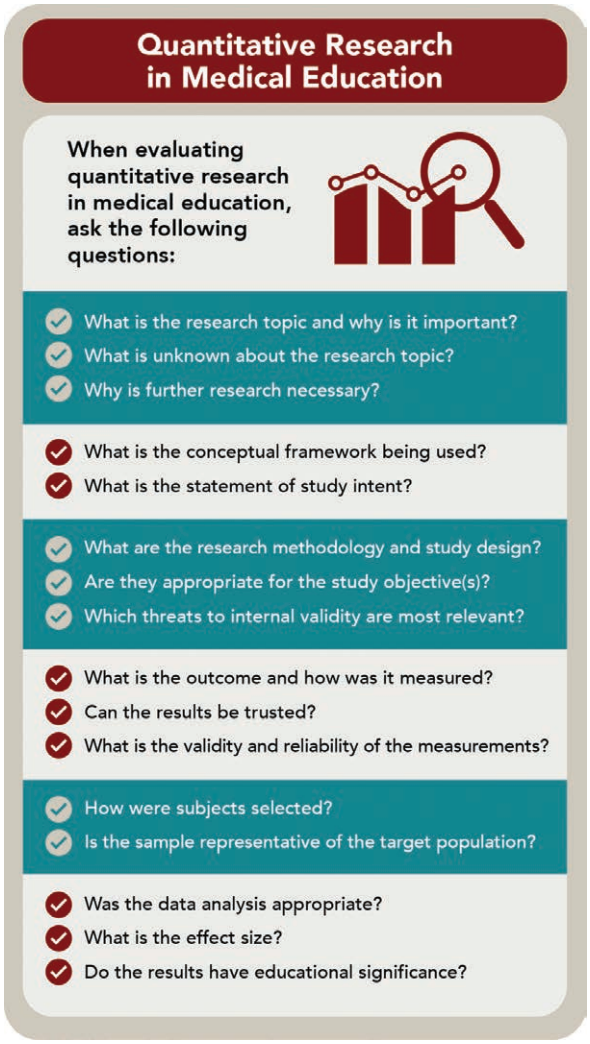
Quantitative Research Methods in Medical Education

John T. Ratelle, M.D., Adam P. Sawatsky, M.D., M.S., Thomas J. Beckman, M.D.

(ANESTHESIOLOGY 2019; 131:23–35)



SUMMARY

There has been a dramatic growth of scholarly articles in medical education in recent years. Evaluating medical education research requires specific orientation to issues related to format and content. Our goal is to review the quantitative aspects of research in medical education so that clinicians may understand these articles with respect to framing the study, recognizing methodologic issues, and utilizing instruments for evaluating the quality of medical education research. This review can be used both as a tool when appraising medical education research articles and as a primer for clinicians interested in pursuing scholarship in medical education.

(ANESTHESIOLOGY 2019; 131:23–35)

There has been an explosion of research in the field of medical education. A search of PubMed demonstrates that more than 40,000 articles have been indexed under the medical subject heading "Medical Education" since 2010, which is more than the total number of articles indexed under this heading in the 1980s and 1990s combined. Keeping up to date requires that practicing clinicians have the skills to interpret and appraise the quality of research articles, especially when serving as editors, reviewers, and consumers of the literature.

While medical education shares many characteristics with other biomedical fields, substantial particularities exist. We recognize that practicing clinicians may not be familiar with the nuances of education research and how to assess its quality. Therefore, our purpose is to provide a review of quantitative research methodologies in medical education. Specifically, we describe a structure that can be used when conducting or evaluating medical education research articles.

Clarifying the Research Purpose

Clarifying the research purpose is an essential first step when reading or conducting scholarship in medical education.[1] Medical education research can serve a variety of purposes, from advancing the science of learning to improving the outcomes of medical trainees and the patients they care for. However, a well-designed study has limited value if it addresses vague, redundant, or unimportant medical education research questions.

Fortunately, there are steps to ensure that the purpose of a research study is clear and logical. Table 1[2–5] outlines these steps,

Image: J. P. Rathmell and Terri Navarette.

Submitted for publication January 8, 2018. Accepted for publication November 29, 2018. From the Division of Hospital Internal Medicine (J.T.R.), and Division of General Internal Medicine (A.P.S., T.J.B.), Mayo Clinic College of Medicine and Science, Department of Medicine, Mayo Clinic, Rochester, Minnesota.

<zdoi;10.1097/ALN.0000000000002727>

Box 1. What to Look for in Research Using This Method

When evaluating qualitative research in medical education, ask the following questions:

1. What is the research topic and why is it important? What is unknown about the research topic? Why is further research necessary?
2. What is the conceptual framework being used to approach the study?
3. What is the statement of study intent?
4. What are the research methodology and study design? Are they appropriate for the study objective(s)?
5. Which threats to internal validity are most relevant for the study?
6. What is the outcome and how was it measured?
7. Can the results be trusted? What is the validity and reliability of the measurements?
8. How were research subjects selected? Is the research sample representative of the target population?
9. Was the data analysis appropriate for the study design and type of data?
10. What is the effect size? Do the results have educational significance?

which will be described in detail in the following sections. We describe these elements not as a simple "checklist," but as an advanced organizer that can be used to understand a medical education research study. These steps can also be used by clinician educators who are new to the field of education research and who wish to conduct scholarship in medical education.

Literature Review and Problem Statement

A literature review is the first step in clarifying the purpose of a medical education research article.[2,5,6] When conducting scholarship in medical education, a literature review helps researchers develop an understanding of their topic of interest. This understanding includes both existing knowledge about the topic as well as key gaps in the literature, which aids the researcher in refining their study question. Additionally, a literature review helps researchers identify conceptual frameworks that have been used to approach the research topic.[2]

When reading scholarship in medical education, a successful literature review provides background information so that even someone unfamiliar with the research topic can understand the rationale for the study. Located in the introduction of the manuscript, the literature review guides the reader through what is already known in a manner that highlights the importance of the research topic. The literature review should also identify key gaps in the literature so the reader can understand the need for further research. This gap description includes an explicit problem statement that summarizes the important issues and provides a reason for the study.[2,4] The following is one example of a problem statement:

> "Identifying gaps in the competency of anesthesia residents in time for intervention is critical to patient safety and an effective learning system… [However], few available instruments relate to complex behavioral performance or provide descriptors…that could inform subsequent feedback, individualized teaching, remediation, and curriculum revision."[7]

This problem statement articulates the research topic (identifying resident performance gaps), why it is important (to intervene for the sake of learning and patient safety), and current gaps in the literature (few tools are available to assess resident performance). The researchers have now underscored why further research is needed and have helped readers anticipate the overarching goals of their study (to develop an instrument to measure anesthesiology resident performance).[4]

The Conceptual Framework

Following the literature review and articulation of the problem statement, the next step in clarifying the research

Table 1. Steps in Clarifying the Purpose of a Research Study in Medical Education

Step	Description and Aims
Literature review	• Outline what is known and unknown about the research topic. • Highlight the importance of the research topic. • Provide a background for readers unfamiliar with the research topic. • Identify existing approaches used to study the research topic.
Problem statement	• Culmination of the literature review. • Articulates key gaps in the literature and why further research is needed. • Identifies the problem to be addressed by the research study.
Conceptual framework	• Serves as an approach to understanding the research topic or solving the research problem. • Can be a theory, principle, or model used to understand the research topic. • Can be an evidence-based best practice that can be applied to the research topic.
Statement of study tntent	• Makes the research purpose explicit. • Identifies the study context and population, study variables, and hypothesized relationship between variables. • Can be a research objective, hypothesis, or question.

