20619
3. Streiner DL, Norman GR. Health measurement scales: a practical guide to their development and use. 2nd ed. Oxford, UK: Oxford University Press, 2001.
4. Aranda-Gallardo M, Enriquez de Luna-Rodriguez M, Canca-Sanchez JC, Moya-Suarez AB, Morales-Asencio JM. Validation of the STRATIFY falls risk-assessment tool for acute-care hospital patients and nursing home residents: study protocol. *J Adv Nurs.* 2015;71:1948-57. Available from: https://doi.org/10.1111/jan.12651
5. McKechnie D, Pryor J, Fisher MJ. Predicting falls: considerations for screening tool selection vs. screening tool development. *J Adv Nurs.* 2016;72:2238-50. Available from: https://doi.org/10.1111/jan.12977
6. Davies C, Fisher MJ. Understanding research paradigms. *JARNA.* 2018;21:21-25. Available from: https://search.informit.org/doi/10.3316/informit.160174725752074
7. Fisher MJ, Bloomfield J. Understanding the research process. *JARNA.* 2019;22:22-27. Available from: https://doi.org/10.33235/jarna.22.1.22-27
8. Bloomfield J, Fisher MJ. Quantitative research design. *JARNA.* 2019;22:17-30. Available from: https://doi.org/10.33235/jarna.22.2.27-30
9. McKechnie D, Fisher MJ. Considerations when choosing a statistical method for data analysis. *JARNA.* 2019;22:20-9. Available from: https://doi.org/10.33235/jarna.22.3.20-29
10. Fisher MJ, McKechnie D, Pryor J. Conducting a critical review of the research literature. *JARNA.* 2020;23:20-9. Available from: https://doi.org/10.33235/jarna.23.1.20-29
11. McKechnie D, Fisher JM. Quality appraisal of the research literature in healthcare: a discussion on quality appraisal tools. *JARNA.* 2020;23:25-34. Available from: https://doi.org/10.33235/jarna.23.2.25-34
12. Mokkink LB, Terwee CB, Knol DL, Stratford PW, Alonso J, Patrick DL, et al. Protocol of the COSMIN study: COnsensus-based Standards for the selection of health Measurement INstruments. *BMC Med Res Methodol.* 2006;6:1-7. Available from: https://doi.org/10.1186/1471-2288-6-2
13. Lohr KN. Assessing health status and quality-of-life instruments: attributes and review criteria. *Qual Life Res.* 2002;11:193-205. Available from: https://doi.org/10.1023/A:1015291021312
14. Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL, et al. The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: an international Delphi study. *Qual Life Res.* 2010;19:539-549. Available from: https://doi.org/10.1007/s11136-010-9606-8
15. Kirshner B, Guyatt G. A methodological framework for assessing health indices. *J Chronic Dis.* 1985;38:27-36. Available from: https://doi.org/10.1016/0021-9681(85)90005-0
16. McDowell I. Measuring health: a guide to rating scales and questionnaires. 3rd ed. Oxford, New York: Oxford University Press, Inc., 2006.
17. Thanasegaran G. Reliability and validity issues in research. *Integration & Dissemination* 2009;4:35-40.
18. Clark LA, Watson D. Constructing validity: basic issues in objective scale development. *Psychol Assess.* 1995;7:309-19. Availaible from: https://doi.org/10.1037/14805-012
19. Portney LG. Foundations of clinical research: applications to evidence-based practice. 4th ed. Philadelphia, PA: F. A. Davis Company, 2020.
20. Drost EA. Validity and reliability in social science research. *Educ Res Perspect.* 2011;38:105-23. Available from: https://search.informit.org/doi/10.3316/informit.491551710186460

21. McKechnie D, Fisher MJ, Pryor J, Bonser M, Jesus JD. Development of the Sydney Falls Risk Screening Tool in brain injury rehabilitation: a multisite prospective cohort study. *J Clin Nurs.* 2018;27:958-68. Available from: https://doi.org/10.1111/jocn.14048
22. Field A. Discovering statistics using IBM SPSS Statistics. 4th ed. London: SAGE Publications, Ltd, 2013.
23. Gillespie BM, Chaboyer W. Assessing measurement instruments. In: Schneider Z, Whitehead D, LoBiondo-Wood G, Haber J. editors. Nursing and midwifery research: methods and appraisal for evidence based practice. Sydney: Elsevier Australia; 2016; 197-212.
