Assessing measurement in health: Beyond reliability and validity

Assessing Measurement in Health: Beyond Reliability and Validity

Abstract

BACKGROUND: Psychometric concepts have undergone a transformation in health fields, as articulated in a consensus report by an international panel of health measurement experts: COSMIN, the COnsensus-based Standards for the selection of health Measurement INstruments.

OBJECTIVES: The aims of this paper are to describe emerging ideas relating to the development and testing of new measures in health fields, to present a revised measurement taxonomy that builds upon COSMIN, and to explore the extent to which the new measurement concepts have played a role in psychometric assessments in nursing.

DESIGN: A descriptive analysis of a sample of psychometric papers published in three major nursing journals was undertaken.

METHODS: A new measurement taxonomy is presented and explained. A sample of 105 studies, representing a consecutive sample of studies published in the *International Journal of Nursing Studies, Nursing Research,* and *Research in Nursing & Health* between 2010 and 2014 was reviewed to ascertain the extent to which psychometric assessments in nursing map onto the new taxonomy.

RESULTS: Most nursing studies reviewed adhered to traditional concepts of psychometric assessment, which focus on reliability and validity. The studies in the

sample rarely involved assessments of longitudinal measurement aspects, namely the reliability and validity of change scores (responsiveness).

CONCLUSIONS: Many constructs of interest to nurse researchers are amenable to change—and these constructs are frequently the target of nursing interventions designed to foster change. Future psychometric work by nurse researchers would benefit from assessments of the psychometric adequacy of change scores.

KEYWORDS:

Change scores

COSMIN

Instrument development

Measurement

Measurement error

Psychometrics

Reliability

Responsiveness

Scale development

Validity

New Measurement Concepts in Health: A Proposed Measurement Taxonomy

Measurement concepts have evolved considerably in the past few decades, and new guidance from an expert panel on measurement in health has emerged. The purpose of this paper is to highlight major evolving measurement concepts of relevance to nurse researchers, to present a new measurement taxonomy, and to explore the extent to which psychometric work in nursing journals maps onto current guidelines for rigorous assessment of scales.

The Evolution of Measurement Properties

In many fields in which measures of human attributes are developed and tested, classic ideas established by psychometricians decades ago have prevailed. The two measurement properties that have been the focus of standard psychometric assessment are *reliability* and *validity*. Nurse researchers, who have developed and evaluated hundreds of new scales, have largely followed the guidance of prominent psychometricians. Classic books such as the one by Nunnally and Bernstein (1994) are often cited by nurse researchers, as are books on scale development written by psychometricians such as DeVellis (2012) and Streiner and Norman (2008).

In medicine, however, measurement concepts have evolved beyond what is traditional. Indeed, several new ideas reflect a revolt against some aspects of classic psychometrics. The "revolution" has, however, experienced some turbulence, with a great proliferation of terms, definitions, and operationalizations of emerging measurement properties. In an effort to bring order to the turmoil that characterized

measurement contributions in medicine, a working group in the Netherlands undertook a Delphi study with the goal of arriving at a consensus among an international panel of measurement experts. Their purpose was to identify and define critical measurement properties for health researchers, to array those properties in a taxonomy, and to create checklists for evaluating measurement papers.

