HHS Public Access

Author manuscript

Anesth Analg. Author manuscript; available in PMC 2023 December 08.

Published in final edited form as:

Anesth Analg. 2019 January ; 128(1): 176–181. doi:10.1213/ANE.0000000000003859.

Psychometrics: Trust, but Verify

Thomas R. Vetter, MD, MPH[*], **Catherine Cubbin, PhD**[†]

[*]Department of Surgery and Perioperative Care, Dell Medical School at the University of Texas at Austin, Austin, Texas

[†]Steve Hicks School of Social Work at the University of Texas at Austin, Austin, Texas.

Abstract

There is a continued mandate for practicing evidence-based medicine and the prerequisite rigorous analysis of the comparative effectiveness of alternative treatments. There is also an increasing emphasis on delivering value-based health care. Both these high priorities and their related endeavors require correct information about the outcomes of care. Accurately measuring and confirming health care outcomes are thus likely now of even greater importance. The present basic statistical tutorial focuses on the germane topic of psychometrics. In its narrower sense, psychometrics is the science of evaluating the attributes of such psychological tests. However, in its broader sense, psychometrics is concerned with the objective measurement of the skills, knowledge, and abilities, as well as the subjective measurement of the interests, values, and attitudes of individuals—both patients and their clinicians. While psychometrics is principally the domain and content expertise of psychiatry, psychology, and social work, it is also very pertinent to patient care, education, and research in anesthesiology, perioperative medicine, critical care, and pain medicine. A key step in selecting an existing or creating a new health-related assessment tool, scale, or survey is confirming or establishing the usefulness of the existing or new measure; this process conventionally involves assessing its reliability and its validity. Assessing reliability involves demonstrating that the measurement instrument generates consistent and hence reproducible results—in other words, whether the instrument produces the same results each time it is used in the same setting, with the same type of subjects. This includes interrater reliability, intrarater reliability, test–retest reliability, and internal reliability. Assessing validity is answering whether the instrument is actually measuring what it is intended to measure. This includes content validity, criterion validity, and construct validity. In evaluating a reported set of research data and its analyses, in a similar manner, it is important to assess the overall internal validity of the attendant study design and the external validity (generalizability) of its findings.

Address correspondence to Thomas R. Vetter, MD, MPH, Department of Surgery and Perioperative Care, Dell Medical School at the University of Texas at Austin, Health Discovery Bldg, Room 6.812, 1701 Trinity St, Austin, TX 78712. thomas.vetter@austin.utexas.edu.

DISCLOSURES

Name: Thomas R. Vetter, MD, MPH.
Contribution: This author helped write and revise the manuscript.
Name: Catherine Cubbin, PhD.
Contribution: This author helped write and revise the manuscript.

The authors declare no conflicts of interest.

Reprints will not be available from the authors.

доверяй, но проверяй — "Trust, but verify"

—Russian proverb, often used by Vladimir Lenin (1870–1924), Russian communist revolutionary, politician, and political theorist; as well as Ronald Reagan (1911–2004), American actor, politician, and 40th President of the United States

There is a continued mandate for practicing evidence-based medicine (EBM) and the prerequisite rigorous analysis of the comparative effectiveness of alternative treatments.[1] This is borne out by a recently published EBM "manifesto for better healthcare"—authored in response to perceived systematic bias, waste, error, and fraud in research underpinning patient care.[2]

Contemporary EBM moreover now increasingly (1) stresses the need to combine the critical appraisal of available evidence with patient's values and preferences through shared decision-making and (2) recognizes the crucial role of patient values and preferences in clinical decision-making.[3] This is epitomized by the present-day balancing and attendant frequent tension between the quantity of life and the quality of a patient's life.[4]

There is also an increasing emphasis on delivering value-based health care, in which value can be defined as health outcomes achieved per dollar spent.[5] In this value-based health care quotient, the numerator of achieved health outcomes encompasses not only quality and safety but also patient and provider satisfaction.[6–8]

Both these high priorities and their related endeavors require correct information about the outcomes of care. Accurately measuring and confirming health care outcomes are thus likely now of even greater importance.[1,9]

Earlier tutorials in this ongoing series in *Anesthesia & Analgesia* dealt with types of clinical and research data[10] and agreement analysis.[11] The previous tutorial focused on diagnostic testing and medical decision-making.[12] The present basic statistical tutorial focuses on the related and equally germane topic of psychometrics. It is not intended to provide in-depth coverage but instead to familiarize the reader with these specific psychometric concepts and techniques[13–15]:

- Internal reliability
- Test–retest reliability
- Interrater reliability
- Content validity: including face validity
- Criterion validity: including concurrent validity and predictive validity
- Construct validity: including convergent validity and discriminant validity
- Internal validity and external validity of a study

WHAT IS PSYCHOMETRICS?

