A Call for a Shift in Theory and Terminology for Validation Studies in Urological Education

As the use and adoption of simulation based training and assessment increase in urology, it becomes important for the design and interpretation of education studies to be aligned with behavioral science standards and the greater health care community. In the last 60 years the American Psychological Association, American Educational Research Association and the National Council on Measurement in Education have developed and introduced 6 editions of the "standards" to guide the validation process.[1] Adoption of the 1974 standards has guided much of the urology simulation education based literature to date as a result of the sentinel and invaluable article for surgical disciplines by Gallagher et al[2] and McDougall's 2007 article applying these concepts to urology.[3] The crux of these standards includes the division of validity into what are now well-known subjective concepts such as content validity and face validity, and objective concepts such as construct validity and criterion validity with little to no association with an overall "construct" to wrap around the evidence gained.

Since 1974 there has been a major shift in the concept of validity reflected in several standards for educational and psychological testing released between 1985 and 2014 which has yet to be reflected in the urological simulation education based literature.[4] Cook et al have begun to develop methodologies to translate these concepts into applications for simulation in health care,[5] as are leaders in the surgical disciplines as well. We introduce the updated taxonomy of validity, and encourage scholars in urological education to adopt the new concepts presented in the updated standards for future educational validity studies.

UPDATED VALIDITY CONCEPTS

First and foremost, validity is no longer divided into face, content, construct and criterion, and it is never "established." Validity evidence is accumulated towards a specific purpose. The current definition of validity refers to "the degree to which evidence and theory support the interpretation of simulator data/scores for measuring a certain construct."[4] Therefore, validity is a hypothesis, and in order to support or refute it, evidence should be collected. The construct is defined as "the concepts or characteristics that a simulator is designed to measure." Therefore, all of the validity evidence should be assessed around how well the data obtained by an educational intervention measure the intended construct.[4]

Examples of distinct intended constructs to be tested include but are not limited to: 1) Do the data support selection of medical students for a career in urology? 2) Do the data support facilitating the learning curve for specific procedures/skills for residents? 3) Do the data support facilitating the learning curve for specific procedures/skills for practicing urologists? 4) Do the data support translation of skills to practice and improved patient outcomes? Thus, a major shift occurred from validation of the simulator itself to interpretation of the data/scores of simulation based curricula to an intended use for a specific population and purpose. In addition, validation is now considered an ongoing process and not a target that is obtained or achieved once assessments are developed. Therefore, validity is initiated at commencement of the design and continues throughout the development and implementation processes. More appropriately, would be to state that there is "validity evidence" for a specific purpose (eg training) and for whom (eg residents).[6,7]

Second, according to the 2014 Standards for Educational and Psychological Testing,[4] the validation process starts with building a "conceptual framework," delineating the knowledge, skills, abilities, traits, interests and competencies that either support or refute the construct. In order to build a validity argument, a clear hypothesis, which is the proposed inference, should first be declared. Then, the weakest, not the strongest, assumption(s) is tested to either accept or refute the hypothesis (fig.1). Therefore, questions such as, "what do we want to assess?", "what should be included in the simulation based curriculum to conduct this assessment?" and "what is the decision(s) that has

0022-5347/18/1993-0617/0
THE JOURNAL OF UROLOGY®

https://doi.org/10.1016/j.juro.2017.10.022
Vol. 199, 617-620, March 2018
Printed in U.S.A.

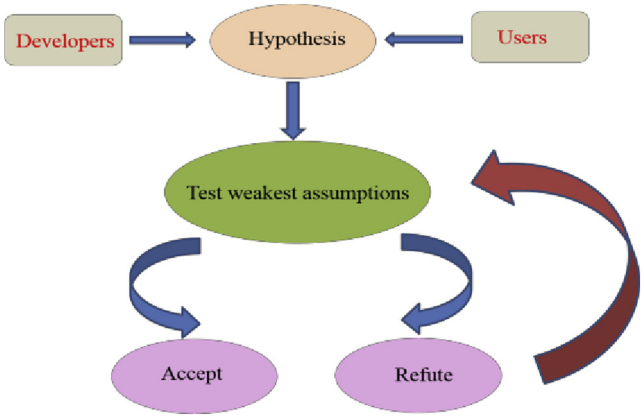

Figure 1. Building argument for validity

to be made based on this/these assessments?" should be asked. As with other scientific studies, the same questions should be asked when interpreting the literature to determine if the results of a given study are aligned with the application the reader has in mind, ie similarity in subject population and specific educational learning/assessment goals and objectives.