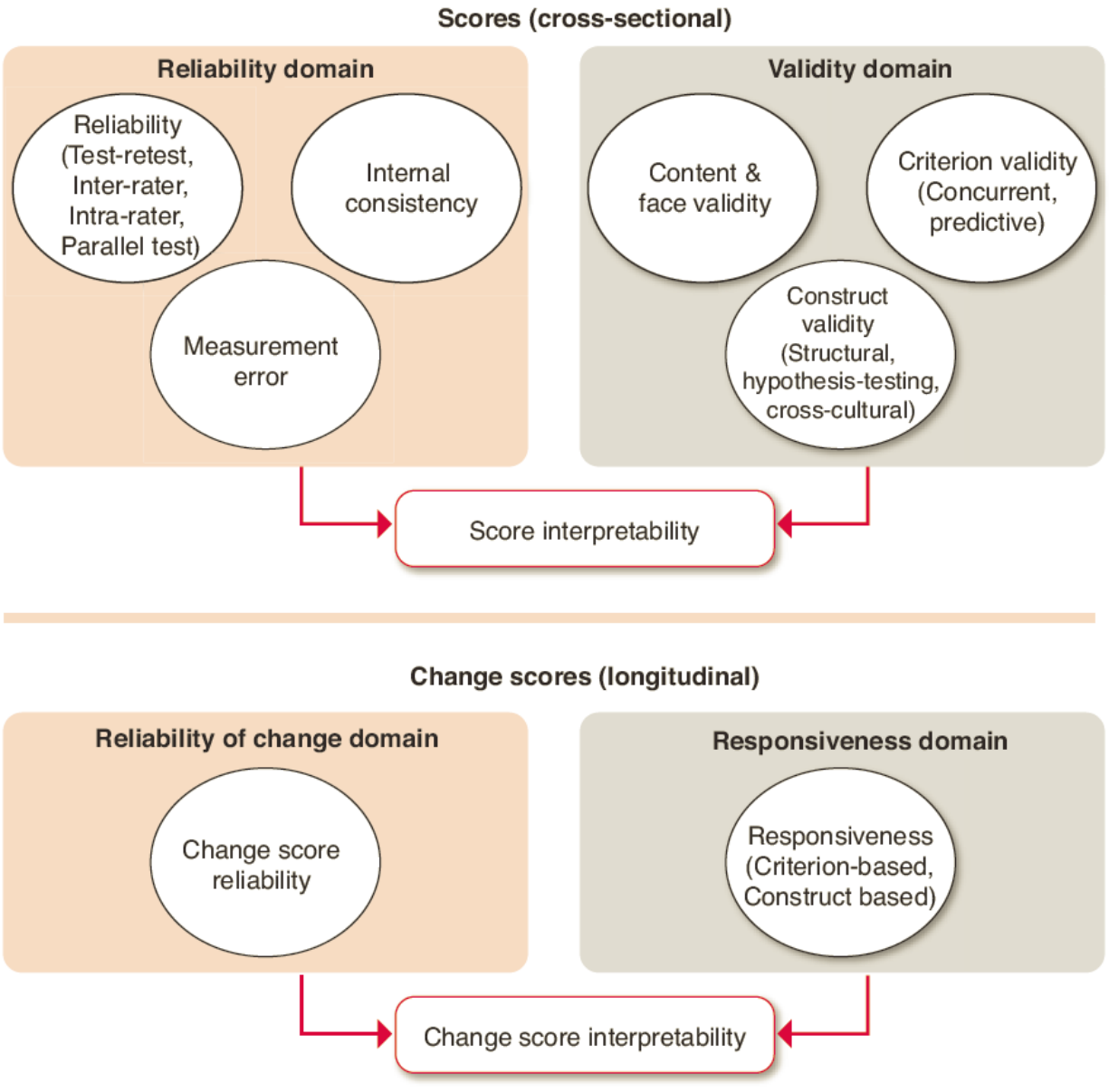
The Delphi study resulted in the creation of COSMIN, the **Co**nsensus-based **S**tandards for the selection of health **M**easurement **In**struments (Mokkink, Terwee, Patrick, Alonso, Stratford, Knol, Bouter, & DeVet, 2010a, 2010b). The new measurement ideas articulated in COSMIN await more widespread adoption in health fields other than medicine and clinical epidemiology.

The COSMIN taxonomy, available on the COSMIN website (www.cosmin.nl) includes three key measurement properties: reliability, validity, and responsiveness. More recently, Polit and Yang (2016) have proposed a slight modification of the COSMIN taxonomy that more specifically incorporates a time element. As shown in Figure 1, the standard properties of reliability and validity are depicted as being relevant for assessments of cross-sectional (point-in-time) measurements. For longitudinal measurement—that is, for measuring *change* in a construct—the two relevant measurement properties in this taxonomy are the reliability of change scores and responsiveness. In both the COSMIN and the Polit-Yang taxonomies, *interpretation* of scores and change scores is an important aspect of measurement rigor.

The current paper reviews some of the new measurement ideas reflected in these two taxonomies. To explore the extent to which new measurement concepts have

penetrated into the nursing literature, a content analysis of a small sample of psychometric papers published in nursing journals was undertaken.

Figure 1 The Polit-Yang taxonomy of measurement properties[a]



[a]Figure 3.1 in Polit & Yang (2016). Reprinted with permission

The Sample of Psychometric Nursing Papers. A sample of 105 papers that reported the psychometric testing of a scale was analyzed. A consecutive sample of papers published between 2010 and 2014 was drawn from three general nursing research journals that frequently publish papers on scale development and testing: *International Journal of Nursing Studies* (*N* = 58), *Nursing Research* (*N* = 19), and *Research in Nursing & Health* (*N* = 28). To be eligible for this analysis, the study had to involve one of the following: the development and psychometric assessment of a new instrument; the translation of an instrument into another language and an evaluation of the translated version; a psychometric evaluation of an existing instrument for a new population; or an evaluation of an instrument adaptation (e.g., the testing of a short form). Papers were excluded if they focused on a narrow psychometric question (e.g., a content validity effort or a factor analysis only), if the focus was to compare multiple scales, or if the paper was a systematic review. Papers were coded for the types of psychometric assessments that were undertaken. Intercoder reliability was assessed for a subsample of papers (*N* = 25), and was found to be high (kappa = .91).

The Reliability Domain

In both the COSMIN and Polit-Yang taxonomies, the reliability domain encompasses three components: reliability, internal consistency, and measurement error. Reliability can be defined as the degree to which "…scores for people *who have not changed* are the same for repeated measurements, under several situations" (Polit & Yang, 2016, p. 25), including repetition on different occasions (test-retest reliability and intra-rater reliability), by different persons (inter-rater reliability), or in the form of different replicates (items) on a multi-item instrument (internal consistency).