Psychological tests are designed to measure the psychological attributes or states of individuals (eg, presence of anxiety or depression).[16] Psychometrics comprises the development, appraisal, and interpretation of psychological tests and other measures used to assess variability in behavior and to link such variability to psychological conditions.[17] In its narrower sense, psychometrics is the science of evaluating the attributes of such psychological tests,[16] specifically:

- The type of information or data generated by the psychological test
- The reliability of the information or data generated by the psychological test
- The validity of the information or data generated by the psychological test

However, in its broader sense, psychometrics is concerned with the objective measurement of the skills, knowledge, and abilities, as well as the subjective measurement of the interests, values, and attitudes of individuals—both patients and their clinicians.[17] It is in this broader context that psychometrics has greater and major applicability in clinical care and health outcomes research.

WHY SHOULD YOU CARE ABOUT PSYCHOMETRICS?

While psychometrics is principally the domain and content expertise of psychiatry, psychology, and social work, it is also very pertinent to patient care, education, and research in anesthesiology, perioperative medicine, critical care, and pain medicine.

An interdisciplinary, biopsychosocial approach to (1) practicing anesthesiology,[18] (2) addressing the needs of the intensive care unit survivor,[19,20] (3) achieving perioperative patient optimization,[21,22] and (4) most certainly, managing acute and chronic pain,[23–26] all fundamentally rely on applying psychometrically reliable and valid clinical assessment scales and tools.

Health care quality improvement is predicated on psychometrically reliable and valid surveys of domains like the patient safety climate,[27] the patient care experience,[28,29] care provider teamwork,[30] as well as care provider health and well-being versus burnout.[31,32]

Last, clinical researchers often seek to develop and to implement a new health assessment tool, scale, or survey instrument—or to apply an existing one in a novel setting or population.[4] Unfortunately, researchers can fail to initially demonstrate that a newly created assessment tool, scale, or survey instrument has adequate psychometric properties —irreparably undermining the veracity of their subsequently collected data and reported findings. Researchers also need to be careful in extrapolating reliability and validity information that had been tested in one setting or population to a novel one.

RELIABILITY VERSUS VALIDITY

A key step in selecting an existing or creating a new health-related assessment tool, scale, or survey is confirming or establishing the usefulness of the existing or new measure; this process conventionally involves assessing its reliability and its validity.[13–15,33,34]

As noted by Streiner and Norman,[35] "The terms reliability and validity have very specific meanings. They have evolved over time, reflecting a greater understanding of the process of scale development and what it is we are trying to accomplish when we assess an instrument's reliability and establish its validity with various groups."

Reliability

"Reliability is a measure of reproducibility and is solely an empirical issue."[34] Assessing reliability involves demonstrating that the measurement instrument generates consistent and hence reproducible results—in other words, whether the instrument produces the same results each time it is used in the same setting, with the same type of subjects.[13,15,33,36,37] Reliability is mainly a function of random, unsystematic error, so as random error with the measure increases, its reliability decreases.[33] There are 4 basic types of reliability: interrater reliability, intrarater reliability, test–retest reliability, and internal reliability.[34]

Validity

"Validity lies at the heart of the measurement process."[34] Assessing validity is answering whether the instrument is actually measuring what it is intended to measure[34]—in other words, how well the tool, scale, or survey really measures the underlying construct of interest.[14,15,33,36] Validity is mainly a function of nonrandom, systematic error, or bias.[33] The 3 basic types of validity are content validity, criterion validity, and construct validity, referred to as the "three C's of validity."[14,33,35,38] Although one cannot have a valid measure without it being reliable, it is quite possible to have a reliable measure that is not valid (eg, a scale that has been calibrated at minus 5 pounds).