See Beckman and Cook, 2007[2]; Bordage, 2009[3]; Dine *et al.*, 2015[4]; and Hart, 2018.[5]

purpose is to select a conceptual framework that can be applied to the research topic. Conceptual frameworks are "ways of thinking about a problem or a study, or ways of representing how complex things work."[3] Just as clinical trials are informed by basic science research in the laboratory, conceptual frameworks often serve as the "basic science" that informs scholarship in medical education. At a fundamental level, conceptual frameworks provide a structured approach to solving the problem identified in the problem statement.

Conceptual frameworks may take the form of theories, principles, or models that help to explain the research problem by identifying its essential variables or elements. Alternatively, conceptual frameworks may represent evidence-based best practices that researchers can apply to an issue identified in the problem statement.[3] Importantly, there is no single best conceptual framework for a particular research topic, although the choice of a conceptual framework is often informed by the literature review and knowing which conceptual frameworks have been used in similar research.[8] For further information on selecting a conceptual framework for research in medical education, we direct readers to the work of Bordage[3] and Irby *et al.*[9]

To illustrate how different conceptual frameworks can be applied to a research problem, suppose you encounter a study to reduce the frequency of communication errors among anesthesiology residents during day-to-night handoff. Table 2[10,11] identifies two different conceptual frameworks researchers might use to approach the task. The first framework, cognitive load theory, has been proposed as a conceptual framework to identify potential variables that may lead to handoff errors.[12] Specifically, cognitive load theory identifies the three factors that affect short-term memory and thus may lead to communication errors:

- Intrinsic load: Inherent complexity or difficulty of the information the resident is trying to learn (*e.g.*, complex patients).
- Extraneous load: Distractions or demands on short-term memory that are not related to the information the resident is trying to learn (*e.g.*, background noise, interruptions).
- Germane load: Effort or mental strategies used by the resident to organize and understand the information he/she is trying to learn (*e.g.*, teach back, note taking).

Using cognitive load theory as a conceptual framework, researchers may design an intervention to reduce extraneous load and help the resident remember the overnight to-do's. An example might be dedicated, pager-free handoff times where distractions are minimized.

The second framework identified in table 2, the I-PASS (Illness severity, Patient summary, Action list, Situational awareness and contingency planning, and Synthesis by receiver) handoff mnemonic,[11] is an evidence-based best practice that, when incorporated as part of a handoff bundle, has been shown to reduce handoff errors on pediatric wards.[13] Researchers choosing this conceptual framework may adapt some or all of the I-PASS elements for resident handoffs in the intensive care unit.

Note that both of the conceptual frameworks outlined above provide researchers with a structured approach to addressing the issue of handoff errors; one is not necessarily better than the other. Indeed, it is possible for researchers to use both frameworks when designing their study. Ultimately, we provide this example to demonstrate the necessity of selecting conceptual frameworks to clarify the research purpose.[3,8] Readers should look for conceptual frameworks in the introduction section and should be wary of their omission, as commonly seen in less well-developed medical education research articles.[14]

Statement of Study Intent

After reviewing the literature, articulating the problem statement, and selecting a conceptual framework to address the research topic, the final step in clarifying the research purpose is the statement of study intent. The statement of study intent is arguably the most important element of framing the study because it makes the research purpose explicit.[2] Consider the following example:

> This study aimed to test the hypothesis that the introduction of the BASIC Examination was associated

Table 2. Conceptual Frameworks to Address the Issue of Handoff Errors in the Intensive Care Unit

Conceptual Framework	Type of Framework	Key Elements	Example Intervention
Cognitive load theory[12]	Theory of working memory used to identify variables that may lead to communication errors during handoff.	• Intrinsic load (complexity of the patients being discussed during handoff) • Extraneous load (distraction during handoff) • Germane load (mental strategies, such as note taking or "read-back" used by handoff receiver)	Dedicated pager-free time during shift handoffs to reduce extraneous load.
I-PASS mnemonic[13]	Evidence-based best practice for reducing communication errors during handoff.	• I-Illness severity • P-Patient summary • A-Action list • S-Situational awareness and contingency planning • S-Synthesis by receiver	Implement or adapt the I-PASS handoff mnemonic in the intensive care unit.

I-PASS, Illness severity, Patient summary, Action list, Situational awareness and contingency planning, and Synthesis by receiver.

> with an accelerated knowledge acquisition during residency training, as measured by increments in annual ITE scores.[15]

This statement of study intent succinctly identifies several key study elements including the population (anesthesiology residents), the intervention/independent variable (introduction of the BASIC Examination), the outcome/dependent variable (knowledge acquisition, as measure by in In-training Examination [ITE] scores), and the hypothesized relationship between the independent and dependent variable (the authors hypothesize a positive correlation between the BASIC examination and the speed of knowledge acquisition).[6,14]

The statement of study intent will sometimes manifest as a research objective, rather than hypothesis or question. In such instances there may not be explicit independent and dependent variables, but the study population and research aim should be clearly identified. The following is an example:

> "In this report, we present the results of 3 [years] of course data with respect to the practice improvements proposed by participating anesthesiologists and their success in implementing those plans. Specifically, our primary aim is to assess the frequency and type of improvements that were completed and any factors that influence completion."[16]

The statement of study intent is the logical culmination of the literature review, problem statement, and conceptual framework, and is a transition point between the Introduction and Methods sections of a medical education research report. Nonetheless, a systematic review of experimental research in medical education demonstrated that statements of study intent are absent in the majority of articles.[14] When reading a medical education research article where the statement of study intent is absent, it may be necessary to infer the research aim by gathering information from the Introduction and Methods sections. In these cases, it can be useful to identify the following key elements[6,14,17]:

1. Population of interest/type of learner (*e.g.*, pain medicine fellow or anesthesiology residents)
2. Independent/predictor variable (*e.g.*, educational intervention or characteristic of the learners)
3. Dependent/outcome variable (*e.g.*, intubation skills or knowledge of anesthetic agents)
4. Relationship between the variables (*e.g.*, "improve" or "mitigate")

Occasionally, it may be difficult to differentiate the independent study variable from the dependent study variable.[17] For example, consider a study aiming to measure the relationship between burnout and personal debt among anesthesiology residents. Do the researchers believe burnout might lead to high personal debt, or that high personal debt may lead to burnout? This "chicken or egg" conundrum reinforces the importance of the conceptual framework which, if present, should serve as an explanation or rationale for the predicted relationship between study variables.