24. Terwee CB, Bot SD, de Boer MR, van der Windt DAWM, Knol DL, Dekker J, et al. Quality criteria were proposed for measurement properties of health status questionnaires. *J Clin Epidemiol.* 2007;60:34-42. Available from: https://doi.org/10.1016/j.jclinepi.2006.03.012
25. Charter RA. Sample size requirements for precise estimates of reliability, generalizability, and validity coefficients. *J Clin Exp Neuropsychol.* 1999;21: 559-66. Available from: https://doi.org/10.1076/jcen.21.4.559.889
26. Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL, et al. The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. *J Clin Epidemiol.* 2010;63:737-45. Available from: https://doi.org/10.1016/j.jclinepi.2010.02.006
27. Salkind NJ. 100 questions (and answers) about research methods. Thousand Oaks, California: SAGE Publications, Inc; 2012.
28. Sherman EM, Brooks BL, Iverson GL, Slick DJ, Strauss E. Reliability and validity in neuropsychology. In: Schoenberg MR, Scott JG, editors. The little black book of neuropsychology: a syndrome-based approach. New York: Springer, 2011; 873-92.
29. Haddow JE, Palomaki GE. An introduction to assessing genomic screening and diagnostic tests. *Nutr Today.* 2011;46:162-8. Available from: https://doi.org/10.1097/NT.0b013e3182261d7f
30. DeVon HA, Block ME, Moyle-Wright P, Ernst DM, Hayden SJ, Lazzara DJ, et al. A psychometric toolbox for testing validity and reliability. *J Nurs Scholarsh.* 2007;39:155-64. Available from: https://doi.org/10.1111/j.1547-5069.2007.00161.x
31. Heale R, Twycross A. Validity and reliability in quantitative studies. *Evid Based Nurs* 2015;18:66-7. Available from: https://doi.org/10.1136/eb-2015-102129
32. Jebb AT, Ng V, Tay L. A review of key Likert Scale development advances: 1995-2019. *Front Psychol.* 2021;12:1-14. Available from: https://doi.org/10.3389/fpsyg.2021.637547
33. Barbaranelli C, Lee CS, Vellone E, Riegel B. Dimensionality and reliability of the self-care of heart failure index scales: further evidence from confirmatory factor analysis. *Res Nurs Health.* 2014;37:524-37. Available from: https://doi.org/10.1002/nur.21623
34. Dunn TJ, Baguley T, Brunsden V. From alpha to omega: a practical solution to the pervasive problem of internal consistency estimation. *Br J Psychol* 2014;105:399-412. Available from: https://doi.org/10.1111/bjop.12046
35. Zinbarg RE, Revelle W, Yovel I, Li W. Cronbach's α, Revelle's β, and Mcdonald's ωH: their relations with each other and two alternative conceptualizations of reliability. *Psychometrika.* 2005;70:123-33. Available from: https://doi.org/10.1007/s11336-003-0974-7
36. Salter K, Jutai JW, Teasell R, Foley NC, Bitensky J. Issues for selection of outcome measures in stroke rehabilitation: ICF Body Functions. *Disabil Rehabil.* 2005;27:191-207. Available from: https://doi.org/10.1080/09638280400008537
37. Chen X, Luo L, Jiang L, Shi L, Yang L, Zeng Y, et al. Development of the nurse's communication ability with angry patients scale and evaluation of its psychometric properties. *J Adv Nurs.* 2021;00:1-9. Available from: https://doi.org/10.1111/jan.14788
38. Potkin SG, Bugarski-Kirola D, Edgar CJ, Soliman S, Le Scouiller S, Kunovac J, et al. Psychometric evaluation of the Work Readiness Questionnaire in schizophrenia. *CNS Spectr.* 2016;21:199-206. Available from: https://doi.org/10.1017/S1092852914000352
39. Bland J, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet.* 1986;1:307-10. Available from: https://doi.org/10.1016/S0140-6736(86)90837-8
40. Bonett DG, Price RM. Inferential methods for the tetrachoric correlation coefficient. *J Educ Behav Stat.* 2016;30:213-25. Available from: https://doi.org/10.3102/10769986030002213
41. Brown JA, Cooper AL, Albrecht MA. Development and content validation of the Burden of Documentation for Nurses and Midwives (BurDoNsaM) survey. *J Adv Nurs.* 2020;76:1273-81. Available from: https://doi.org/10.1111/jan.14320
42. Huang CY, Chen SS, Chen CT, Lee PS, Yu TY, Chen KL. Psychometric properties and efficiency of the Computerized Adaptive Testing System for measuring Self-Care Performance in Taiwanese children with developmental disabilities. *Arch Phys Med Rehabil.* 2020;101:1332-37. Available from: https://doi.org/10.1016/j.apmr.2020.01.014
43. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *J Chiropr Med.* 2016;15:155-63. Available from: https://doi.org/10.1016/j.jcm.2016.02.012