INTERRATER RELIABILITY

Interrater or interobserver reliability refers to the reproducibility of the individual scores or answers on the same measurement instrument or survey by different raters or observers.[15,34,37] Interrater reliability focuses on the variation in scores and error that results from different observers' perceptions of the same behavior.[13]

Such agreement between raters and observers about a dichotomous (binary) variable is commonly reported as the Cohen kappa statistic (κ). The kappa statistic represents a quantitative measure of the magnitude of agreement between observers beyond what would be expected simply by chance.[11,13,33,39–41] With >2 raters or observers, the Fleiss' kappa can be applied.[11,42]

Cohen weighted kappa is typically used to assess the level of interrater agreement, beyond expected simply by chance, between raters and observers with ordinal variables and data.[11,43]

The intraclass correlation or intraclass correlation coefficient (ICC) is a commonly applied measure of agreement for continuous data.[11] The ICC is designed to determine the agreement or consistency between ≥2 assessments or measurements that share the same metric.[44,45] The ICC can be validly applied to assess interrater reliability, when multiple raters or observers, for example, evaluate the same patients in a clinical study or practice setting.[11,46,47]

INTRARATER RELIABILITY

Intrarater reliability refers to the reproducibility of the individual scores or answers on the measurement instrument or survey by the same, single rater or observer on 2 different occasions.[15,34,37] Intrarater reliability focuses on the variation in scores and error that results from the same observer's changing standards and perceptions over time of the same behavior.[13]

The ICC can also be validly applied specifically to assess intrarater reliability, when the same, single rater or observer evaluates the same patients in a clinical study or practice setting at different times.[11]

TEST–RETEST RELIABILITY

With patient self-rated tests of psychological function, pain, or disease severity or impact, there is no external observer; however, reliability of the scale is still a concern.[13] Test–retest reliability refers to the reproducibility of same group of respondents' scores or answers on the measurement instrument or survey over some logical interval of time (typically, 2–14 days apart).[13,15,34,37] A key concern with test–retest reliability is that the amount of time between tests is not so brief that respondents recall their answers on the first test, but not so long that change in the measure is likely to occur. Testing conditions on the different occasions should also be similar.

The ICC can appropriately also be applied to assess for test–retest reliability, when the study subjects or patients repeatedly complete the same measurement instrument.[11,48] Of note, while the Pearson correlation coefficient is typically applied to assess the association between 2 distinct variables, it has been posited that it can be validly applied to assess for test–retest reliability.[11,47] However, Pearson correlation coefficient can generate a liberal, overestimate of reliability.[13] Pearson correlation coefficient is likely only appropriate in situations in which the underlying condition of the study subjects is expected to change between the test–retest measurements. Otherwise, the ICC or other measures of agreement are indicated.[11]

Low test–retest reliability values can have 3 causes[13]:

- The test is innately unreliable.
- The test is reliable, but the underlying condition has changed relatively quickly over time.

- The test is “reactive” such that completing the test on 1 occasion influences some study subjects’ or patients’ subsequent responses with readministrating it.

INTERNAL RELIABILITY

Internal reliability refers to the reproducibility of individual scores or answers across similar items or questions within the measurement instrument or survey. The basic posed question is how closely each item in a scale is related to the overall scale.[34] Internal reliability is also referred to as “internal consistency.”[37]

Cronbach alpha coefficient (α),[49,50] a derivation of the ICC, is commonly applied to assess for internal reliability or internal consistency.[33] The Kuder-Richardson Formula 20 (K-R-20) is applicable to assess for test–retest reliability with dichotomous (binary) data.[33,51] Both Cronbach α and K-R-20 assess internal reliability in terms of the internal consistency or homogeneity of the items on the scale.[33]

CONTENT VALIDITY AND FACE VALIDITY

Content validity refers to the comprehensiveness of the measure and answers the question, “Do the items contained in the measure adequately cover the domain of interest or under investigation?”[14,15,33,52] A measure that includes a wider, more representative sample of targeted behaviors, beliefs, traits, or characteristics intuitively generates inferences that are more accurate and likely true under a wider range of circumstances.[14,33] Content validity is a subjective assessment by experts in the domain.