Methodology

Research methodology is the "…design or plan that shapes the methods to be used in a study."[1] Essentially, *methodology* is the general strategy for answering a research question, whereas *methods* are the specific steps and techniques that are used to collect data and implement the strategy. Our objective here is to provide an overview of quantitative methodologies (*i.e.*, approaches) in medical education research.

The choice of research methodology is made by balancing the approach that best answers the research question against the feasibility of completing the study. There is no perfect methodology because each has its own potential caveats, flaws and/or sources of bias. Before delving into an overview of the methodologies, it is important to highlight common sources of bias in education research. We use the term *internal validity* to describe the degree to which the findings of a research study represent "the truth," as opposed to some alternative hypothesis or variables.[18] Table 3[18–20] provides a list of common threats to internal validity in medical education research, along with tactics to mitigate these threats.

Experimental Research

The fundamental tenet of experimental research is the manipulation of an independent or experimental variable to measure its effect on a dependent or outcome variable.

True Experiment

True experimental study designs minimize threats to internal validity by randomizing study subjects to experimental and control groups. Through ensuring that differences between groups are—beyond the intervention/variable of interest—purely due to chance, researchers reduce the internal validity threats related to subject characteristics, time-related maturation, and regression to the mean.[18,19]

Quasi-experiment

There are many instances in medical education where randomization may not be feasible or ethical. For instance, researchers wanting to test the effect of a new curriculum among medical students may not be able to randomize learners due to competing curricular obligations and schedules. In these cases, researchers may be forced to assign subjects to experimental and control groups based upon some other criterion beyond randomization, such as different classrooms or different sections of the same course. This process, called quasi-randomization, does not inherently lead to internal validity threats, as long as research investigators are mindful of measuring and controlling for extraneous variables between study groups.[19]

Table 3. Threats to Internal Validity and Strategies to Mitigate Their Effects

Threat	Description	Strategies to Mitigate the Threat
Subject characteristics (*i.e.*, selection bias)	Differences in study subjects beyond the variables being studied.	• Collect information on study subjects. • Use randomization.
Mortality	Loss of study subjects.	• Collect baseline information on study subjects. • Identify and compare reasons for missing data between study groups (if applicable). • Use multiple imputation to account for missing data (if applicable).
Location	The location of the study may influence the results. May be related to differences in educational resources or learning environment unrelated to study variables.	• Collect information on study setting. • Standardize study conditions.
Instrumentation	Variation in test administration and/or scoring may influence study results.	• Collect information on study methods including test administration and data collection. • Standardize study conditions.
Testing	A pretest at the outset of a study, in and of itself, affects the results (*e.g.*, practice effect)	• Avoid pretest. • Use "post-then-pre" assessment. (At the conclusion of the study, subjects rate the outcome as a result of intervention. Then, subjects are asked to rate the outcome prior to the intervention.)[20]
History	Unplanned influences or events outside of the study variables can affect the results.	• Collect information on study methods, including unforeseen influences or deviations. • Standardize study conditions.
Maturation	Study subjects may experience change related to passage of time, rather than being related to study variables.	• Collect information on study subjects. • Use randomization.
Subject attitude	The attitude of the study subject toward participation in the study may affect the results (*e.g.*, Hawthorne effect).	• Standardize conditions ("blinding," if possible). • Collect information on study subjects, including baseline attitudes. • Use randomization.
Regression	Study subjects with extremely low or high performance at baseline are likely to "regress" to the mean on subsequent testing, regardless of the intervention.	• Collect information on study subjects. • Use randomization.
Implementation	Variation in the treatment or experience of study subjects in experimental *vs.* control groups, outside of the intervention.	• Standardize conditions. • Collect information on the experience of study subjects, both within and without of the research intervention. • Collect information on potential cointerventions.

Adapted with permission from Cook and Beckman, 2010.[19]

Single-group Methodologies

All experimental study designs compare two or more groups: experimental and control. A common experimental study design in medical education research is the single-group pretest–posttest design, which compares a group of learners before and after the implementation of an intervention.[21] In essence, a single-group pre–post design compares an experimental group (*i.e.*, postintervention) to a "no-intervention" control group (*i.e.*, preintervention).[19] This study design is problematic for several reasons. Consider the following hypothetical example: A research article reports the effects of a year-long intubation curriculum for first-year anesthesiology residents. All residents participate in monthly, half-day workshops over the course of an academic year. The article reports a positive effect on residents' skills as demonstrated by a significant improvement in intubation success rates at the end of the year when compared to the beginning.

This study does little to advance the science of learning among anesthesiology residents. While this hypothetical report demonstrates an improvement in residents' intubation success before *versus* after the intervention, it does not tell why the workshop worked, how it compares to other educational interventions, or how it fits in to the broader picture of anesthesia training.

Single-group pre–post study designs open themselves to a myriad of threats to internal validity.[20] In our hypothetical example, the improvement in residents' intubation skills may have been due to other educational experience(s) (*i.e.*, implementation threat) and/or improvement in manual dexterity that occurred naturally with time (*i.e.*, maturation threat), rather than the airway curriculum. Consequently, single-group pre–post studies should be interpreted with caution.[18]

Repeated testing, before and after the intervention, is one strategy that can be used to reduce the some of the inherent limitations of the single-group study design. Repeated pretesting can mitigate the effect of regression toward the mean, a statistical phenomenon whereby low pretest scores tend to move closer to the mean on subsequent testing (regardless of intervention).[20] Likewise, repeated posttesting at multiple time intervals can provide potentially useful information about the short- and long-term effects of an intervention (*e.g.*, the "durability" of the gain in knowledge, skill, or attitude).

Observational Research

Unlike experimental studies, observational research does not involve manipulation of any variables. These studies often involve measuring associations, developing psychometric instruments, or conducting surveys.

Association Research

Association research seeks to identify relationships between two or more variables within a group or groups (correlational research), or similarities/differences between two or more existing groups (causal–comparative research). For example, correlational research might seek to measure the relationship between burnout and educational debt among anesthesiology residents, while causal–comparative research may seek to measure differences in educational debt and/or burnout between anesthesiology and surgery residents. Notably, association research may identify relationships between variables, but does not necessarily support a causal relationship between them.

Psychometric and Survey Research

Psychometric instruments measure a psychologic or cognitive construct such as knowledge, satisfaction, beliefs, and symptoms. Surveys are one type of psychometric instrument, but many other types exist, such as evaluations of direct observation, written examinations, or screening tools.[22] Psychometric instruments are ubiquitous in medical education research and can be used to describe a trait within a study population (*e.g.*, rates of depression among medical students) or to measure associations between study variables (*e.g.*, association between depression and board scores among medical students).

