Abstract:
Simulation is becoming a standard assessment modality in pediatric emergency medicine, but its use for high-stakes assessment has not been well described. We aimed to explore literature pertinent to the use of simulation for high-stakes assessment and describe applicable assessment instruments. In this article, we describe potential means by which simulation can be used in a high-stakes manner, along with future developments in assessment methodology for pediatric emergency medicine. A wide array of potentially useful simulation-based assessment instruments exists, although further validity evidence will be needed for their use to be recommended in most cases. Although many simulation-based assessment modalities exist, the evidence is limited for the majority of them. Care must be taken when choosing an appropriate instrument.

Keywords:
assessment; high stakes; pediatric emergency medicine; emergency medicine; simulation; summative assessment; summative; validity; validation; simulation-based medical education

*University of Louisville School of Medicine, Louisville, Kentucky, USA;
†Faculty of Medicine, McGill University, Montreal, Quebec, Canada;
‡McMaster University School of Medicine, Hamilton, Ontario, Canada;
§University of British Columbia, Vancouver, British Columbia, Canada.
Reprint requests and correspondence: Aaron W Calhoun, MD, Division of Pediatric Critical Care Medicine, 571 S Floyd St, STE 332, Louisville, KY 40202.
aaron.calhoun@louisville.edu

1522-8401

Simulation for High-Stakes Assessment in Pediatric Emergency Medicine

Aaron W. Calhoun, MD*, Farhan Bhanji, MD, MSc (Ed), FRCPC†, Jonathan Sherbino, MD, MEd‡, Rose Hatala, MD§

Over the past decades, graduate medical training has shifted away from traditional time-based models toward a competency-based approach in which learners must demonstrate achievement of the competencies (ie, abilities) necessary to practice independently before they can be certified.[1] Determining when residents have achieved these competencies requires a programmatic approach to assessment that uses reliable and valid methods. In the fields of emergency medicine (EM) and pediatric emergency medicine (PEM), simulation has been increasingly used as a critical part of this assessment program.[2,3]

In this article, we provide a critical review of the literature addressing the use of simulation for high-stakes assessment in PEM with the goals of describing existing assessment instruments that may be useful for this purpose and commenting on future directions and innovations.[4]

WHAT MAKES A GOOD ASSESSMENT: THE NATURE OF VALIDITY

How do we know whether an assessment is appropriate for making decisions about the progress of learners? A recent consensus statement details several criteria for "good assessment," including the validity and reproducibility of the assessment's scores, its equivalence across different institutions or cycles of testing, its feasibility and acceptability to key stakeholders (eg, the learners, teachers and educational institutions, and society), the effect of preparation for the assessment on a learner's education, and its ability to catalyze future learning.[5] These criteria clearly illustrate that the process of assessment is not simply about quantifying learning but that it also contributes to learning.[6] This approach also acknowledges that an assessment's value is not based solely on psychometric or statistical features but depends on creating an argument for a particular use in a particular environment.[5]

We next focus specifically on the concept of validity evidence. One common misunderstanding is that once an instrument is "validated," it can be generalized to multiple environments. Validity, however, is not a property of a specific instrument but rather a relationship between an intended decision, an instrument, a specific group of learners, and an environment of use. Thus, when an assessment methodology is applied to a new context, its validity must be reassessed. The strength of validity evidence is of particular importance for high-stakes assessments, as these decisions have more significant consequences for learners and must be defensible.[7] These concepts also apply to comprehensive programs of assessment that include multiple instruments connected by a systematic process.

As an example, consider a third-year PEM fellow in his or her last month of training before graduation, who, as part of the criteria for completion, needs to demonstrate competency at independently leading resuscitations. In your program, you currently assess this competency in the simulation laboratory using a global rating scale of 1 to 5, with 3 representing "competent" resuscitation performance. Deciding whether or not the global rating scale has sufficient validity evidence to support this competency assessment requires a number of steps.

The validation process begins by articulating the decision that will be made with the assessment results (also called the *intended use* of the instrument).[8,9] As an illustration, imagine the global rating of resuscitation skill described above used in 2 settings: at the start of fellowship training to ensure a fellow is safe to lead a resuscitation under supervision in the ED and as the final competency assessment as part of an examination certifying the fellow for independent practice. In these 2 contexts, the same score of "3" (defined in this context as acceptable performance) will imply 2 very different things and lead to 2 very different decisions. Because of this, different types and levels of validity evidence will be needed to support each potential use.[8–10]

When considering the specific types of evidence that contribute to the validity argument, the use of a framework can assist in categorizing the possibilities. In this article, we primarily adopt Messick's 5 sources of evidence, although other frameworks exist.[10,11] The first source of evidence is content: does the test content reflect the underlying construct it intends to measure? The next, response process, refers to test security and quality control as well as raters' understanding of scoring. Internal structure concerns the reliability of the instrument and contains the typical psychometric evidence presented in assessment studies including internal consistency (do all questions assess the same construct?) and interrater reliability (do scores remain stable between raters?). These can be quantified using a variety of statistical methods, including Cronbach α for internal consistency and Cohen κ, or other correlation coefficients (Pearson, Spearman, intraclass) for interrater reliability, or a generalizability study. *Relation with other variables* refers to relationships that may exist between the scores of the assessment and other performance measures. *Consequence* refers to the impact of the assessment on the learner, educational system, or patient.[11] Once the most salient threads of evidence are selected, the data are then synthesized into an argument supporting the use of the instrument in the proposed manner, a principle articulated clearly by Kane.[8,10]

From the example above, content and response process could best be supported by clearly describing how the instrument was developed from our current best knowledge of resuscitation practice (content) and how the individual items and rating scales reflect the decision that the instrument was created to assist (response process). A pilot study, using multiple raters per learner and conducted in the simulation-based context where the instrument will be used, could next be performed. Care must be taken at this phase to assure that the simulated environment and case allow for adequate demonstration of the skills to be assessed. The data from this study can then be used to calculate measures of