However, in the health care setting, it is frequently impractical, or even impossible for a measure to sample the entire domain of interest or under investigation due to the inherent complexity of the domain or topic.[33] Therefore, for many health outcome measures, content validity is distilled down to so-called face validity, also a subjective assessment, in which the larger community of clinicians and/or researchers (including journal editors and peer reviewers) judges whether the measure really measures the domain or topic[15,33]—“Do the selected and included items appear on the surface to be measuring what they actually are?”[53]

“Face validity simply indicates whether, on the face of it, the instrument appears to be assessing the desired qualities.”[37] Primary considerations for face validity include its basic supporting evidence, coherence of content, and inclusion of suitable subjects to whom the measure is directed (ie, patients with the diagnosis of interest).[33]

Reporting content validity includes a description of the steps taken to create the measurement instrument and who contributed to the development of the instrument (eg, a group of local, organizational level individuals with content expertise; a panel of national or international content experts), along with any other information that supports that the instrument contains appropriate content (eg, a similarly designed and previously applied, reported instrument).[36]

CRITERION VALIDITY

The conventional definition of criterion validity is the correlation of a new health-related assessment scale with another, already shown to be valid and reliable measure of the same targeted behavior, disorder, or other clinical outcome of interest.[14] Criterion validity is operationally assessed by correlating the measure of interest with a "gold standard" measure or an already well-established and widely used measure of the same characteristic (ie, the criterion).[14,15,33] Criterion validity is in turn typically divided into 2 subcategories: concurrent validity and predictive validity.[14,33]

This definition of criterion validity naturally begs the question, "Why, if a good criterion already exists, are we going through the often laborious process of developing a new instrument?"[14] Legitimate reasons can be either (1) that the existing, "gold standard" test is expensive, invasive, dangerous, or time consuming—the usual rationale for establishing concurrent validity; or (2) that the health outcome may not become manifest or apparent until too late in its natural course for effective treatment and/or secondary prevention—the usual basis for predictive validity.[14]

CONSTRUCT VALIDITY

What Is a Construct?

Like most practicing clinicians, anesthesiologists are accustomed to dealing with physical attributes that are readily observable (eg, height, weight) or can be operationally defined by their method of observation (eg, QT interval on an electrocardiogram, systolic blood pressure on a manometer).[14]

A construct is an abstract theory, idea, belief, theme, or item that cannot be directly observed but that a clinician or researcher nevertheless seeks to measure.[54] A construct attempts to explain the relationships among various observed behaviors and a set of underlying, contributing factors, including subjective traits, attitudes, and beliefs. Most psychological tests and many health outcome measures are designed to explore and to describe aspects of a theoretical or hypothetical construct.[14]

The 2 primary reasons for developing an instrument that taps into a construct are (1) the construct is a newly proposed or hypothesized one, and no scale is currently available to measure it; or (2) the existing tool is missing some key aspect of the construct. This entails more than replacing an existing tool with another that is cheaper, less invasive, safer, or shorter—the above-stated rationale for criterion validity.[14] The researcher instead uses the underlying theoretical model to devise a better instrument that can "explain a broader range of findings, explain them in a more parsimonious manner, or make more accurate predictions" about an individual's behaviors or beliefs.[14]

Establishing Construct Validity

Construct validity focuses on the relationship between a measurement instrument and ≥1 postulated but unobservable constructs. Because these constructs cannot be directly observed and lack an established criterion for validation, establishing construct validity

involves hypothesis testing within the context of the underlying theoretical or conceptual model.[14,15,33,55,56]

This process includes specifying and elaborating this underlying model, choosing a research design and methods, and collecting empirical observation.[56]

In practice, a researcher can first examine group differences on the measurement instrument, with one group known to have the characteristic and another group known not to have the characteristic. The observed differences in the scores on the instrument scale are statistically compared.[33] Doing so supports the presence of construct validity by demonstrating so-called known groups validity.[57]

Second, measures of similar constructs should be related and thus highly and significantly correlated—referred to as "convergent validity." Measures of dissimilar constructs should be unrelated and thus not highly and not significantly correlated—referred to as "discriminant validity."[14,33] For example, to have construct validity, a novel measure of quality of life should be highly correlated with other established measures of quality of life, but not so highly correlated with other constructs.