Psychometric and survey research studies are prone to the internal validity threats listed in table 3, particularly those relating to mortality, location, and instrumentation.[18] Additionally, readers must ensure that the instrument scores can be trusted to truly represent the construct being measured. For example, suppose you encounter a research article demonstrating a positive association between attending physician teaching effectiveness as measured by a survey of medical students, and the frequency with which the attending physician provides coffee and doughnuts on rounds. Can we be confident that this survey administered to medical students is truly measuring teaching effectiveness? Or is it simply measuring the attending physician's "likability"? Issues related to measurement and the trustworthiness of data are described in detail in the following section on measurement and the related issues of validity and reliability.

Measurement

Measurement refers to "the assigning of numbers to individuals in a systematic way as a means of representing properties of the individuals."[23] Research data can only be trusted insofar as we trust the measurement used to obtain the data. Measurement is of particular importance in medical education research because many of the constructs being measured (*e.g.*, knowledge, skill, attitudes) are abstract and subject to measurement error.[24] This section highlights two specific issues related to the trustworthiness of data: the validity and reliability of measurements.

Validity

Validity regarding the scores of a measurement instrument "refers to the degree to which evidence and theory support the interpretations of the [instrument's results] for the proposed use of the [instrument]."[25] In essence, do we believe the results obtained from a measurement really represent what we were trying to measure? Note that validity evidence for the scores of a measurement instrument is separate from the internal validity of a research study. Several frameworks for validity evidence exist. Table 4[2,22,26] represents the most commonly used framework, developed by Messick,[27] which identifies sources of validity evidence—to support the target construct—from five main categories: content, response process, internal structure, relations to other variables, and consequences.

Reliability

Reliability refers to the consistency of scores for a measurement instrument.[22,25,28] For an instrument to be reliable, we would anticipate that two individuals rating the same object of measurement in a specific context would provide the same scores.[25] Further, if the scores for an instrument are reliable between raters of the same object of measurement, then we can extrapolate that any difference in scores between two objects represents a true difference across the sample, and is not due to random variation in measurement.[29] Reliability can be demonstrated through a variety of methods such as internal consistency (*e.g.*, Cronbach's alpha), temporal stability (*e.g.*, test–retest reliability), interrater agreement (*e.g.*, intraclass correlation coefficient), and generalizability theory (generalizability coefficient).[22,29]

Example of a Validity and Reliability Argument

This section provides an illustration of validity and reliability in medical education. We use the signaling questions outlined in table 4 to make a validity and reliability argument for the Harvard Assessment of Anesthesia Resident Performance (HARP) instrument.[7] The HARP was developed by Blum *et al.* to measure the performance of anesthesia trainees that is required to provide safe anesthetic care to patients. According to the authors, the HARP is designed to be used "…as part of a multiscenario, simulation-based assessment" of resident performance.[7]

Content Validity: Does the Instrument's Content Represent the Construct Being Measured?

To demonstrate content validity, instrument developers should describe the construct being measured and how the instrument was developed, and justify their approach.[25] The HARP is intended to measure resident performance in the

Table 4. Sources of Validity Evidence for Measurement Instruments

Category	Signaling Question(s)	Sample Sources	Example
Content	Does the instrument's content represent the construct being measured?	• Develop content based upon conceptual frameworks, literature review, and/or previous instruments. • Have content reviewed and modified by experts in the field of the construct.	Develop an instrument to measure nurse-physician teamwork in the operating room from theoretical models of teamwork in other fields.
Response process	Are raters interpreting the instrument items as intended?	• Ask subjects to "think aloud" while completing the instrument. • Monitor response times to various items and compare between subjects.	Students verbalized the rationale for their responses when filling out a new survey to measure teaching effectiveness of their attending physician. Items were reworded to ensure students were focused purely on teaching effectiveness, not likeability of the attending.
Internal structure	Do instrument items measuring similar constructs yield homogeneous results? Do instrument items measuring different constructs yield heterogeneous results?	• Factor analysis. • Internal consistency reliability (*e.g.* Cronbach's alpha).	Factor analysis demonstrates that scores from a five-item measure resident well-being form a single factor, and that the internal consistency reliability of the item scores is high (Cronbach alpha = 0.95)
Relationships to other variables	Do instrument scores correlate with other measures of similar or different constructs as expected?	• Demonstrate positive correlations between the instrument and other measures of the same construct, or negative correlations with instruments that measure dissimilar constructs. • Instrument scores predict outcomes related to the construct.	A new instrument to measure depression among medical students correlates with student dropout rates.
Consequence	Are instrument results being used as intended? Are there unintended or negative uses of the instrument results?	• Report the anticipated and unanticipated impact of instrument results.	A new screening tool for depression among medical students leads to creation of new support system and counseling resources for depressed students.

Adapted with permission from Beckman, 2007.[2]

critical domains required to provide safe anesthetic care. As such, investigators note that the HARP items were created through a two-step process. First, the instrument's developers interviewed anesthesiologists with experience in resident education to identify the key traits needed for successful completion of anesthesia residency training. Second, the authors used a modified Delphi process to synthesize the responses into five key behaviors: (1) formulate a clear anesthetic plan, (2) modify the plan under changing conditions, (3) communicate effectively, (4) identify performance improvement opportunities, and (5) recognize one's limits.[7,30]

Response Process Validity: Are Raters Interpreting the Instrument Items as Intended?

In the case of the HARP, the developers included a scoring rubric with behavioral anchors to ensure that faculty raters could clearly identify how resident performance in each domain should be scored.[7]

Internal Structure Validity: Do Instrument Items Measuring Similar Constructs Yield Homogenous Results? Do Instrument Items Measuring Different Constructs Yield Heterogeneous Results?

Item-correlation for the HARP demonstrated a high degree of correlation between some items (*e.g.*, formulating a plan and modifying the plan under changing conditions) and a lower degree of correlation between other items (*e.g.*, formulating a plan and identifying performance improvement opportunities).[30] This finding is expected since the items within the HARP are designed to assess separate performance domains, and we would expect residents' functioning to vary across domains.

Relationship to Other Variables' Validity: Do Instrument Scores Correlate with Other Measures of Similar or Different Constructs as Expected?

As it applies to the HARP, one would expect that the performance of anesthesia residents will improve over the course of training. Indeed, HARP scores were found to be generally higher among third-year residents compared to first-year residents.[30]

Consequence Validity: Are Instrument Results Being Used as Intended? Are There Unintended or Negative Uses of the Instrument Results?

While investigators did not intentionally seek out consequence validity evidence for the HARP, unanticipated consequences of HARP scores were identified by the authors as follows:

> "Data indicated that CA-3s had a lower percentage of worrisome scores (rating 2 or lower) than CA-1s… However, it is concerning that any CA-3s had *any* worrisome scores…low performance of some CA-3 residents, albeit in the simulated environment, suggests opportunities for training improvement."[30]

That is, using the HARP to measure the performance of CA-3 anesthesia residents had the unintended consequence of identifying the need for improvement in resident training.

Reliability: Are the Instrument's Scores Reproducible and Consistent between Raters?