Internal Consistency. Internal consistency concerns the degree to which the items on a scale are measuring the same underlying construct. Nurse researchers have tended to follow the psychometric tradition of emphasizing internal consistency as the most important aspect of reliability, and typically rely on Cronbach's alpha as the measurement parameter to be estimated. In the sample of 105 nursing studies reviewed for this paper, coefficient alpha was computed in all but three studies. In one study, the omission was appropriate, because the measure was a formative index and not a reflective scale (Streiner, 2003). In the other two papers, the rationale for failing to evaluate internal consistency was either not stated or could be considered misguided (it was said that the scale had too few items to compute alpha). Interestingly, coefficient alpha was computed in the three studies that used Item Response Theory (IRT) or Rasch models rather than Classical Test Theory (e.g., Stump, Husman, & Brem, 2012)—even though different reliability parameters are considered more relevant with these "modern" methods of scale construction. The researchers likely understood that most readers would expect information about internal consistency in a psychometric report.

Reliability. For scales that involve self-reports, test-retest reliability assesses the extent to which scores are stable and reproducible. In test-retest reliability, replication takes the form of administering a measure to the same people twice. The assumption is that for traits that have not changed, differences in people's scores on the two testings reflect measurement error. Psychometricians have favored internal consistency over test-retest reliability and, indeed, Nunnally and Bernstein (1994) gave this explicit advice: "We recommend that the retest method generally not be used to estimate reliability" (p. 255). Their argument, which was cited by one of the papers in the sample

of nursing articles (Fogg, Mawn, & Porell, 2011), was based on theoretical and methodological considerations. These include concerns about the stability of the attribute being measured, as well as apprehensions about possible carryover effects in a second testing. Nunnally and Bernstein's advice has been widely embraced, perhaps in part because the assessment of internal consistency is easy and convenient. Calculating coefficient alpha requires only one administration of an instrument and can be readily computed in standard statistical software packages.

Medical researchers, however, have stressed the importance of retest reliability. Psychologists often measure attributes with high temporal stability (e.g., intelligence, personality) and have worried that retest assessments could actually introduce bias. For example, in a retest there is a risk that responses in a second administration of an instrument will be affected by respondents' memory of previous responses, or by their desire to appear consistent (Polit, 2014). Health care researchers, by contrast, are typically interested in attributes that are modifiable—indeed, health professionals specifically hope to improve (change) health states and behaviors through intervention, and seek to measure those changes. Consequently, medical researchers have tended to focus on a measure's ability to differentiate between random temporal fluctuations (unreliability) and true change on an attribute.

Another issue is that the appropriate measurement parameter for estimating retest reliability with continuous measures (scale scores) is the intraclass correlation coefficient (ICC), and not Pearson's correlation coefficient (DeVet, Terwee, Mokkink, & Knol, 2011). There are two basic classes of intraclass correlation coefficients—for Agreement and for Consistency. In reporting test-retest reliability, researchers should

report which intraclass correlation coefficient was used, because the one for Consistency is more liberal than the one for Agreement.

In the sample of 105 psychometric papers in nursing journals reviewed for this article, test-retest reliability was assessed in 46 studies (43.8%). Pearson's *r* was reported as the retest reliability coefficient in 18 of these 46 studies (39.1%), and 2 reports failed to state which retest statistic was used. Among those researchers who computed the intraclass correlation coefficient in the retest analysis, only two explained which intraclass correlation coefficient was used (e.g., Kalisch, Lee, & Salas, 2010). It might also be noted that the retest intervals (the time between the first and second administration) in these studies varied from 1 day to about 90 days, and only a handful of studies offered a rationale for the chosen interval (e.g., Zomorodi & Lynn, 2010). Polit (2014) and Watson (2004) have suggested strategies for selecting an appropriate retest interval.

For scales that rely on observational ratings (e.g., a measure of infant pain), the appropriate reliability parameter is inter-rater reliability—the degree to which a score can be repeated by two independent observers—or, rarely, intra-rater reliability when the same observer provides ratings on two separate occasions. For inter-rater and intra-rater reliability assessment, the appropriate parameter is either the intraclass correlation coefficient (for continuous scores) or Cohen's kappa (for nominal-level scores). In the sample of 105 nursing studies, only four involved an observational scale. In the assessments of inter-rater reliability, intraclass correlation coefficients were computed in only one of three studies with continuous scores (Chen, Lin, & Watson, 2010); the other two papers relied on correlation coefficients. In the one paper with

nominal-level scoring, kappas were appropriately computed (Roets-Merken, Zuidema, Vernooij-Dassen, & Kempen, 2014).

Measurement Error. A third component of the reliability domain in the COSMIN and Polit-Yang taxonomies is measurement error. Measurement error and reliability are related, but low measurement error does not guarantee high reliability because reliability coefficients are affected by sample variability—i.e., low variation in a set of scores dampens reliability estimates. Parameters of measurement error are useful for explaining individual scores because, unlike reliability coefficients that range from 0.0 to 1.0, measurement error statistics are in the units of the measure itself.