INTERNAL VALIDITY AND EXTERNAL VALIDITY OF A STUDY

In evaluating a reported set of research data and its analyses, it is important to assess the overall internal validity of the attendant study design and the external validity its findings.

Just like assessing the validity of an instrument necessitates answering whether it is actually measuring what it is intended to measure,[34] the internal validity of a study design is the degree to which it successfully generated results that are correct for its sample of subjects and hence the corresponding population of interest.[58] Internal validity refers the extent to which a study design permits making strong cause-and-effect inferences.[59] The main strength of experimental research designs is its potential for high internal validity.

There are numerous possible threats to the internal validity of a study design (eg, recall, observational, attrition, misclassification or informational, and selection),[59] which are discussed in the earlier tutorial in this series in *Anesthesia & Analgesia* dealing with bias, confounding, and interaction.[60]

The external validity of a study is the degree to which its results are applicable to other populations, settings, and times.[59] For the reader and practicing clinician, it answers, "Assuming that the results of this study are true, do they apply to my patients as well?"[58] If a clinical study is internally valid, its findings and conclusions are then generalizable to patients who are very similar to those enrolled in the study—but not assuredly so to less similar patients or to nonclinical populations or samples. External validity is accordingly also referred to as "generalizability."[58] Where experimental research is typically strong on internal validity, it is typically weak on external validity.

CONCLUSIONS

This tutorial is not intended to promote a cookbook approach to psychometrics or to provide a simplistic, routine checklist of the types of reliability and validity. It is instead intended to raise awareness of the importance of psychometrics in patient care and clinical research in anesthesiology, perioperative medicine, critical care, and pain medicine.

The need for greater psychometric rigor is exemplified by the recent efforts of the Sedation Consortium on Endpoints and Procedures for Treatment, Education, and Research, which was established by the Analgesic, Anesthetic, and Addiction Clinical Trial Translations, Innovations, Opportunities, and Networks, a public–private partnership with the US Food and Drug Administration.[61,62]

Analgesic, Anesthetic, and Addiction Clinical Trial Translations, Innovations, Opportunities, and Networks has concluded that "the development of improved interventions for procedural sedation [in adults and children] will be facilitated by additional research on existing measures or development of novel measures using state-of-the-art methods for developing patient-reported outcomes and other clinical measures. Such efforts will require the identification of clinically important outcome domains by individuals with clinical and research expertise working collaboratively with patients and other stakeholders, followed by the development of measures that validly and reliably assess these sedation outcomes."[63,64]

We have elected to focus here on classic test theory (CTT), which dates back to the turn of the 20th century and the work of Karl Pearson, and which underlies traditional measurement scale construction and psychometrics.[13] It should be noted that starting in the late 1960s, item response theory has evolved as an alternate approach that seeks to address the posited problematic assumptions of CTT and with measurement scales constructed using CTT.[65]

For the sake of brevity, we have not included but instead refer the reader to cogent material on (1) designing questionnaires, interviews, and online surveys,[66] and (2) the advantages versus disadvantages of the different available methods of their administration—including face-to-face interviews, telephone questionnaires, mailed questionnaires, and computerized administration, using e-mail and the Web.[67]

As noted by Streiner et al,[4] "Our position, always, is not to bring a new scale into the world unless it absolute necessary." The so-inclined reader—and the aptly-motivated researcher—is referred to the definitive yet practical textbook on the development and use of health measurement scales, by Streiner et al.[4] When faced with an identified gap in research or clinical practice, these authors provide a rigorous "roadmap or guide" to the complex process of (1) deciding whether an existing instrument can be used/modified or (2) undertaking the construction and evaluation (testing) of a new scale.[4]

REFERENCES

1. Kane RL, Radosevich DM. Introduction to outcomes research. Conducting Health Outcomes Research. 1st ed. Sudbury, MA: Jones and Bartlett Learning, 2011:1–23.