The HARP was applied by two raters for every resident in the study across seven different simulation scenarios. The investigators conducted a generalizability study of HARP scores to estimate the variance in assessment scores that was due to the resident, the rater, and the scenario. They found little variance was due to the rater (*i.e.*, scores were consistent between raters), indicating a high level of reliability.[7]

Sampling

Sampling refers to the selection of research subjects (*i.e.*, the sample) from a larger group of eligible individuals (*i.e.*, the population).[31] Effective sampling leads to the inclusion of research subjects who represent the larger population of interest. Alternatively, ineffective sampling may lead to the selection of research subjects who are significantly different from the target population. Imagine that researchers want to explore the relationship between burnout and educational debt among pain medicine specialists. The researchers distribute a survey to 1,000 pain medicine specialists (the population), but only 300 individuals complete the survey (the sample). This result is problematic because the characteristics of those individuals who completed the survey and the entire population of pain medicine specialists may be fundamentally different. It is possible that the 300 study subjects may be experiencing more burnout and/or debt, and thus, were more motivated to complete the survey. Alternatively, the 700 nonresponders might have been too busy to respond and even more burned out than the 300 responders, which would suggest that the study findings were even more amplified than actually observed.

When evaluating a medical education research article, it is important to identify the sampling technique the researchers employed, how it might have influenced the results, and whether the results apply to the target population.[24]

Sampling Techniques

Sampling techniques generally fall into two categories: probability- or nonprobability-based. Probability-based sampling ensures that each individual within the target population has an equal opportunity of being selected as a research subject. Most commonly, this is done through random sampling, which should lead to a sample of research subjects that is similar to the target population. If significant differences between sample and population exist, those differences should be due to random chance, rather than systematic bias. The difference between data from a random sample and that from the population is referred to as sampling error.[24]

Nonprobability-based sampling involves selecting research participants such that inclusion of some individuals may be more likely than the inclusion of others.[31] Convenience sampling is one such example and involves selection of research subjects based upon ease or opportuneness. Convenience sampling is common in medical education research, but, as outlined in the example at the beginning of this section, it can lead to sampling bias.[24] When evaluating an article that uses nonprobability-based sampling, it is important to look for participation/response rate. In general, a participation rate of less than 75% should be viewed with skepticism.[21] Additionally, it is important to determine whether characteristics of participants and nonparticipants were reported and if significant differences between the two groups exist.

Data Analysis and Interpretation

Interpreting medical education research requires a basic understanding of common ways in which quantitative data are analyzed and displayed. In this section, we highlight two broad topics that are of particular importance when evaluating research articles.

The Nature of the Measurement Variable

Measurement variables in quantitative research generally fall into three categories: nominal, ordinal, or interval.[24] Nominal variables (sometimes called categorical variables) involve data that can be placed into discrete categories without a specific order or structure. Examples include sex (male or female) and professional degree (M.D., D.O., M.B.B.S., *etc.*) where there is no clear hierarchical order to the categories. Ordinal variables can be ranked according to some criterion, but the spacing between categories may not be equal. Examples of ordinal variables may include measurements of satisfaction (satisfied *vs.* unsatisfied), agreement (disagree *vs.* agree), and educational experience (medical student, resident, fellow). As it applies to educational experience, it is noteworthy that even though education can be quantified in years, the spacing between years (*i.e.*, educational "growth") remains unequal. For instance, the difference in performance between second- and third-year medical students is dramatically different than third- and fourth-year medical students. Interval variables can also be ranked according to some criteria, but, unlike ordinal variables, the spacing between variable categories is equal. Examples of interval variables include test scores and salary. However, the conceptual boundaries between these measurement variables are not always clear, as in the case where ordinal scales can be assumed to have the properties of an interval scale, so long as the data's distribution is not substantially skewed.[32]

Understanding the nature of the measurement variable is important when evaluating how the data are analyzed and reported. Medical education research commonly uses measurement instruments with items that are rated on Likert-type scales, whereby the respondent is asked to assess their level of agreement with a given statement. The response is often translated into a corresponding number (*e.g.*, 1 = strongly disagree, 3 = neutral, 5 = strongly agree). It is remarkable that scores from Likert-type scales are sometimes not normally distributed (*i.e.*, are skewed toward one end of the scale), indicating that the spacing between scores is unequal and the variable is ordinal in nature. In these cases, it is recommended to report results as frequencies or medians, rather than means and SDs.[33]

Consider an article evaluating medical students' satisfaction with a new curriculum. Researchers measure satisfaction using a Likert-type scale (1 = very unsatisfied, 2 = unsatisfied, 3 = neutral, 4 = satisfied, 5 = very satisfied). A total of 20 medical students evaluate the curriculum, 10 of whom rate their satisfaction as "satisfied," and 10 of whom rate it as "very satisfied." In this case, it does not make much sense to report an average score of 4.5; it makes more sense to report results in terms of frequency (*e.g.*, half of the students were "very satisfied" with the curriculum, and half were not).

Effect Size and CIs

In medical education, as in other research disciplines, it is common to report statistically significant results (*i.e.*, small *P* values) in order to increase the likelihood of publication.[34,35] However, a significant *P* value in itself does necessarily represent the educational impact of the study results. A statement like "Intervention x was associated with a significant improvement in learners' intubation skill compared to education intervention y ($P < 0.05$)" tells us that there was a less than 5% chance that the difference in improvement between interventions x and y was due to chance. Yet that does not mean that the study intervention necessarily caused the nonchance results, or indicate whether the between-group difference is educationally significant. Therefore, readers should consider looking beyond the *P* value to effect size and/or CI when interpreting the study results.[36,37]

Effect size is "the magnitude of the difference between two groups," which helps to quantify the educational significance of the research results.[37] Common measures of effect size include Cohen's d (standardized difference between two means), risk ratio (compares binary outcomes between two groups), and Pearson's r correlation (linear relationship between two continuous variables).[37] CIs represent "a range of values around a sample mean or proportion" and are a measure of precision.[31] While effect size and CI give more useful information than simple statistical significance, they are commonly omitted from medical education research articles.[35] In such instances, readers should be wary of overinterpreting a *P* value in isolation. For further information effect size and CI, we direct readers the work of Sullivan and Feinn[37] and Hulley *et al.*[31]

Tools for Evaluating the Quality of Medical Education Research

In this final section, we identify instruments that can be used to evaluate the quality of quantitative medical education research articles. To this point, we have focused on framing the study and research methodologies and identifying potential pitfalls to consider when appraising a specific article. This is important because how a study is framed and the choice of methodology require some subjective interpretation. Fortunately, there are several instruments available for evaluating medical education research methods and providing a structured approach to the evaluation process.

