

# Developing scholarly projects in education: A primer for medical teachers

THOMAS J. BECKMAN & DAVID A. COOK

Mayo Clinic College of Medicine, Rochester, MN, USA

## Abstract

Boyer and Glassick's broad definition of and standards for assessing scholarship apply to all aspects of education. Research on the quality of published medical education studies also reveals fundamentally important elements to address. In this article a three-step approach to developing medical education projects is proposed: refine the scholarly question, identify appropriate designs and methods, and select outcomes. Refining the scholarly question requires careful attention to literature review, conceptual framework, and statements of problem and study intent. The authors emphasize statement of study intent, which is a study's focal point, and conceptual framework, which situates a project within a theoretical context and provides a means for interpreting the results. They then review study designs and methods commonly used in education projects. They conclude with outcomes, which should be distinguished from assessment methods and instruments, and are separated into Kirkpatrick's hierarchy of reaction, learning, behavior and results.

## Introduction

In 1990 Boyer proposed a four-category framework of scholarship: Discovery, Integration, Application and Teaching (Boyer 1990). As the name implies, *Discovery* scholarship is finding new knowledge, usually through experimental research. *Integration* scholarship is synthesizing knowledge across disciplines and showing the relationships between individual parts of the whole. *Application* scholarship is harnessing knowledge in a useful and practical fashion. *Teaching* scholarship is the ability to communicate knowledge in ways that deepen learners' understanding and allow them to place information into a larger context. Scholarship, regardless of the category, should be disseminated in peer-reviewed forums and expanded upon by a wide community of scholars (Hutchings & Schulman 1999; Beattie 2000; Fincher & Work 2006).

Glassick advanced Boyer's work by defining six standards for assessing scholarship (Glassick et al. 1997; Glassick 2000). Establishing *Clear Goals* includes outlining the aims of a project and articulating statements of problem and study intent. *Adequate Preparation* requires a critical and thorough literature review. *Appropriate Methods* reflects the use of proper study design and the selection of meaningful outcomes. *Effective Communication* is apparent in a logically organized manuscript, with, for instance, an Introduction that progresses from broad concepts to specific. *Reflective Critique* uncovers threats to validity and shows how a given project increases knowledge and understanding in education. The last

## Practice points

- Glassick's standards are the yardstick for measuring education scholarship.
- Research indicates that fundamentals of education scholarship are neglected.
- Successful projects have well-crafted study questions, situated within convincing conceptual frameworks. Outcomes should strike a balance between feasibility and meaningfulness.
- Understanding and identifying appropriate methods and proceeding in a systematic way will improve the quality of scholarly projects.

standard, *Outstanding Results*, will be achieved only when the other standards have been carefully addressed. Whether developing a curriculum, writing a review article, or conducting research, Glassick's standards are the yardstick by which education scholarship is measured.

With the emergence of medical education as a field of scientific inquiry, authorities have requested the application of traditional research designs (Norman 2003; Carney et al. 2004) and expert guidelines (Moher et al. 1999; Moher et al. 2001; Des Jarlais et al. 2004). Authorities have also identified the need to address essential elements of education scholarship and research. For example, Bordage's analysis of manuscripts submitted to the Association of American Medical Colleges'

Research in Medical Education Proceedings revealed that manuscripts were rejected due to inappropriate instrumentation, insufficient problem statement, and incomplete, inaccurate or outdated literature review (Bordage 2001). Conversely, strengths of accepted manuscripts were importance of the problem studied and sound study design. Chubin and Hackett (1990) showed that manuscripts were rejected from *Social Studies of Science* owing to poor argumentation and ignorance of the literature. Our systematic review of articles published in six major medical education and general medical/surgical journals demonstrated that essential elements of scientific reporting were often missing (Cook et al. in press). Specifically, only 55% presented a conceptual framework, 45% critically reviewed the literature and 16% presented a statement of study design. While most presented a statement of study intent (purpose, research question or hypothesis), the majority of these statements were incomplete. All this underscores the need for greater attention to the scholarly assessment of education studies, such as formulating statements of problem and study intent (clear goals); consulting expert guidelines, carefully reviewing the literature and reflecting on conceptual framework (adequate preparation); and utilizing sound instruments and study designs (appropriate methods).