2. Heneghan C, Mahtani KR, Goldacre B, Godlee F, Macdonald H, Jarvies D. Evidence based medicine manifesto for better healthcare: a response to systematic bias, wastage, error and fraud in research underpinning patient care. Evid Based Med. 2017;22:120–122. [PubMed: 28720642]

3. Djulbegovic B, Guyatt GH. Progress in evidence-based medi-cine: a quarter century on. Lancet. 2017;390:415–423. [PubMed: 28215660]

4. Streiner DL, Norman GR, Cairney J. Introduction to health measurement scales. Health Measurement Scales: A Practical Guide to Their Development and Use. 5th ed. Oxford, United Kingdom: Oxford University Press, 2015:1–6.

5. Porter ME. What is value in health care? N Engl J Med. 2010;363:2477–2481. [PubMed: 21142528]

6. Vetter TR, Ivankova NV, Pittet JF. Patient satisfaction with anesthesia: beauty is in the eye of the consumer. Anesthesiology. 2013;119:245–247. [PubMed: 23676455]

7. Mohammed K, Nolan MB, Rajjo T, et al. Creating a patient-centered health care delivery system: a systematic review of health care quality from the patient perspective. Am J Med Qual. 2016;31:12–21. [PubMed: 25082873]

8. Friedberg MW, Chen PG, Van Busum KR, et al. Factors Affecting Physician Professional Satisfaction and Their Implications for Patient Care, Health Systems, and Health Policy. Santa Monica, CA: RAND Corporation; 2013. Available at: www.rand.org/pubs/research_reports/RR439.html. Accessed August 19, 2018.

9. Roger VL. Outcomes research and epidemiology: the synergy between public health and clinical practice. Circ Cardiovasc Qual Outcomes. 2011;4:257–259. [PubMed: 21586721]

10. Vetter TR. Fundamentals of research data and variables: the devil is in the details. Anesth Analg. 2017;125:1375–1380. [PubMed: 28787341]

11. Vetter TR, Schober P. Agreement analysis: what he said, she said versus you said. Anesth Analg. 2018;126:2123–2128. [PubMed: 29677066]

12. Vetter TR, Schober P, Mascha EJ. Diagnostic testing and decision-making: beauty is not just in the eye of the beholder. Anesth Analg. 2018;127:1085–1091. [PubMed: 30096083]

13. Streiner DL, Norman GR, Cairney J. Reliability. Health Measurement Scales: A Practical Guide to Their Development and Use. 5th ed. Oxford, United Kingdom: Oxford University Press, 2015:159–199.

14. Streiner DL, Norman GR, Cairney J. Validity. Health Measurement Scales: A Practical Guide to Their Development and Use. 5th ed. Oxford, United Kingdom: Oxford University Press, 2015:227–253.

15. Rubin A, Babbie ER. Measurement. Research Methods for Social Work. 9th ed. Boston, MA: Cengage Learning, 2017: 191–217.

16. Furr RM. Psychometrics and the importance of psychological measurement. Psychometrics: An Introduction. 3rd ed. Thousand Oaks, CA: SAGE Publications, Inc, 2018:1–17.

17. Committee on Psychological Testing, Including Validity Testing for Social Security Administration Disability Determinations, Board on the Health of Select Populations, Institute of Medicine. Overview of psychological testing. Psychological Testing in the Service of Disability Determination. Washington, DC: The National Academies Press, 2015:87–116.

18. Bajwa SJ, Kalra S. A deeper understanding of anesthesiology practice: the biopsychosocial perspective. Saudi J Anaesth. 2014;8:4–5. [PubMed: 24665231]

19. Khan BA, Lasiter S, Boustani MA. CE: critical care recovery center: an innovative collaborative care model for ICU survivors. Am J Nurs. 2015;115:24–31.

20. Sevin CM, Bloom SL, Jackson JC, Wang L, Ely EW, Stollings JL. Comprehensive care of ICU survivors: development and implementation of an ICU recovery center. J Crit Care. 2018;46:141–148. [PubMed: 29929705]

21. Levett DZ, Edwards M, Grocott M, Mythen M. Preparing the patient for surgery to improve outcomes. Best Pract Res Clin Anaesthesiol. 2016;30:145–157. [PubMed: 27396803]

22. Grocott MPW, Plumb JOM, Edwards M, Fecher-Jones I, Levett DZH. Re-designing the pathway to surgery: better care and added value. Perioper Med (Lond). 2017;6:9. [PubMed: 28649376]