The Medical Education Research Study Quality Instrument (MERSQI)[21] and the Newcastle Ottawa Scale-Education (NOS-E)[38] are two commonly used instruments, both of which have an extensive body of validity evidence to support the interpretation of their scores. Table 5[21,39] provides more detail regarding the MERSQI, which includes evaluation of study design, sampling, data type, validity, data analysis, and outcomes. We have found that applying the MERSQI to manuscripts, articles, and protocols has intrinsic educational value, because this practice of application familiarizes MERSQI users with fundamental principles of medical education research. One aspect of the MERSQI that deserves special mention is the section on evaluating *outcomes* based on Kirkpatrick's widely recognized hierarchy of reaction, learning, behavior, and results (table 5; fig.).[40] Validity evidence for the scores of the MERSQI include its operational definitions to improve response process, excellent reliability, and internal consistency, as well as high correlation with other measures of study quality, likelihood of publication, citation rate, and an association between MERSQI score and the likelihood of study funding.[21,41] Additionally, consequence validity for the MERSQI scores has been demonstrated by its utility for identifying and disseminating high-quality research in medical education.[42]

The NOS-E is a newer tool to evaluate the quality of medication education research. It was developed as a modification of the Newcastle-Ottawa Scale[43] for appraising the quality of nonrandomized studies. The NOS-E includes items focusing on the representativeness of the experimental group, selection and compatibility of the control group, missing data/study retention, and blinding of outcome assessors.[38,39] Additional validity evidence for NOS-E scores includes operational definitions to improve response process, excellent reliability and internal consistency, and its correlation with other measures of study quality.[39] Notably, the complete NOS-E, along with its scoring rubric, can found in the article by Cook and Reed.[39]

A recent comparison of the MERSQI and NOS-E found acceptable interrater reliability and good correlation between the two instruments[39] However, noted differences exist between the MERSQI and NOS-E. Specifically, the MERSQI may be applied to a broad range of study designs, including experimental and cross-sectional research.

Table 5. The Medical Education Research Study Quality Instrument for Evaluating the Quality of Medical Education Research

MERSQI Domain*	MERSQI Items	MERSQI Score†	Comments‡
Study design	Single-group cross-sectional or single group posttest only	1	• Cross-sectional: Study of a single group at one point in time • Nonrandomized two-group: An intervention applied to two separate groups of subjects, who are not randomly assigned
	Single-group pretest and posttest	1.5	
	Nonrandomized, two-group	2	
	Randomized controlled trial	3	
Sampling: Number of institutions studied	1	0.5	• An institution is a separate medical center (*e.g.*, two hospitals within the same academic medical center would *not* be considered separate institutions)
	2	1	
	> 2	1.5	
Sampling: Response rate (%)	Not applicable		• Response rate is the proportion of eligible participants who completed the survey, posttest, *etc.* • For intervention studies, this is the proportion of enrolled participants who completed the intervention • For studies in which there is >1 response rate (*e.g.*, pretest and posttest completions), use the score for the highest response rate
	< 50 or not reported	0.5	
	50–74	1	
	≥ 75	1.5	
Type of data	Assessment by study participant	1	• Observer ratings are considered anything other than assessment by the study subject.
	Objective measurement	3	
Validity evidence for evaluation instrument scores	Not applicable		• Content evidence includes theory, literature, expert opinions, and previous instruments that were used to create the instrument • Internal structure evidence includes reliability (*e.g.*, internal consistency, interrater, test–retest), and measures of dimensionality (*e.g.*, factor analysis) • Relations to other variables evidence includes correlation with other variables that represent similar constructs (concordant) or dissimilar constructs (discordant) • Use "not applicable" only if the study does not measure a psychologic construct
	Content	1	
	Internal structure	1	
	Relationships to other variables	1	
Data analysis: Complexity	Descriptive analysis only	1	• Descriptive analyses include frequency, mean, and median • Any test of statistical inference involving associations or comparisons is considered "beyond descriptive"
	Beyond descriptive analysis	2	
Data analysis: Appropriateness	Data analysis inappropriate for study design and type of data	0	• Considered "inappropriate" if there is a flaw in the analysis that invalidates the results
	Data analysis appropriate for study design and type of data	1	
Outcomes	Satisfaction, attitudes, perceptions, opinions, general facts	1	• General facts include participant characteristics and basic data such as instrument score reliability • Behaviors are actions that occur in real-life practice
	Knowledge, skills	1.5	
	Behaviors	2	
	Patient/healthcare outcomes	3	

*The Medical Education Research Study Quality Instrument (MERSQI) has six domains: Study Design, Sampling, Type of Data, Validity Evidence, Data Analysis, and Outcomes. †The maximum score within each MERSQI domain is 3 and the maximum total score is 18. ‡Comments regarding the MERSQI are intended to provide guidance on how to interpret MERSQI domains and items; for more information on the scoring of specific MERSQI items, please see Reed *et al.*, 2007[21]; adapted with permission from Cook and Reed, 2015.[39]

Additionally, the MERSQI addresses issues related to measurement validity and data analysis, and places emphasis on educational outcomes. On the other hand, the NOS-E focuses specifically on experimental study designs, and on issues related to sampling techniques and outcome assessment.[39] Ultimately, the MERSQI and NOS-E are complementary tools that may be used together when evaluating the quality of medical education research.

Conclusions

This article provides an overview of quantitative research in medical education, underscores the main components of education research, and provides a general framework for evaluating research quality. We highlighted the importance of framing a study with respect to purpose, conceptual framework, and statement of study intent. We reviewed the most common research methodologies, along with threats to the validity of a study and its measurement instruments. Finally, we identified two complementary instruments, the MERSQI and NOS-E, for evaluating the quality of a medical education research study.

Research Support

Support was provided solely from institutional and/or departmental sources.

Fig. Kirkpatrick's hierarchy of outcomes as applied to education research. Reaction = Level 1, Learning = Level 2, Behavior = Level 3, Results = Level 4. Outcomes become more meaningful, yet more difficult to achieve, when progressing from Level 1 through Level 4. Adapted with permission from Beckman and Cook, 2007.[2]

Box 2. Where to Find More Information on This Topic

For more information on quantitative research in medical education:

1. Bordage G: Conceptual frameworks to illuminate and magnify. Medical education. 2009; 43(4):312–9.
2. Cook DA, Beckman TJ: Current concepts in validity and reliability for psychometric instruments: Theory and application. The American journal of medicine. 2006; 119(2):166. e7–166. e116.
3. Franenkel JR, Wallen NE, Hyun HH: How to Design and Evaluate Research in Education. 9th edition. New York, McGraw-Hill Education, 2015.
4. Hulley SB, Cummings SR, Browner WS, Grady DG, Newman TB: Designing clinical research. 4th edition. Philadelphia, Lippincott Williams & Wilkins, 2011.
5. Irby BJ, Brown G, Lara-Alecio R, Jackson S: The Handbook of Educational Theories. Charlotte, NC, Information Age Publishing, Inc., 2015
6. Standards for Educational and Psychological Testing (American Educational Research Association & American Psychological Association, 2014)
7. Swanwick T: Understanding medical education: Evidence, theory and practice, 2nd edition. Wiley-Blackwell, 2013.
8. Sullivan GM, Artino Jr AR: Analyzing and interpreting data from Likert-type scales. Journal of graduate medical education. 2013; 5(4):541–2.
9. Sullivan GM, Feinn R: Using effect size—or why the P value is not enough. Journal of graduate medical education. 2012; 4(3):279–82.
10. Tavakol M, Sandars J: Quantitative and qualitative methods in medical education research: AMEE Guide No 90: Part II. Medical teacher. 2014; 36(10):838–48.