Two parameters of measurement error can be estimated. The first is the standard error of measurement (SEM), which is a function of both variability in scores and the measure's reliability. The standard error of measurement can be computed using either coefficient alpha or the intraclass correlation coefficient as the reliability coefficients. Medical researchers prefer calculating the standard error of measurement based on test-retest reliability rather than on internal consistency, but the opposite is true among psychometricians.

A second measurement error parameter is the limits of agreement (LOA), derived from work by Bland and Altman (1986). Bland-Altman plots are widely used by medical researchers to examine aspects of both reliability and validity of measures. A Bland-Altman plot is a useful device for visually inspecting and differentiating random measurement error and systematic error (bias) in retest assessments. In a Bland-Altman plot for retest reliability, the X (horizontal) axis is used to graph the mean of the test and retest scores for each person. The score difference on the two administrations for each

person is plotted on the Y (vertical) axis. The limits of agreement designate a confidence interval around the mean overall difference in the test and retest scores, and can easily be calculated using information from a paired *t*-test analysis.

Even though the influential psychometric guidelines, the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME Joint Committee, 2014), recommend that the standard error of measurement be reported in instrument development papers, those who follow the psychometric tradition (including nurse researchers) tend not to do so. The COSMIN group also recommended that a parameter of measurement error be communicated in measurement papers. In the sample of 105 psychometric papers reported in nursing journals, measurement error statistics were reported in two of the three papers involving a Rasch or item response theory model (Stump et al., 2012; Lerdal & Kottorp, 2011), and in only one paper using a Classical Test Theory framework (Laker, Rose, Flach, Csipke, McCrone, Craig, Kelland, & Wykes, 2012). One barrier to reporting measurement error might be that standard statistical software such as SPSS does not directly compute the standard error of measurement or limits of agreement. However, these statistics are extremely easy to calculate from available information about the scale's reliability and variability.

In summary, current advice stemming from the COSMIN initiative suggests some important changes in the reliability domain for nurse researchers who develop or adapt instruments. First, retest reliability should almost always be evaluated—although care is needed in designing a retest study, especially with regard to the interval between testings. Second, the parameter computed and reported in retest studies should be the intraclass correlation coefficient. Finally, researchers should estimate and report a

parameter of measurement error. As adoption of the COSMIN guidelines and assessment checklists expands, those who make instrument selection decisions are likely to expect such information.

The Validity Domain

Validity is defined by COSMIN as "the degree to which an instrument truly measures the construct(s) it purports to measure" (Mokkink et al., 2010a, p.743). In both the COSMIN and the Polit-Yang taxonomies, there are three validity components: content/face validity, criterion validity, and construct validity. Construct validity, an especially important aspect of validity, has multiple elements: hypothesis-testing construct validity, structural validity, and cross-cultural validity, which is relevant for the validation of a cultural or linguistic adaptation of an instrument. Features of these types of measurement validity are shown in Table 1. There have been fewer debates and controversies with regard to the validity domain than with the reliability domain, but a few points are worthy of note.

Content Validity. Content validity refers to the degree to which the content of an instrument adequately reflects the construct being measured. Nurse researchers have been leaders in content validity methods, but instrument developers in other health fields have recently begun to pay increasing attention to it. This new interest may be the result of an influential document by the United States Food and Drug Administration (2009), which stated that in self-report instruments used for pharmaceutical labeling claims (for example, for a measure of quality of life) content validity is of paramount importance. Content validity represents an early effort to enhance the construct validity

of an instrument, but claims about the validity of an instrument should never be based exclusively on evidence of adequate content validity.

Table 1 Types of Measurement-Related Validity[a]

Validity Class	Specific Type of Validity	Explanation
Content and Face Validity		
	Content validity	Concerns the adequacy of content coverage and relevance for multi-item measures of a construct
	Face validity	Concerns the extent to which an instrument looks as though it is a measure of the target construct
Criterion Validity		
	Concurrent validity	Tests whether a measure is consistent with a criterion (gold standard), measured at the same time
	Predictive validity	Tests whether a measure is consistent with a criterion (gold standard), measured at a future point in time
Construct Validity		
Hypothesis-Testing	Convergent validity	In the absence of a gold standard, tests hypotheses about the correlation between scores on the focal measure and scores on a measure of a construct with which conceptual convergence is expected
	Divergent (discriminant) validity	Tests hypotheses that the focal measure is not a measure of a different construct other than the one intended
	Known groups (discriminative) validity	Tests hypotheses about the degree to which a measure can discriminate between groups known to differ with regard to the focal construct
Structural validity	--	Tests whether a measure captures the hypothesized dimensionality of a construct, using factor analysis
Cross-cultural validity	--	Concerns the extent to which a translated or adapted measure is equivalent to the original

[a]Adapted from Table 13.1, Polit & Yang (2016)

In the sample of nursing reports analyzed for this paper, content validity was relevant in the 48 studies that involved developing a new instrument. Of these 48, content validity was formally evaluated in all but four (e.g., Lundman, et al., 2011). Content validity was not relevant for 57 studies whose purpose was to test an existing instrument in a new population (e.g., Lerdal & Kottorp, 2011), adapt or revise an existing instrument (e.g., Chou, Rau, & Lin, 2011), or translate an existing instrument (e.g., Hsu, Kao, Wang, Chang, & Tsai, 2014)—although some studies that involved a translation did re-evaluate the content validity of the translated scale (e.g., Ryu, Kim, Choi, Cleland, & Fu, 2013).