23. Gatchel RJ, McGeary DD, McGeary CA, Lippe B. Interdisciplinary chronic pain management: past, present, and future. Am Psychol. 2014;69:119–130. [PubMed: 24547798]

24. Meissner W, Coluzzi F, Fletcher D, et al. Improving the management of post-operative acute pain: priorities for change. Curr Med Res Opin. 2015;31:2131–2143. [PubMed: 26359332]
25. Turk DC, Monarch ES. Biopsychosocial perspective on chronic pain. In: Turk DC, Gatchel RJ, eds. Psychological Approaches to Pain Management: A Practitioner's Handbook. 3rd ed. New York, NY: Guilford Press, 2018:3–24.
26. Darnall BD, Carr DB, Schatman ME. Pain psychology and the biopsychosocial model of pain treatment: ethical imperatives and social responsibility. Pain Med. 2017;18:1413–1415. [PubMed: 27425187]
27. Colla JB, Bracken AC, Kinney LM, Weeks WB. Measuring patient safety climate: a review of surveys. Qual Saf Health Care. 2005;14:364–366. [PubMed: 16195571]
28. Beattie M, Murphy DJ, Atherton I, Lauder W. Instruments to measure patient experience of healthcare quality in hospitals: a systematic review. Syst Rev. 2015;4:97. [PubMed: 26202326]
29. Anhang Price R, Elliott MN, Zaslavsky AM, et al. Examining the role of patient experience surveys in measuring health care quality. Med Care Res Rev. 2014;71:522–554. [PubMed: 25027409]
30. Valentine MA, Nembhard IM, Edmondson AC. Measuring teamwork in health care settings: a review of survey instruments. Med Care. 2015;53:e16–e30. [PubMed: 24189550]
31. Hall LH, Johnson J, Watt I, Tsipa A, O'Connor DB. Healthcare staff wellbeing, burnout, and patient safety: a systematic review. PLoS One. 2016;11:e0159015.
32. Brand SL, Thompson Coon J, Fleming LE, Carroll L, Bethel A, Wyatt K. Whole-system approaches to improving the health and wellbeing of healthcare workers: a systematic review. PLoS One. 2017;12:e0188418.
33. Kane RL, Radosevich DM. Measurement. Conducting Health Outcomes Research. 1st ed. Sudbury, MA: Jones and Bartlett Learning, 2011:49–84.
34. Frytak JR, Kane RL. Measurement. In: Kane RL, ed. Understanding Health Care Outcomes Research. 2nd ed. Sudbury, MA: Jones and Bartlett Publishers, 2006:83–120.
35. Streiner DL, Norman GR. "Precision" and "accuracy": two terms that are neither. J Clin Epidemiol. 2006;59:327–330. [PubMed: 16549250]
36. Sullivan GM. A primer on the validity of assessment instruments. J Grad Med Educ. 2011;3:119–120. [PubMed: 22655129]
37. Streiner DL, Norman GR, Cairney J. Basic concepts. Health Measurement Scales: A Practical Guide to Their Development and Use. 5th ed. Oxford, United Kingdom: Oxford University Press, 2015:7–18.
38. Cronbach LJ, Meehl PE. Construct validity in psychological tests. Psychol Bull. 1955;52:281–302. [PubMed: 13245896]
39. Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Meas. 1960;20:37–46.
40. Viera AJ, Garrett JM. Understanding interobserver agreement: the kappa statistic. Fam Med. 2005;37:360–363. [PubMed: 15883903]
41. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977;33:159–174. [PubMed: 843571]
42. Fleiss JL. Measuring nominal scale agreement among many raters. Psychol Bull. 1971;76:378–382.
43. Cohen J. Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. Psychol Bull. 1968;70:213–220. [PubMed: 19673146]
44. Müller R, Büttner P. A critical discussion of intraclass correlation coefficients. Stat Med. 1994;13:2465–2476. [PubMed: 7701147]
45. McGraw KO, Wong SP. Forming inferences about some intraclass correlation coefficients. Psychol Methods. 1966;1:30–46.
46. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. Psychol Bull. 1979;86:420–428. [PubMed: 18839484]
47. Rousson V, Gasser T, Seifert B. Assessing intrarater, interrater and test-retest reliability of continuous measurements. Stat Med. 2002;21:3431–3446. [PubMed: 12407682]

