OPEN

SPECIAL TOPIC

Special Topic

Validity, Reliability, and the Questionable Role of Psychometrics in Plastic Surgery

Eric Swanson, MD

Summary: This report examines the meaning of validity and reliability and the role of psychometrics in plastic surgery. Study titles increasingly include the word "valid" to support the authors' claims. Studies by other investigators may be labeled "not validated." Validity simply refers to the ability of a device to measure what it intends to measure. Validity is not an intrinsic test property. It is a relative term most credibly assigned by the independent user. Similarly, the word "reliable" is subject to interpretation. In psychometrics, its meaning is synonymous with "reproducible." The definitions of valid and reliable are analogous to accuracy and precision. Reliability (both the reliability of the data and the consistency of measurements) is a prerequisite for validity. Outcome measures in plastic surgery are intended to be surveys, not tests. The role of psychometric modeling in plastic surgery is unclear, and this discipline introduces difficult jargon that can discourage investigators. Standard statistical tests suffice. The unambiguous term "reproducible" is preferred when discussing data consistency. Study design and methodology are essential considerations when assessing a study's validity. *(Plast Reconstr Surg Glob Open 2014;2:e161; doi: 10.1097/GOX.0000000000000103; Published online 3 June 2014.)*

Increasingly, investigators use the term "validation" in labeling their studies.[1–14] Before we pass judgment on a study or test and label it "validated" or "not validated," we need to know the meaning of validated. According to the dictionary,[15] to validate means "to support or corroborate on a sound basis." Evidence-based medicine examines the soundness of a study using quality criteria such as randomization, prospective study design, and the use of controls.[16] It represents the best existing structure from which to judge the soundness of a study. The importance of methodological considerations has been emphasized.[16] This report examines the meaning of validity and reliability and the role of psychometrics in plastic surgery.

From the Swanson Center, Leawood, Kans.

Received for publication February 3, 2014; accepted April 3, 2014.

Copyright © 2014 The Authors. Published by Lippincott Williams & Wilkins on behalf of The American Society of Plastic Surgeons. PRS Global Open is a publication of the American Society of Plastic Surgeons. This is an open-access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives 3.0 License, where it is permissible to download and share the work provided it is properly cited. The work cannot be changed in any way or used commercially.

DOI: 10.1097/GOX.0000000000000103

VALIDITY

In general usage, "valid" means well grounded or justifiable, being at once relevant and meaningful.[17] Statistically, its meaning is more specific. Statistics starts with the numbers. This discipline does not concern itself necessarily with how they came to be. (Hence, the old expression "There are lies, damned lies, and statistics.")[18] A valid test is simply one that measures what it intends to measure.[19–21] In this sense, validity is similar to accuracy, a term that also has a more restricted statistical meaning. Validity is not an absolute[19]; few tests are either perfectly valid or entirely invalid. For example, Caprini scores likely have some degree of validity in identifying patients at higher risk for thromboembolism, although perhaps not enough validity to justify their use as a screening measure for ordering anticoagulation.[22]

Disclosure: *The author has no financial interest to declare in relation to the content of this article. The Article Processing Charge was paid for by the author.*

Importantly, validity is not an intrinsic property of a test.[19] Validation is not a seal of authenticity obtained at the end of a lengthy stepwise process analogous to drug testing for Food and Drug Administration approval. Because validity is not an inherent test property, it is not transferrable. For example, a quality of life scale that has proved useful in osteoporosis treatment cannot be assumed to be valid for evaluating breast reduction patients.[23] A questionnaire designed for breast reduction patients cannot be considered validated for assessment of other types of breast surgery.[24] Validity is a quality most credibly assigned by independent investigators. Einstein's theory of relativity was tested and found to be valid by independent astronomers, not by Einstein.[25] He did not title his famous publication "General Relativity: A Valid Theory."

Adding to the confusion, many subtypes of validity have been described[19]—content, construct, criterion, convergent, and face validity, to name a handful. Ironically, the more validity is defined, the less clear is its meaning. Investigators often use the terms validity and reliability interchangeably when referencing their scales.[6,9,12] Researchers may claim validity on the basis of reproducibility alone.[6,9,12] A correlation between survey questions that are expected to have similar responses (eg, in the case of the FACE-Q, patients who believe they look younger also rate their appearance higher)[14] may be offered as evidence of validity. Such a comparison, of course, is made at the discretion of the investigator. Validity is not a quantifiable term; it is not expressed in units. When an investigator attaches units, a measure of reproducibility (ie, test-retest repeatability or internal consistency) is being reported, not validity per se.