44. Bland J. An introduction to medical statistics. 3rd ed. Oxford, United Kingdom: Oxford University Press; 2000.
45. Prinsen CA, Vohra S, Rose MR, Boers M, Tugwell P, Clarke M, et al. How to select outcome measurement instruments for outcomes included in a "Core Outcome Set" – a practical guideline. *Trials.* 2016;17:1-10. Available from: https://doi.org/10.1186/s13063-016-1555-2
46. Cronbach LJ. Coefficient alpha and the internal structure of tests. *Psychometrika.* 1951;16:297-334. Available from: https://doi.org/10.1007/BF02310555
47. De Vaus DA. Surveys in Social Research. 5th ed. NSW, Australia: Allen & Unwin; 2002.
48. Fitzpatrick R, Davey C, Buxton MJ, Jones D. Evaluation of patient-based outcome measures for use in clinical trials. *Health Technol Assess.* 1998;2:1-74. Available from: https://doi.org/10.3310/hta2140
49. Bland J, Altman DG. Statistics notes: Cronbach's alpha. *BMJ.* 1997;314:572. Available from: https://doi.org/10.1136/BMJ.314.7080.572
50. Fisher MJ, Pryor J, Capell J, et al. The psychometric properties of a modified client-centred rehabilitation questionnaire in an Australian population. *Disabil Rehabil* 2020; 42: 122-129. 2018/09/29. DOI: https://doi.org/10.1080/09638288.2018.1494214.
51. Landis RT, Koch GG. The measurement of observer agreement for categorical data. *Biometrics.* 1977;33:159-74. Available from: https://doi.org/10.2307/2529310

52. Fleiss JL, Levin B, Paik MC. Statistical methods for rates and proportions. 3rd ed. Hoboken, New Jersey: John Wiley & Sons, Inc; 2003.
53. Cicchetti DV. Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychol Assess.* 1994;6:284-90. Available from: https://doi.org/10.1037/1040-3590.6.4.284
54. Nunnally JC. Psychometric theory-25 years ago and now. *Educ Res.* 1975;4:7-21. Available from: https://doi.org/10.3102/0013189X004010007.
55. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. *Psychol Bull.* 1979;86:420-8. Available from: https://doi.org/10.1037/0033-2909.86.2.420
56. Bland J, Altman DG. A note on the use of the intraclass correlation coefficient in the evaluation of agreement between two methods of measurement. *Comput Riol Med.* 1990;20:337-40. Available from: https://doi.org/10.1016/0010-4825(90)90013-F
57. McKechnie D, Pryor J, Fisher MJ. An examination of patient characteristics that contribute to falls in the inpatient traumatic brain injury rehabilitation setting. *Disabil Rehabil.* 2017;39:1864-71. Available from: https://doi.org/10.1080/09638288.2016.1212112
58. McKechnie D, Pryor J, Fisher MJ. Falls and fallers in traumatic brain injury (TBI) rehabilitation settings: an integrative review. *Disabil Rehabil.* 2015;37:2291-9. Available from: https://doi.org/10.3109/09638288.2014.1002578
59. McKechnie D, Fisher MJ, Pryor J. A case-control study examining the characteristics of patients who fall in an inpatient traumatic brain injury rehabilitation setting. *J Head Trauma Rehabil.* 2016;31:E59-70. Available from: https://doi.org/10.1097/HTR.0000000000000146
60. Toth-Kiraly I, Bothe B, Rigo A, Orosz G. An illustration of the Exploratory Structural Equation Modeling (ESEM) framework on the Passion Scale. *Front Psychol.* 2017;8:1-15. Available from: https://doi.org/10.3389/fpsyg.2017.01968
61. Lynn MR. Determination and quantification of content validity. *Nurs Res.* 1986;35:382-5. Available from: https://doi.org/10.1097/00006199-198611000-00017
62. Polit DF, Beck CT. The content validity index: are you sure you know what's being reported? Critique and recommendations. *Res Nurs Health.* 2006;29: 489-497. Available from: https://doi.org/10.1002/nur.20147
63. Waltz CF, Strickland OL, Lenz ER. Measurement in nursing and health research. 3rd ed. New York, USA: Springer Publishing Company, Inc; 2005.