Criterion Validity. The second component in the validity domain is criterion validity, two forms of which are concurrent validity and predictive validity (Table 1). The key feature of a criterion validity approach is that there must be a "gold standard" criterion against which scores on the focal measure can be assessed. For most self-report measures, a gold standard criterion is difficult to identify—indeed, researchers from the COSMIN group expressed the belief that patient-reported outcomes (PROs) "almost always lack a gold standard" (DeVet, Terwee, Mokkink, & Knol, 2011, p. 161). However, identifying a criterion is often possible for certain patient reported outcomes (Polit & Yang, 2016). For example, in validating a self-report measure of suicide ideation, a diagnosis from a clinician could be used as the gold standard in a concurrent validity assessment. A measure of a person's intention to behave in a certain way (e.g., to quit smoking) could be assessed against actual subsequent behavior (e.g., smoking) in a predictive validity assessment.

In the sample of 105 nursing studies, there were good examples of concurrent (e.g., Chou et al., 2011) and predictive validation (e.g., Ho & McGrath, 2011). However, many researchers who lacked a gold standard criterion inappropriately claimed to have done a criterion validity assessment. When researchers examine the correlation between scores on a focal measure and scores on measures of other constructs considered to be relevant, they are assessing convergent (not criterion) validity.

Construct Validity. Construct validity is the degree to which evidence about a measure's scores supports the inference that the construct has been appropriately represented. Hypothesis-testing construct validity is a particularly important form of construct validity, and can take various forms, as shown in Table 1. In the sample of 105 psychometric nursing papers, 19 studies (18.1%) involved neither hypothesis-testing construct validation nor criterion validation. In most of these 19 papers, the researchers relied instead on factor analysis as their approach to assessing construct validity.

In COSMIN, structural validity is considered one aspect of construct validity, but is viewed as supplementing (not substituting for) hypothesis-testing validity. The COSMIN group recommends that structural validations be undertake using confirmatory factor analysis (CFA), rather than exploratory factor analysis (EFA) because confirmatory factor analysis can be used to test explicit hypotheses about an instrument's structure. Exploratory factor analysis, by contrast, is largely a tool for exploring the dimensionality of a set of items, and for identifying items to revise or eliminate. In the sample of nursing articles, factor analysis was undertaken in 85 studies (81.0%), and exploratory factor analysis was used exclusively in 51 of these (60.0%). The remaining studies that involved factor analysis used either confirmatory

factor analysis alone (18 studies), or used both exploratory and confirmatory factor analyses (16 studies).

Cross-cultural validity, the third type of construct validity, concerns the extent to which evidence supports the inference that the original and a translated or culturally adapted scale are equivalent. In the sample of 105 nursing studies, a full 36 (34.3%) of them involved efforts to assess the cross-cultural validity of a translated scale.

In summary, nurse researchers could strengthen their validity claims in instrument studies by testing thoughtful, theory-based hypotheses about the extent to which the measure yields scores that "behave" as predicted in relation to other constructs—or by identifying an appropriate gold standard for a criterion validation. Factor analysis alone as a construct validity strategy does not directly answer the central validity question: Does the scale measure the construct it purports to measure? Exploratory factor analysis is an important tool for finalizing or refining a multi-dimensional instrument, but confirmatory factor analysis should be the method of choice for structural validation.

The Reliability of Change Score Domain

In the Polit-Yang taxonomy (Figure 1), the third domain is the reliability of change scores. Change scores are calculated by subtracting scores at one point in time such as at baseline (T1) from scores at another point in time (T2), such as after an intervention. The unreliability of a measure can be compounded when change scores are computed. The difference between an imperfectly reliable score at T1 and another imperfectly reliable score at T2 potentially can mask a large change or magnify a small one. The greater the degree of unreliability, the greater the risk that a change score will

be misleading. Thus, for scales that will be used to assess change, information about the reliability of change scores ideally should be assessed.