48. Yen M, Lo LH. Examining test-retest reliability: an intra-class correlation approach. Nurs Res. 2002;51:59–62. [PubMed: 11822570]
49. Cronbach LJ. Coefficient alpha and the internal structure of tests. Psychometrika. 1951;16:297–334.
50. Cronbach LJ, Shavelson RJ. My current thoughts on coefficient alpha and successor procedures. Educ Psychol Meas. 2004;64:391–418.
51. Cronbach LJ. Essentials of Psychological Testing. New York, NY: Harper & Row; 1990.
52. Streiner DL, Norman GR, Cairney J. Devising the items. Health Measurement Scales: A Practical Guide to Their Development and Use. 5th ed. Oxford, United Kingdom: Oxford University Press, 2015:19–37.
53. Streiner DL, Norman GR, Cairney J. Selecting the items. Health Measurement Scales: A Practical Guide to Their Development and Use. 5th ed. Oxford, United Kingdom: Oxford University Press, 2015:74–99.
54. Construct Dew D.. In: Lavrakas P, ed. Encyclopedia of Survey Research Methods. Vol. 1. Thousand Oaks, CA: Sage Publications, Inc, 2008:134.
55. Smith GT. On construct validity: issues of method and measurement. Psychol Assess. 2005;17:396–408. [PubMed: 16393005]
56. Smith GT. On the complexity of quantifying construct validity. Psychol Assess. 2005;17:413–414. [PubMed: 16393007]
57. Davidson M. Known-groups validity. In: Michalos AC, ed. Encyclopedia of Quality of Life and Well-Being Research. Dordrecht, the Netherlands: Springer Netherlands, 2014:3481–3482.
58. Fletcher RH, Fletcher SW, Fletcher GS. Introduction. Clinical Epidemiology: The Essentials. 5th ed. Philadelphia, PA: Wolters Kluwer/Lippincott Williams & Wilkins, 2014:1–16.
59. Kane RL, Radosevich DM. Outcomes research study design. Conducting Health Outcomes Research. 1st ed. Sudbury, MA: Jones and Bartlett Learning, 2011:39–48.
60. Vetter TR, Mascha EJ. Bias, confounding, and interaction: lions and tigers, and bears, oh my! Anesth Analg. 2017;125:1042–1048. [PubMed: 28817531]
61. Williams MR, Ward DS, Carlson D, et al. Evaluating patient-centered outcomes in clinical trials of procedural sedation, part 1 efficacy: sedation consortium on endpoints and procedures for treatment, education, and research recommendations. Anesth Analg. 2017;124:821–830. [PubMed: 27622720]
62. Ward DS, Williams MR, Berkenbosch JW, et al. Evaluating patient-centered outcomes in clinical trials of procedural sedation, part 2 safety: sedation consortium on endpoints and procedures for treatment, education, and research recommendations. Anesth Analg. 2018;127:1146–1154. [PubMed: 29782404]
63. Williams MR, McKeown A, Dexter F, et al. Efficacy outcome measures for procedural sedation clinical trials in adults: an ACTTION systematic review. Anesth Analg. 2016;122:152–170. [PubMed: 26678470]
64. Williams MR, Nayshtut M, Hoefnagel A, et al. Efficacy outcome measures for pediatric procedural sedation clinical trials: an ACTTION systematic review. Anesth Analg. 2018;126:956–967. [PubMed: 28922236]
65. Streiner DL, Norman GR, Cairney J. Item response theory. Health Measurement Scales: A Practical Guide to Their Development and Use. 5th ed. Oxford, United Kingdom: Oxford University Press, 2015:273–303.
66. Hulley SB, Cummings SR, Browner WS, Grady DG, Newman TB. Designing questionnaires, interviews, and online surveys. In: Cummings SR, Kohn MA, Hulley SB, eds. Designing Clinical Research. 4th ed. Philadelphia, PA: Wolters Kluwer/Lippincott Williams & Wilkins, 2013:223–236.
67. Streiner DL, Norman GR, Cairney J. Methods of administration. Health Measurement Scales: A Practical Guide to Their Development and Use. 5th ed. Oxford, United Kingdom: Oxford University Press, 2015:304–339.