Competing Interests

The authors declare no competing interests.

Correspondence

Address correspondence to Dr. Ratelle: Mayo Clinic Rochester, 200 First Street SW, Rochester, Minnesota 55905. Ratelle.John@mayo.edu. Information on purchasing reprints may be found at www.anesthesiology.org or on the masthead page at the beginning of this issue. ANESTHESIOLOGY's articles are made freely accessible to all readers, for personal use only, 6 months from the cover date of the issue.

References

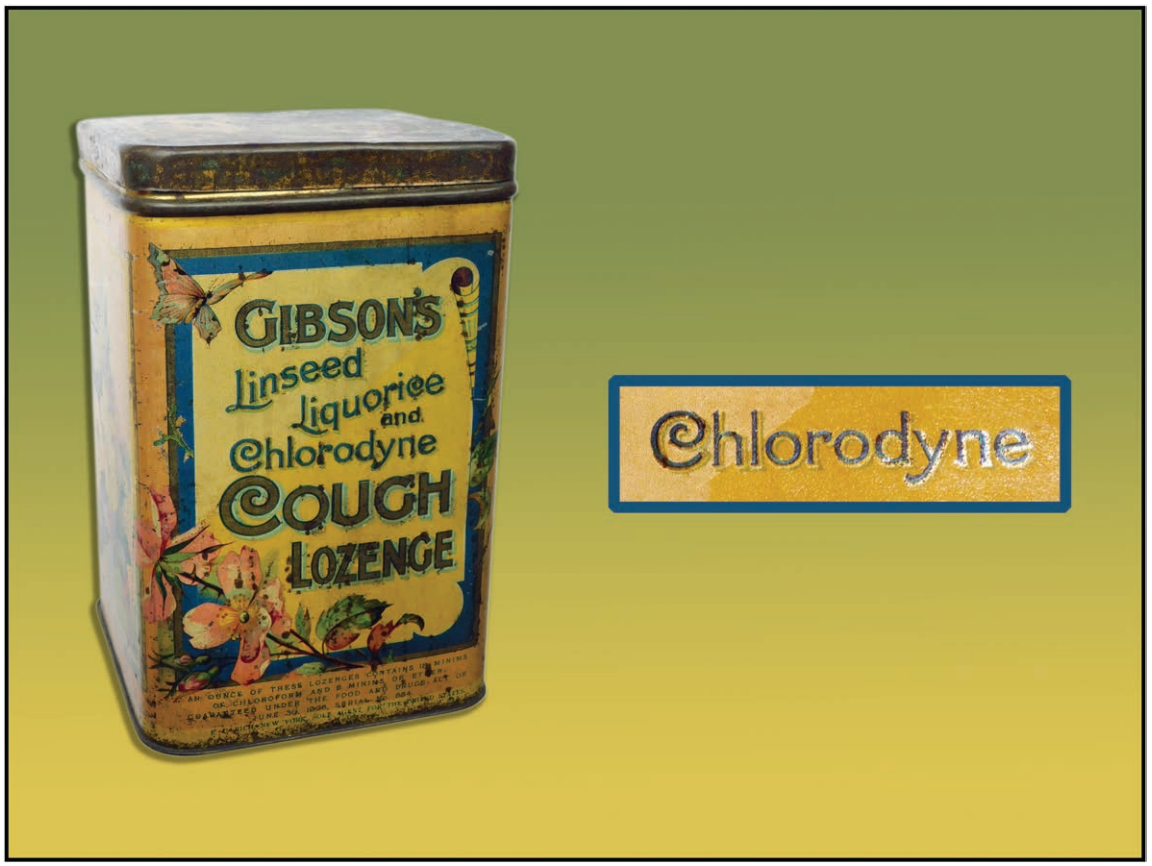
1. Swanwick T: Understanding medical education: Evidence, theory and practice, 2nd edition. Wiley-Blackwell, 2013
2. Beckman TJ, Cook DA: Developing scholarly projects in education: A primer for medical teachers. Med Teach 2007; 29:210–8
3. Bordage G: Conceptual frameworks to illuminate and magnify. Med Educ 2009; 43:312–9
4. Dine CJ, McGaghie WC, Bordage G, Shea JA: Chapter 6: Problem statement, conceptual framework, and research question, Review Criteria for Research Manuscripts, 2nd edition. Edited by Durning SJ, Carline JD. Washington, DC, Association of American Medical Colleges, 2015, pp 19–20
5. Hart C: Doing a Literature Review: Releasing the Research Imagination, 2nd edition. London, SAGE Publications Ltd, 2018
6. Bordage G, Dawson B: Experimental study design and grant writing in eight steps and 28 questions. Med Educ 2003; 37:376–85
7. Blum RH, Boulet JR, Cooper JB, Muret-Wagstaff SL; Harvard Assessment of Anesthesia Resident Performance Research Group: Simulation-based assessment to identify critical gaps in safe anesthesia resident performance. ANESTHESIOLOGY 2014; 120:129–41
8. Bordage G, Lineberry M, Yudkowsky R: Conceptual frameworks to guide research and development (R&D) in health professions education. Acad Med 2016; 91:e2
9. Irby BJ, Brown G, Lara-Alecio R, Jackson S: The Handbook of Educational Theories. Charlotte, NC, Information Age Publishing, Inc., 2015
10. Young JQ, Van Merrienboer J, Durning S, Ten Cate O: Cognitive load theory: Implications for medical education: AMEE Guide No. 86. Med Teach 2014; 36:371–84
11. Starmer AJ, Spector ND, Srivastava R, Allen AD, Landrigan CP, Sectish TC; I-PASS Study Group: I-pass, a mnemonic to standardize verbal handoffs. Pediatrics 2012; 129:201–4