A parameter of change score reliability can be estimated using either the standard error of measurement or the limits of agreement. Psychotherapists developed an index called the Reliable Change Index (RCI), which is calculated with a formula that uses the standard error of measurement (Jacobson & Truax, 1991). By contrast, medical researchers most often compute an index called the Smallest Detectable Change (SDC) (or minimal detectable change, MDC), which is defined as any value outside the limits of agreement (De Vet et al., 2011; Polit and Yang, 2016). Both indexes are easy to compute using information about reliability and variability. The Reliable Change Index and Smallest Detectable Change tend to yield similar but not identical values, in part because of differences in how variability is captured as well as differences in which reliability parameter is used. None of the 105 papers in the sample of nursing papers reported a value for the Smallest Detectable Change or Reliable Change Index.

It should be noted that in the COSMIN taxonomy, the Smallest Detectable Change was regarded as a mechanism for *interpreting* change scores. Polit and Yang (2016) view the Smallest Detectable Change and the Reliable Change Index as longitudinal extensions of measurement error, which is a component in the reliability domain. The Smallest Detectable Change and Reliable Change Index were thus conceptualized as parameters in a reliability domain for longitudinal measurement assessments.

The Responsiveness Domain

Responsiveness is another measurement property that concerns change scores, and so is relevant for instruments that might be used to assess improvement or deterioration on a health construct. The COSMIN initiative was probably undertaken largely to arrive at a consensus with regard to responsiveness, because it is a measurement property that has stirred debates.

The noted clinical epidemiologist Alvan Feinstein (1987), father of an approach to clinical measurement called clinimetrics, rejected standard psychometric assessments as being insufficient for evaluating clinical instruments. He argued that an important aspect of a measure's quality is its *sensitivity* to change, because a core feature of clinical evaluation is to assess whether patients have improved. Gordon Guyatt and his colleagues (1986, 1987) appear to have been the first to suggest using the term *responsiveness*, to avoid confusion with sensitivity as an index of diagnostic accuracy (i.e., sensitivity versus specificity).

In the years after Guyatt introduced the term, responsiveness gained recognition as an important measurement property in medicine. Yet, there was no agreement about what responsiveness is or how to evaluate it. In their systematic review of the quality-of-life literature, Terwee, Dekker, Wiersinga, Prummel, and Bossuyt (2003) found 25 definitions of responsiveness and 31 methods for assessing it. They organized the definitions of responsiveness into three major categories: as a measure's ability to (1) detect a change, in general; (2) detect clinically important change; and (3) capture change consistent with a true change in the construct being measured. The COSMIN panel, in their effort to resolve the controversy, achieved consensus in defining responsiveness in this third manner—i.e., as the *validity* of change scores.

Responsiveness as a measurement property has been rejected by psychometricians, including ones who have been prominent in health measurement (e.g., Hays & Hadorn, 1992; Streiner & Norman, 2008). Their argument is that responsiveness is simply longitudinal construct validity and does not require a separate label. Yet, having a separately named measurement property does have the advantage of calling attention to the need to evaluate it in new measures. Few instrument developers in the psychometric tradition (including ones in nursing) have specifically sought to gather evidence about an instrument's longitudinal validity.

It is beyond the scope of this paper to describe the myriad methods that can be used to assess responsiveness, which include both a criterion-based approach and a construct-based approach. Most methods are similar to ones used in hypothesis-testing construct validity, except that the focus is on testing hypotheses about change scores on the focal measure. For example, to assess the responsiveness of a self-report measure of physical function, we might hypothesize that change scores would reflect significantly greater improvement among mobility-impaired patients who had hip replacement surgery, compared to similarly diagnosed patients who did not have the surgery.

Responsiveness appears to be evaluated infrequently by nurse researchers. In the sample of 105 articles reviewed for this paper, only 8 research teams (7.6%) conducted assessments of the responsiveness (as defined by COSMIN) of their measures. For example, Chen, Narsavage, Culp, and Weaver (2010) tested the hypothesis that patients undergoing pulmonary rehabilitation would have improved scores at follow-up on a new measure of pulmonary function, the Short-Form Pulmonary Functional Status Scale (PFSS-11), and the hypothesis was supported.

In summary, nurse researchers who develop a new instrument that might be used to evaluate patients' progress or stability over time should give strong consideration to evaluating the measure's responsiveness, as well as the reliability of change scores.

Conclusions

This brief paper was written to heighten the visibility of recent developments in health measurement. The ideas in COSMIN emerged within the medical community, and many of the members of the 45-member COSMIN panel were working in medicine or clinical epidemiology. However, the products of the COSMIN group—the definitions of measurement properties and the checklists for reviewing measurement papers—have relevance to researchers in all health disciplines.

The COSMIN venture makes clear that there is a fundamental difference between classic psychometrics and measurement ideas that have emerged in medicine, and that difference largely centers on the measurement of change. Psychometricians have for the most part been wary of measuring change and have not put energy into devising ways to evaluate reliable and valid changes in scores. Among health care researchers, by contrast, measuring change is central to understanding the trajectories of health outcomes and the efficacy of health interventions. Interest in change is why retest reliability is held in high esteem, why measurement error is considered important (especially in relation to assessments of reliable change), and why longitudinal validity (responsiveness) is valued.

