There is a continued mandate for practicing evidence-based medicine and the prerequisite rigorous analysis of the comparative effectiveness of alternative treatments. 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. 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. In this value-based health care, the numerator of achieved health outcomes encompasses not only quality and safety but also patient and provider satisfaction.

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. Earlier tutorials in this ongoing series in Anesthesia & Analgesia dealt with types of clinical and research data and agreement analysis. The present tutorial focused on diagnostic testing and medical decision-making. 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:

- 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

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. 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. In this value-based health care, the numerator of achieved health outcomes encompasses not only quality and safety but also patient and provider satisfaction.

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.

Earlier tutorials in this ongoing series in Anesthesia & Analgesia dealt with types of clinical and research data and agreement analysis. The present tutorial focused on diagnostic testing and medical decision-making. 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:

- 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

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.
WHAT IS PSYCHOMETRICS?

Psychological tests are designed to measure the psychological attributes or states of individuals (eg, presence of anxiety or depression). 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 coordination,30 the patient’s perception of respect and empathy,31 the patient’s perception of the professional conduct of caregivers,32 the patient’s perception of the treatment and management of pain,33 the patient care experience,34,35 professional culture,36 the patient’s perception of the treatment and management of pain,33,34,36,37 and the patient’s perception of the treatment and management of pain.33

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 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—in other words, how well the tool, scale, or survey really measures the underlying construct of interest.13,14,33,35,36 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.”14,34,36,37 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).

INTRARATER 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.13,15,34,37 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.13,14,15,33,38–41 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
SPECIAL ARTICLE

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.13,14,37 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.33

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.31

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).33,34,37 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 test–retest reliability, when the study subjects or patients repeatedly complete the same measurement instrument.11,14 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 causes13:

• 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.33 Internal reliability is also referred to as “internal consistency.”57

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.33,35 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?”14,15,33,32 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 topic15,33—“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.”53 Primary considerations for face validity include its basic supporting evidence, coherence of content, and inclusion of appropriate items 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

178 www.anesthesia-analgesia.org ANESTHESIA & ANALGESIA
Copyright © 2018 International Anesthesia Research Society. Unauthorized reproduction of this article is prohibited.
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.14 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.13,33,35,36

This process includes specifying and elaborating this underlying model, choosing a research design and methods, and collecting empirical observation.36

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.35 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 to 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, attribution, 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?”59 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.30,82

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.4,6,6,4

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 is absolutely necessary.” The so-inclined reader—and the aptly-motivated researcher—is referred to the definitive yet practical text-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 is absolutely 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

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.

This manuscript was handled by: Jean-Francois Pittet, MD.

REFERENCES