Rigorous methodology helps to ensure scientific soundness and therefore validity.[16] Essential considerations include consecutive patients, an adequate inclusion rate and reasonable eligibility criteria, and efforts to control or eliminate confounders.[16] Surprisingly, many studies that claim validity do not report consecutive patients,[1–6,8–12] inclusion rates,[1,3,4,6,8–12] or eligibility criteria.[1,3,6,9–12] Contrary to the suggestion of some researchers,[26] rigorous methodology does not recognize patient interviews, focus groups, extensive field-testing, or expert panels as quality criteria in themselves. Evidence-based medicine holds a low regard for expert opinion.[16] Expert panels invite the influence of conventional wisdom, which is notoriously undependable.

RELIABILITY AND REPRODUCIBILITY

What about the meaning of reliable? According to the dictionary,[27] reliable means dependable and trustworthy. It follows that reliable data are data that have been accumulated in a scientifically sound manner, a meaning that strongly resembles validity. In psychometrics, however, the definition of reliable is somewhat different (reflecting an unfortunate past error in nomenclature). Reliability means consistency—the ability to provide reproducible scores.[19,21] In this context, reliability is analogous to precision. A test may be precise (results consistent) but inaccurate (the mean result is not the true mean). This concept is illustrated using targets and bullet holes in Figure 1. When considering a study, it is important to know whether the investigators are using "reliable" to mean consistent or as a synonym for validity. Notably, the International Vocabulary of Metrology[20] does not define reliability as a measurement parameter. "Reproducibility" is used instead, and this is a measurable quantity.[20]

Using the metrological definitions of accuracy and precision (also called "measurement accuracy" and "measurement precision"),[20] a test might be accurate but not precise. Such a statement has long confused statistics students because it is counterintuitive. Statistically speaking, an "accurate" test may produce the correct mean result, even though some data points are substantially different from the true value. (A third term, "trueness," is similarly defined.)[20] In general usage, however, no one would call a marksman accurate (or true) if his or her bullet patterns are widely dispersed, even if the bullet holes are centered on the bull's eye (Fig. 1). How does this discussion relate to validity and reliability? Reproducibility (or reliability as psychometrically defined) is a prerequisite for validity of a test that consistently produces true measurements,[19] just as precision is a prerequisite for the truly accurate marksman who consistently hits the bull's eye. A measuring device cannot be inconsistent and still valid. Validity is the final estimation of the usefulness of a test.

VALIDITY REQUIRES RELIABLE DATA

Regardless of the reproducibility of results, a valid test cannot be based on unreliable data.[16] This fact is both intuitive and correct. We are all familiar with the simple adage, "garbage in, garbage out." No statistical maneuver can compensate for poor-quality data.[16] Although this observation may seem obvious, even sophisticated investigators can become preoccupied with psychometric tests and lose sight of methodological safeguards that ensure that the data are sound in the first place.[28] Even with trustworthy

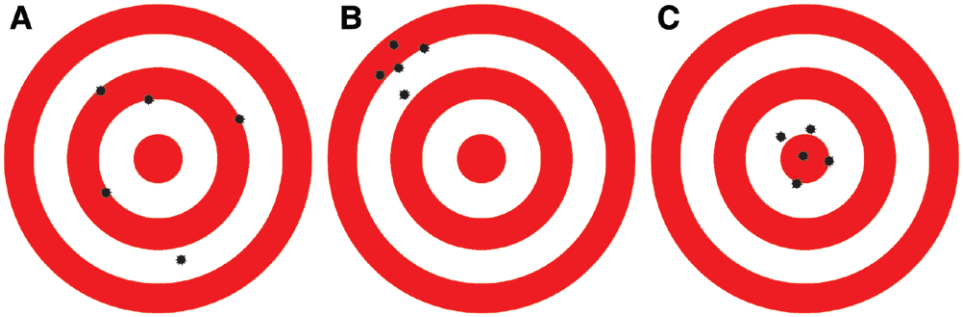

Fig. 1. Illustrations of a target and bullet holes. A marksman may shoot in the correct direction, but not consistently on the bull's eye (A). Statistically, such a shooter might be called accurate but not precise. In common usage, this marksman is neither. A marksman may shoot with a tight cluster that is not centered—precise but inaccurate (B). The skilled sharpshooter has a good aim and consistently shoots close to the bull's eye (C), precise and accurate. Reproducibility is demonstrated in targets B and C, but validity is represented only in target C.

data, validity may be compromised if a survey asks for information that differs from what it intends to measure.[28]