The COSMIN and Polit-Yang taxonomies, and the COSMIN checklists, illuminate the many challenges facing contemporary scale developers in their pursuit of evidence regarding their instruments' quality. Those challenges may be especially daunting to

those who have followed a more traditional psychometric program of instrument development and testing, including nurse researchers.

Three challenges are noteworthy because of resource implications. First, to undertake test-retest reliability assessments, researchers need to ask members of the research sample (or a subsample) to complete the instrument twice at an interval that is appropriate both substantively and methodologically. Substantively, the researchers must have a firm understanding of the construct so that an appropriate interval between the two administrations can be selected; the interval must be one in which the construct would not be expected to change. Methodologically, the challenge is selecting an interval that does not introduce bias from carryover effects and that minimizes the risk of attrition, which could bias the reliability estimate. The new interest in change scores is likely to result in increased expectations that retest reliability assessments for self-report measures will be conducted, because it is necessary to distinguish random measurement error stemming from the instability of responses (as estimated in a test-retest study) from true change.

A second challenge is that nurse researchers will be increasingly expected to undertake confirmatory factor analyses to provide evidence of structural validity. In turn, this means that larger samples will need to be recruited—a minimum of 200-300 is sometimes suggested, but ideally confirmatory factor analyses would be performed with even larger samples. (In the 105 psychometric nursing studies, sample size ranged from 108 to 1758 among the 34 studies that involved a confirmatory factor analysis; 20.6% of the confirmatory factor analysis studies had a sample of fewer than 200 participants). The use of confirmatory factor analysis might also mean that some nurse

researchers will have to collaborate with psychometricians or statisticians who are proficient in such analyses.

The third challenge concerns assessments of responsiveness and change score reliability, both of which require a longitudinal design. This in turn implies the need for a longer assessment schedule and more resources. The interval between administrations in a responsiveness study also needs careful consideration, and the requirements are opposite to the ones relevant in a retest study. In a retest study, the interval must be one is which no change on the construct is expected, whereas in a responsiveness study the interval must be one in which improvement or deterioration on the construct would be expected for many sample members.

Limitations. This paper has several limitations that should be noted. First, it was beyond the scope of this paper to provide formulas or detailed explanations for all of the new measurement concepts and measurement parameters discussed. References cited in this paper can be consulted for greater detail (e.g., DeVet et al., 2011; Polit & Yang, 2016).

Second, the measurement domains in the COSMIN and Polit-Yang taxonomies were summarized, but this paper does not cover an important topic, the *interpretation* of scores and change scores. New approaches to interpreting change scores (notably, by means of an index called the *minimal important change* or MIC) have been developed. These new methods merit scrutiny by nurse researchers because they involve establishing benchmarks for *clinical significance* and thus provide mechanisms to enhance the interpretability of research findings. In the sample of 105 nursing studies,

none of the researchers took steps to create a benchmark for meaningful (clinically significant) change scores.

Third, this paper did not address several other aspects of a measurement "revolution." For example, the paper did not provide information about the growing popularity of latent trait theory (item response theory or Rasch models) as an alternative to Classical Test Theory for scale development.

Finally, it cannot be claimed that the sample of studies used for illustrative purposes in this paper is representative of measurement research in nursing. The 105 papers in the sample are likely to be above average in quality, however, given the relatively high impact factors of the three selected journals and their strong history of publishing psychometric research. A scrutiny of these papers suggests that the measurement concepts articulated in the new measurement taxonomies have achieved only modest penetration in nursing, but the exact percentages reported here are unlikely to be accurate.

References

AERA, APA, & NCME Joint Committee. (2014). *Standards for educational and psychological testing (5th rev.)*. Washington: American Psychological Association.

Bland, J. M. & Altman, D. G. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet, 327,* 307-310.

Chen, Y., Lin, L., & Watson, R. (2010). Evaluation of the psychometric properties and the clinical feasibility of a Chinese version of the Doloplus-2 scale among cognitively

impaired older people with communication difficulty. *International Journal of Nursing Studies, 47,* 78-88.

Chen, Y., Narsavage, G. L., Culp, S., & Weaver, T. (2010). The development and psychometric analysis of the Short-Form Pulmonary Functional Status Scale (PFSS-11). *Research in Nursing & Health, 33,* 477-485.

Chou, P., Rau, K., & Lin, C. (2011). Development and psychometric testing of a short version of the Barriers Questionnaire-Taiwan form for cancer patients. *International Journal of Nursing Studies, 48,* 1071-1079.

DeVellis, R. F. (2012). *Scale development: Theory and application*. (3rd ed.). Thousand Oaks, CA: Sage Publications.

DeVet, H. C. W., Terwee, C., Mokkink, L. B., & Knol, D. L. (2011). *Measurement in medicine: A practical guide*. Cambridge: Cambridge University Press.

Feinstein, A. R. (1987). *Clinimetrics*. New Haven, CT: Yale University Press.

Fogg, C. J., Mawn, B., & Porell, F. (2011). Development of the Fogg Intent-to-Screen for HIV (ITS HIV Questionnaire). *Research in Nursing & Health, 34,* 73-84.