Next, the appropriate validity evidence must be collected. According to the latest standards,[4] the 5 sources of validity evidence that collectively challenge the degree to which accumulated evidence supports an intended interpretation of simulator data/scores for proposed use are content, response processes, internal structure, relations with other variables and consequences (fig. 2).

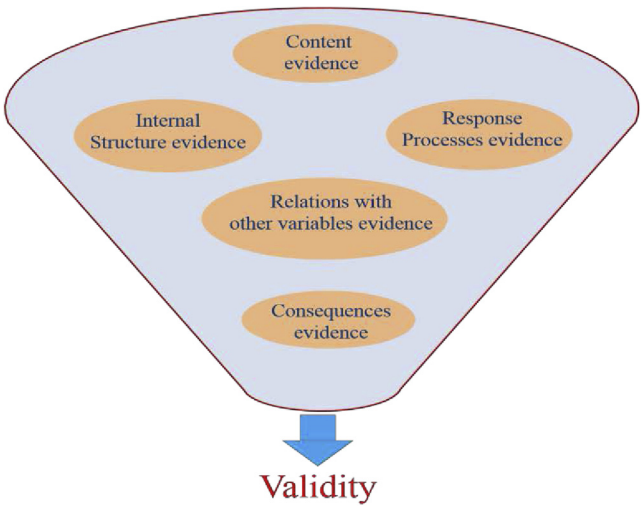

Figure 2. Validity evidence. Modified from Dogan E: The future is here: the (new) Standards for Educational and Psychological Testing. Available at https://www.niss.org/sites/default/files/news_attachments/New%20standards%20Jan%2012%202015%20final_0.pdf. Accessed February 15, 2017.

Content

This source evaluates the relationship between the content of the simulation based curriculum and the construct being examined. This type of evidence could be provided based on logical or empirical analyses about how adequately training/assessment content represents the domain being measured, trained or assessed. Often, content is derived from consensus conferences and clinical guidelines. A concerted effort should be made to focus on content items with evidence-based metrics and avoid technique or dogmatic content in the curriculum.

During collection of content evidence, identification of areas of construct underrepresentation where "the simulator fails to capture important aspects of the construct" and areas of construct irrelevance where "the simulator scores affected by processes that are not part of the construct" can be defined. In an iterative process these areas are formative to the design of curricula whereby items could be added or removed accordingly.

Response Processes

Response processes evidence refers mainly to the extent to which the performance of those assessed (eg trainees) and actions/interpretations of whomever assesses (eg raters) align with the intended construct (eg psychomotor skills). Analyzing individual responses of trainees about their performance strategies for certain simulator tasks can enrich the definition of the construct. For instance, if the individual response processes for a psychomotor task greatly differ among trainees and do not align with raters, developers should reexamine certain items or formats of this task. This can be assessed by documenting and tracking the different aspects of trainees' performance and/or raters' thoughts/actions during the performance and performing theoretical and empirical analyses. Documentation of test security and rater training processes is useful.

Internal Structure

Internal structure evidence focuses on whether the specific items, individual factor (eg learner demographics) and/or clustered composite scores around skills objectives are consistent, reproducible and align with the construct. Therefore, internal structure evidence includes the reliability and some elements of generalizability of simulation derived data/scores. Reliability of simulator data/scores is optimally assessed during technical skills assessment as the scores should be consistent.

Relations with Other Variables

Relations with other variables is a widely used source of validity evidence. It is statistical and correlational, and involves measuring assessment scores of interest