64. Polit DF, Beck CT, Owen SV. Is the CVI an acceptable indicator of content validity? Appraisal and recommendations. *Res Nurs Health.* 2007;30:459-67. Available from: https://doi.org/10.1002/nur.20199
65. Davis LL. Instrument review: getting the most from a panel of experts. *Appl Nurse Res.* 1992;5:194-7. Available from: https://doi.org/10.1016/S0897-1897(05)80008-4
66. Whittaker TA, Worthington RL. Item Response Theory in scale development research: a critical analysis. *Couns Psychol.* 2016;44:216-25. Available from: https://doi.org/10.1177/0011000015626273
67. Prinsen CA, Mokkink LB, Bouter LM, Alonso J, Patrick DL, de Vet HCW, et al. COSMIN guideline for systematic reviews of patient-reported outcome measures. *Qual Life Res.* 2018;27:1147-57. Available from: https://doi.org/10.1007/s11136-018-1798-3
68. Comrey AL, Lee HB. A first course in factor analysis. 2nd ed. New York: Psychology Press; 1992.
69. DiStefano C, Hess B. Using confirmatory factor analysis for construct validation: an empirical review. *J Psychoeduc Assess.* 2005;23:225-41. Available from: https://doi.org/10.1177/073428290502300303
70. Hu L, Bentler PM. Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. *Struct Equ Model.* 1999;6:1-55. Available from: https://doi.org/10.1080/10705519909540118
71. Hosmer DW, Lemeshow S, Sturdivant RX. Applied Logistic Regression. 3rd ed. Hoboken, New Jersey: John Wiley & Sons; 2013.
72. Husted JA, Cook RJ, Farewell VT, Gladman DD. Methods for assessing responsiveness: a critical review and recommendations. *J Clin Epidemiol.* 2000;53:459-68. Available from: https://doi.org/10.1016/s0895-4356(99)00206-1
73. McHorney CA, Ware JE, Lu JF, Sherbourne CD. The MOS 36-Item Short-Form Health Survey (SF-36): III. Tests of data quality, scaling assumptions, and reliability across diverse patient groups. *Med Care.* 1994;32:40-66. Available from: http://www.jstor.org/stable/3766189
74. Cohen J. Statistical Power Analysis for the Behavioral Sciences. 2nd ed. New York, USA: Lawrence Erlbaum Associates; 1988.
75. Cohen J. A power primer. *Psychol Bull.* 1992;112:155-9. Available from: https://doi.org/10.1037/0033-2909.112.1.155
76. McHorney CA, Tarlov AR. Individual-patient monitoring in clinical practice: are available health status surveys adequate? *Qual Life Res.* 1995;4:293-307. Available from: https://doi.org/10.1007/BF01593882
77. Hobart JC, Lamping DL, Freeman JA, Langdon DW, McLellan DW, Greenwood RJ, et al. Evidence-based measurement: Which disability scale for neurologic rehabilitation? *Neurology.* 2011;57:639-44. Available from: https://doi.org/10.1212/WNL.57.4.639
78. Holmes WC, Shea JA. Performance of a new, HIV/AIDS-targeted quality of life (HAT-QoL) instrument in asymptomatic seropositive individuals. *Qual Life Res.* 1997;6:561-71. Available from: https://doi.org/10.1023/A:1018464200708
79. Bossuyt PM, Reitsma JB, Linnet K, Moons KGM. Beyond diagnostic accuracy: the clinical utility of diagnostic tests. *Clin Chem.* 2012;58:1636-43. Available from: https://doi.org/10.1373/clinchem.2012.182576
80. Lesko LJ, Zineh I, Huang SM. What is clinical utility and why should we care? *Clin Pharmacol Ther.* 2010;88:729-33. Available from: https://doi.org/10.1038/clpt.2010.229
81. Grosse SD, Khoury MJ. What is the clinical utility of genetic testing? *Genet Med.* 2006;8:448-50. Available from: https://doi.org/10.1097/01.gim.0000227935.26763.c6
82. Asberg A, Mikkelsen G, Odsaeter IH. A new index of clinical utility for diagnostic tests. *Scand J Clin Lab Invest.* 2019;79:560-5. Available from: https://doi.org/10.1080/00365513.2019.1677938