Guyatt, G.H., Bombardier, C., & Tugwell, P. (1986). Measuring disease-specific quality of life in clinical trials. *Canadian Medical Association Journal, 134,* 889-895.

Guyatt, G.H., Walter, S., & Norman, G. (1987). Measuring change over time: Assessing the usefulness of evaluative instruments. *Journal of Chronic Diseases, 40,* 171-176.

Kalisch, B., Lee, H., & Salas, E. (2010). The development and testing of the Nursing Teamwork Survey. *Nursing Research, 59,* 42-50.

Hays, R. D., & Hadorn, D. (1992). Responsiveness to change: An aspect of validity, not a separate dimension. *Quality of Life Research, 1,* 73-75.

Ho, Y. G., & McGrath, J. (2011). A Chinese version of Iowa Infant Feeding Attitude Scale: Reliability and validity assessment. *International Journal of Nursing Studies, 48,* 475-478.

Hsu, L., Kao, C., Wang, M., Chang, C., & Tsai, P. (2014). Psychometric testing of a Mandarin Chinese version of the Clinically Useful Depression Outcome Scale for patient diagnosed with type 2 diabetes mellitus. *International Journal of Nursing Studies, 51,* 1595-1604.

Jacobson, N. S., & Truax, P. (1991). Clinical significance: A statistical approach to defining meaningful change in psychotherapy research. *Journal of Consulting and Clinical Psychology, 59,* 12-19.

Laker, C., Rose, D., Flach, C., Csipke, E., McCrone, P., Craig, T., Kelland, H., & Wykes,T. (2012). Views of the Therapeutic Environment (VOTE): Stakeholder

involvement in measuring staff perceptions of acute in-patient care. *International Journal of Nursing Studies, 49,* 1403-1410.

Lerdal, A., & Kottorp, A. (2011). Psychometric properties of the Fatigue Severity Scale—Rasch analysis of individual responses in a Norwegian stroke cohort, *International Journal of Nursing Studies, 48,* 1258-1265.

Lundman, B., Viglund, K., Aléx, L., Jonsén, E., Norberg, A., Fischer, R., Strandberg, G., & Nygren, B. (2011). Development and psychometric properties of the Inner Strength Scale. *International Journal of Nursing Studies, 48,* 1266-1274.

Mokkink, L. B., Terwee, C., Patrick, D., Alonso, J., Stratford, P., Knol, D. L., Bouter, L., & DeVet, H. C. W. (2010a). The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. *Journal of Clinical Epidemiology, 63,* 737-745.

Mokkink, L. B., Terwee, C., Patrick, D., Alonso, J., Stratford, P., Knol, D. L., Bouter, L., & DeVet, H. C. W. (2010b). The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status instruments: An international Delphi study. *Quality of Life Research, 19,* 539-549.

Nunnally, J., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York: McGraw-Hill.

Polit, D. F. (2014). Getting serious about test-retest reliability: A critique of retest research and some recommendations. *Quality of Life Research, 23,* 1713-1720.

Polit, D. F., & Yang, F. (2016). *Measurement and the measurement of change: A primer for health professionals*. Philadelphia: Lippincott Williams & Wilkins.

Roets-Merken, L., Zuidema, S., Vernooij-Dassen, M., & Kempen, G. (2014). Screening for hearing, visual, and dual sensory impairment in older adults using behavioural cues: A validation study. *International Journal of Nursing Studies, 51,* 1434-1440.

Ryu, E., Kim, K., Choi, S., Cleland, C., & Fu, M. (2013). The Korean version of the Syptom Experience Index: A psychometric study. *International Journal of Nursing Studies, 50,* 1098-1107.

Streiner, D. L. (2003). Being inconsistent about consistency: When coefficient alpha does and doesn't matter. *Journal of Personality Assessment, 80,* 217-222.

Streiner, D. L., & Norman, G. R. (2008). *Health measurement scales: A practical guide to their development and use* (4th ed.). Oxford: Oxford University Press.

Stump, G. S., Husman, J., & Brem, S. (2012). The Nursing Student Self-Efficacy Scale: Development using item response theory. *Nursing Research, 61,* 149-158.

Terwee, C. B., Dekker, F., Wiersinga, W., Prummel, M., & Bossuyt, P. (2003). On assessing responsiveness of health-related quality of life instruments: Guidelines for instrument evaluation. *Quality of Life Research, 12,* 349-362.

United States Food and Drug Administration. (2009). *Guidance for industry patient-reported outcome measures: Use in medical product development to support labeling claims.* Washington, DC: U.S. Department of Health and Human Services.

Watson, D. (2004). Stability versus change, dependability versus error: Issues in the assessment of personality over time. *Journal of Research in Personality, 8,* 319-350.

Zomorodi, M., & Lynn, M. (2010). Instrument development measuring critical care nurses' attitudes and behaviors with end-of-life care. *Nursing Research, 59,* 234-240.