12. Young JQ, Ten Cate O, O'Sullivan PS, Irby DM: Unpacking the complexity of patient handoffs through the lens of cognitive load theory. Teach Learn Med 2016; 28:88–96
13. Starmer AJ, Spector ND, Srivastava R, West DC, Rosenbluth G, Allen AD, Noble EL, Tse LL, Dalal AK, Keohane CA, Lipsitz SR, Rothschild JM, Wien MF, Yoon CS, Zigmont KR, Wilson KM, O'Toole JK, Solan LG, Aylor M, Bismilla Z, Coffey M, Mahant S, Blankenburg RL, Destino LA, Everhart JL, Patel SJ, Bale JF Jr, Spackman JB, Stevenson AT, Calaman S, Cole FS, Balmer DF, Hepps JH, Lopreiato JO, Yu CE, Sectish TC, Landrigan CP; I-PASS Study Group: Changes in medical errors after implementation of a handoff program. N Engl J Med 2014; 371:1803–12
14. Cook DA, Beckman TJ, Bordage G: Quality of reporting of experimental studies in medical education: A systematic review. Med Educ 2007; 41:737–45
15. Zhou Y, Sun H, Lien CA, Keegan MT, Wang T, Harman AE, Warner DO: Effect of the BASIC examination on knowledge acquisition during anesthesiology residency. Anesthesiology 2018; 128:813–20
16. Steadman RH, Burden AR, Huang YM, Gaba DM, Cooper JB: Practice improvements based on participation in simulation for the maintenance of certification in anesthesiology program. Anesthesiology 2015; 122:1154–69
17. Tavakol M, Sandars J: Quantitative and qualitative methods in medical education research: AMEE Guide No 90: Part I. Med Teach 2014; 36:746–56
18. Franenkel JR, Wallen NE, Hyun HH: How to Design and Evaluate Research in Education. 9th edition. New York, McGraw-Hill Education, 2015
19. Cook DA, Beckman TJ: Reflections on experimental research in medical education. Adv Health Sci Educ Theory Pract 2010; 15:455–64
20. Marsden E, Torgerson CJ: Single group, pre-and post-test research designs: Some methodological concerns. Oxf Rev Educ 2012; 38:583–616
21. Reed DA, Cook DA, Beckman TJ, Levine RB, Kern DE, Wright SM: Association between funding and quality of published medical education research. JAMA 2007; 298:1002–9
22. Cook DA, Beckman TJ: Current concepts in validity and reliability for psychometric instruments: Theory and application. Am J Med 2006; 119:166.e7–16
23. Allen MJ, Yen WM: Introduction to Measurement Theory. Long Grove, IL, Waveland Press, 2002
24. Tavakol M, Sandars J: Quantitative and qualitative methods in medical education research: AMEE Guide No 90: Part II. Med Teach 2014; 36:838–48
25. American Educational Research Association, American Psychological Association, National Council on Measurement in Education: Standards for Educational and Psychological Testing. Washington, DC, American Educational Research Association, 2014
26. Cook DA, Lineberry M: Consequences validity evidence: Evaluating the impact of educational assessments. Acad Med 2016; 91:785–95
27. Messick S: Validity, Educational Measurement, 3rd edition. The American Council on Education/Macmillan Series on Higher Education. New York, Macmillan, 1989, pp 13–103
28. Tavakol M, Dennick R: Making sense of Cronbach's alpha. Int J Med Educ 2011; 2:53–5
29. Downing SM: Validity: On meaningful interpretation of assessment data. Med Educ 2003; 37:830–7
30. Blum RH, Muret-Wagstaff SL, Boulet JR, Cooper JB, Petrusa ER, Baker KH, Davidyuk G, Dearden JL, Feinstein DM, Jones SB, Kimball WR, Mitchell JD, Nadelberg RL, Wiser SH, Albrecht MA, Anastasi AK, Bose RR, Chang LY, Culley DJ, Fisher LJ, Grover M, Klainer SB, Kveraga R, Martel JP, McKenna SS, Minehart RD, Mitchell JD, Mountjoy JR, Pawlowski JB, Pilon RN, Shook DC, Silver DA, Warfield CA, Zaleski KL; Harvard Assessment of Anesthesia Resident Performance Research Group: Simulation-based assessment to reliably identify key resident performance attributes. Anesthesiology 2018; 128:821–31
31. Hulley SB, Cummings SR, Browner WS, Grady DG, Newman TB: Designing Clinical Research. 4th edition. Philadelphia, Lippincott Williams & Wilkins, 2011
32. Streiner DL, Norman GR, Cairney J: Health Measurement Scales: A Practical Guide to Their Development and Use. New York, Oxford University Press, 2015
33. Sullivan GM, Artino AR Jr: Analyzing and interpreting data from Likert-type scales. J Grad Med Educ 2013; 5:541–2
34. Chan AW, Altman DG: Identifying outcome reporting bias in randomised trials on PubMed: Review of publications and survey of authors. BMJ 2005; 330:753
35. Cook DA, Levinson AJ, Garside S: Method and reporting quality in health professions education research: A systematic review. Med Educ 2011; 45:227–38
36. Connor JT: The value of a p-valueless paper. Am J Gastroenterol 2004; 99:1638–40
37. Sullivan GM, Feinn R: Using effect size-or why the P value is not enough. J Grad Med Educ 2012; 4:279–82
38. Cook DA, Levinson AJ, Garside S, Dupras DM, Erwin PJ, Montori VM: Internet-based learning in the health professions: A meta-analysis. JAMA 2008; 300:1181–96
39. Cook DA, Reed DA: Appraising the quality of medical education research methods: The Medical Education Research Study Quality Instrument and the Newcastle-Ottawa Scale-Education. Acad Med 2015; 90:1067–76
40. Kirkpatrick D: Revisiting Kirkpatrick's four-level model. Training and Development 1996; 50: 54–9
41. Sawatsky AP, Beckman TJ, Edakkanambeth Varayil J, Mandrekar JN, Reed DA, Wang AT: Association between study quality and publication rates of medical

education abstracts presented at the Society of General Internal Medicine Annual Meeting. J Gen Intern Med 2015; 30:1172–7
42. Eaton JE, Reed DA, Aboff BM, Call SA, Chelminski PR, Thanarajasingam U, Post JA, Thomas KG, Dupras DM, Beckman TJ, West CP, Wittich CM, Halvorsen AJ, McDonald FS: Update in internal medicine residency education: A review of the literature in 2010 and 2011. J Grad Med Educ 2013; 5:203–10
43. Wells G, Shea B, O'Connell D, Peterson J, Welch V, Losos M, Tugwell P: The Newcastle-Ottawa Scale (NOS) for assessing the quality of nonrandomised studies in meta-analyses. Available at: http://www.ohri.ca/programs/clinical_epidemiology/oxford.asp. Accessed July 10, 2018.

ANESTHESIOLOGY REFLECTIONS FROM THE WOOD LIBRARY-MUSEUM

Chloroforming Coughs? A Tin of Gibson's Linseed, Liquorice, and Chlorodyne



During an 1848 cholera outbreak in India, British Army Surgeon J. Collis Browne, M.R.C.S. (1819 to 1884) used Chlorodyne—his formulation of laudanum, cannabis, and chloroform—as an antidiarrheal remedy. Years later, he partnered with London pharmacist J. T. Davenport to mass-market Chlorodyne as a panacea. From Manchester, England, Robert Gibson & Sons combined Chlorodyne with Linseed and Liquorice in decorative tins of cough lozenges (*above*). By 1901 these "beautifully enameled counter show tins" were advertised to American professionals and the public as filled with cough lozenges that "act like magic." (Copyright © the American Society of Anesthesiologists' Wood Library-Museum of Anesthesiology.)

George S. Bause, M.D., M.P.H., Honorary Curator and Laureate of the History of Anesthesia, Wood Library-Museum of Anesthesiology, Schaumburg, Illinois, and Clinical Associate Professor, Case Western Reserve University, Cleveland, Ohio. UJYC@aol.com.