PSYCHOMETRICS AND PLASTIC SURGERY

Psychometrics pertains to the development of tests that seek to quantify abstract qualities such as intelligence, scholastic aptitude, and personality.[19] There is no known true value, and the final scores can have serious consequences, such as whether an applicant is accepted by a college.[19] Hence, the need for a discipline to try to make the questions as fair as possible. Plastic surgery needs are quite different. Our questionnaires are intended to be surveys, not tests.[28] We seek to evaluate patient satisfaction and improved quality of life, the most important determinants of surgical success.[4,8,16] Patients evaluate our performance in meeting these goals,[29] not the other way around. Patient surveys serve as our report cards. Many questions are practical and quantifiable, such as whether the patient would repeat the surgery (simply yes or no) or the recovery time (measured in days).[29] A cumulative score that mixes survey responses,[4,14] as provided by psychometric scales, is not clinically useful.[28]

Indeed, the relevance of psychometrics to plastic surgery is open to question. Psychometrics introduces jargon that is incomprehensible to plastic surgeons (eg, constructs, Rasch model, item fit, scaling assumptions, targeting, and floor/ceiling effects). The reader could be forgiven for thinking he or she accidentally opened a psychometrics journal. Plastic surgeons may find it impossible to critically appraise such a study and may simply give up.[30] They are likely to lose enthusiasm for any attempts to evaluate their patients using patient-reported outcome measures.

CLINICAL RELEVANCE

Correct terminology is not just an academic issue. If our words are not carefully chosen, their meaning is lost, and our ability to evaluate studies (and properly treat patients) is jeopardized. The issue is serious enough that an international body exists just to provide definitions.[20] None of this is to say that patient interviews, which can be time-intensive, are not useful in developing study questions ("content validity"). Standard statistical tests (eg, *t* tests, chi-square, and Pearson correlations), well known to plastic surgeons, are sufficient to analyze data.[29] There is no need for more sophisticated psychometric modeling.

Practically speaking, where does this discussion leave us? Both the terms valid and reliable have become interchangeable in common usage. Perhaps we should use the more specific terms "reproducible," "repeatable," or "consistent" when discussing the degree of variation in data. When discussing study validity, we should reference the scientific soundness of the data and measuring device using well-known quality criteria related to study design and methodology.[16] We do well to exercise caution when calling one measure or study "validated" and another "not validated." This concern extends to all studies, not just those using psychometrics. We should be especially circumspect when calling our own studies validated. The practice of including the word "valid" in a study title has become so common that some investigators may feel that its inclusion is obligatory and their study may be prejudiced if they do not mention it. In truth, this self-serving designation adds nothing of value to the title. Leave validity

to the judgment of the reader or the independent investigator.

CONCLUSIONS

Practicing plastic surgeons can take comfort in the fact that a presumption of validity cannot be forced upon them by the authors of a study. This determination cannot be outsourced either.[16] Even in a culture of intellectual repression ("conventional wisdom on steroids"), Galileo was able to determine for himself that heavier objects did not fall more quickly to the ground. Four hundred years later, we still need to recognize and challenge dogma. The scientific method inspired by Galileo is all we have to keep us on the true path to knowledge and understanding.

> Jargon is part ceremonial robe, part false beard.—Mason Cooley

Eric Swanson, MD
Swanson Center, 11413 Ash Street
Leawood, KS 66211
E-mail: eswanson@swansoncenter.com

ACKNOWLEDGMENTS

The author thanks Gwendolyn Godfrey for the illustration.

REFERENCES

1. Anderson RC, Cunningham B, Tafesse E, et al. Validation of the breast evaluation questionnaire for use with breast surgery patients. *Plast Reconstr Surg.* 2006;118:597–602.
2. Singer AJ, Arora B, Dagum A, et al. Development and validation of a novel scar evaluation scale. *Plast Reconstr Surg.* 2007;120:1892–1897.
3. Thomson JG, Liu YJ, Restifo RJ, et al. Surface area measurement of the female breast: phase I. Validation of a novel optical technique. *Plast Reconstr Surg.* 2009;123:1588–1596.
4. Pusic AL, Klassen AF, Scott AM, et al. Development of a new patient-reported outcome measure for breast surgery: the BREAST-Q. *Plast Reconstr Surg.* 2009;124:345–353.
5. Brown BC, McKenna SP, Solomon M, et al. The patient-reported impact of scars measure: development and validation. *Plast Reconstr Surg.* 2010;125:1439–1449.
6. Buchner L, Vamvakias G, Rom D. Validation of a photonumeric wrinkle assessment scale for assessing nasolabial fold wrinkles. *Plast Reconstr Surg.* 2010;126:596–601.
7. Creasman CN, Mordaunt D, Liolios T, et al. Four-dimensional breast imaging, part II: clinical implementation and validation of a computer imaging system for breast augmentation planning. *Aesthet Surg J.* 2011;31:925–938.
8. Cano SJ, Klassen AF, Scott AM, et al. The BREAST-Q: further validation in independent clinical samples. *Plast Reconstr Surg.* 2012;129:293–302.
9. Kane MA, Lorenc ZP, Lin X, et al. Validation of a lip fullness scale for assessment of lip augmentation. *Plast Reconstr Surg.* 2012;129:822e–828e.
10. Steffen A, Magritz R, Frenzel H, et al. Psychometric validation of the youth quality of life-facial differences questionnaire in patients following ear reconstruction with rib cartilage in microtia. *Plast Reconstr Surg.* 2012;129:184e–186e.
11. Kane MA, Blitzer A, Brandt FS, et al. Development and validation of a new clinically-meaningful rating scale for measuring lateral canthal line severity. *Aesthet Surg J.* 2012;32:275–285.
12. Lorenc ZP, Bank D, Kane M, et al. Validation of a four-point photographic scale for the assessment of midface volume loss and/or contour deficiency. *Plast Reconstr Surg.* 2012;130:1330–1336.
13. Kececi Y, Sir E, Zengel B. Validation of the Turkish version of the Breast Reduction Assessed Severity Scale. *Aesthet Surg J.* 2013;33:66–74.
14. Klassen AF, Cano SJ, Scott AM, et al. Measuring outcomes that matter to face-lift patients: development and validation of FACE-Q appearance appraisal scales and adverse effects checklist for the lower face and neck. *Plast Reconstr Surg.* 2014;133:21–30.
15. Validated. Available at: http://www.merriam-webster.com/dictionary/validate. Accessed January 13, 2014.
16. Swanson E. Levels of evidence in cosmetic surgery: analysis and recommendations using a new CLEAR classification. *Plast Reconstr Surg Glob Open* 2013;1:e66.
17. Valid. Available at: http://www.merriam-webster.com/dictionary/valid?show=0&t=1389627866. Accessed January 13, 2014.
18. "There are lies, damned lies and statistics." Available at: http://www.brainyquote.com/quotes/quotes/m/marktwain128372.html. Accessed February 1, 2014.
19. Murphy KR, Davidshofer CO. *Psychological Testing: Principles and Applications.* 6th ed. Upper Saddle River, N.J.: Pearson/Prentice Hall; 2005.
20. Joint Committee for Guides in Metrology Working Group 2. International vocabulary of metrology: basic and general concepts and associated terms. *ISO/IEC Guide 99.* 2007:1–92.
21. Pusic AL, Lemaine V, Klassen AF, et al. Patient-reported outcome measures in plastic surgery: use and interpretation in evidence-based medicine. *Plast Reconstr Surg.* 2011;127:1361–1367.
22. Swanson E. Chemoprophylaxis for venous thromboembolism prevention: concerns regarding efficacy and ethics. *Plast Reconstr Surg Glob Open* 2013;1:e23.
23. Swanson E. Randomized controlled trial comparing health-related quality of life in patients undergoing vertical scar versus inverted T-shaped reduction mammaplasty. *Plast Reconstr Surg.* 2014;133:59e–60e.
24. Swanson E. Outcomes analysis of patients undergoing autoaugmentation after breast implant removal. *Plast Reconstr Surg.* 2014;133:216e–218e.
25. Isaacson W. *Einstein: His Life and Universe.* New York: Simon & Schuster; 2007.
26. Ward JA, Potter S, Blazeby JM; BRAVO Study Steering Committee. The BREAST-Q: further validation in independent clinical samples. *Plast Reconstr Surg.* 2012;130:616e–618e.
27. Reliable. Available at: http://www.merriam-webster.com/dictionary/reliable. Accessed January 13, 2014.
28. Swanson E. The FACE-Q: The importance of full disclosure and sound methodology in outcome studies. *Aesthet Surg J.* 2014;34:626–627.
29. Swanson E. Prospective outcome study of 225 cases of breast augmentation. *Plast Reconstr Surg.* 2013;131:1158–1166; discussion 1167–1168.
30. Hammond DC. Discussion. The BREAST-Q: further validation in independent clinical samples. *Plast Reconstr Surg.* 2012;129:303–304.