We propose a three-step approach to designing scholarly education projects (Table 1). Step 1 is *refining the study question*. Arguably, every scholarly endeavor begins with thoughtful reflection that eventually leads to a scholarly question. Consider a clinician who observes that his/her teaching experience is enhanced by discussing teaching

situations and learning principles with his/her experienced colleagues. Subsequently, that teacher wonders: ‘Would attending physicians enjoy increased satisfaction if they staffed residents in a group setting, versus independently?’ Refining this question would require a literature review to determine whether anyone else has investigated this or similar questions. Therefore, gaps in the literature could be identified, leading to a problem statement. Likewise, finding convincing reasons, like existing theories or previous approaches, to suggest that teaching in a group setting would be satisfying would provide a conceptual framework, and ultimately a more refined study question. Step 2 is *identifying a research study design*. In the current example, a randomized controlled design could be implemented by assigning half of the faculty to teach residents in the group setting, and the other half to teach independently. Step 3 is *selecting outcomes*. In this case, the scholarly question identifies teacher satisfaction as the desired outcome. Measuring this outcome could be accomplished by surveying faculty (method) with a Likert-scaled questionnaire (instrument). To follow is a more detailed discussion and practical examples of this three-step process.

### Step 1: Refine the study question

Refining a study question requires attention to a literature review, problem statement, conceptual framework and statement of study intent, all of which are typically found in the introduction portion of a manuscript or study proposal (McGaghie et al. 2001). The topic of interest is developed by progressing from general to specific concepts and from the

**Table 1.** Steps for developing scholarly medical education projects.

Steps	Components and examples	Comments
1. Refine the Study Question	Literature Review	Identifying existing studies and understanding the relevant scholarly environment
	Problem Statement	Describes the overarching context of the study and conveys how it will advance the literature
	Conceptual Framework	A theory, model or approach that situates the study question within a theoretical context and explains the results
	Statement of Study Intent	May be stated as a question, hypothesis or goal
2. Identify Designs and Methods <sup>a</sup>	Experimental	Manipulating an independent variable and studying its effect on a dependent variable
	Observational	Does not involve altering the events under study Includes various methods for determining relationships between variables that are not manipulated
	Validity	Collecting evidence to support valid interpretations of instrument scores
	Qualitative	Data are words Has features of both design and method
	Systematic Reviews	Utilizing explicit methods to identify and summarize previously published studies on a specific subject
3. Select Outcomes	Outcome	Outcomes are conceptual Examples are attitudes, skills and knowledge
	Outcome Methods	Outcome methods (e.g. surveys) are general approaches to assessing a given outcome
	Instruments	Instruments (e.g. questionnaires) are specific devices for systematically collecting data

Notes: <sup>a</sup>Sometimes the distinction between design and method is uncertain. See text for more detailed discussion. Although this table illustrates study designs and methods commonly used in medical education studies, the list is not intended to be exhaustive.

known to the unknown (McGaghie et al. 2001). In this way a well-crafted introduction sets the stage for a credible study question.

#### Literature review

Incomplete literature review is a primary reason for manuscript rejection (Bordage 2001) and a critical literature review was found in only 45% of published education studies (Cook et al. in press). The objectives of a literature review are to integrate knowledge, establish the conceptual framework and scholarly question, clarify the study design and methods, and justify interpretations of study findings (Crandall et al. 2001).

Reviewing the medical education literature is challenging due to a lack of comprehensive databases for education, abundant references outside medical education (e.g. psychology, sociology and general education), and disagreement between medical education themes and medical subject headings and key words (Reed et al. 2005). Nevertheless, solutions to these challenges exist. Numerous databases including MEDLINE, PubMed, PsychINFO, Educational Resource Information Center (ERIC), British Educational Index (BEI), and the Cumulative Index to Nursing and Allied Health Literature (CINAHL) should be searched (Haig & Dozier 2003; Reed et al. 2005). Also, surrogate search terms should be used. For example, when seeking reliable instruments in a given field of study, in addition to the search terms 'reliable' and 'instrument' one should also try 'psychometric', 'validity', and 'evaluation studies'. Finally, consulting experts and reviewing the bibliographies of selected articles will yield valuable references that would otherwise be missed.

#### Problem statement

The Problem Statement is an essential component of any well-written introduction and its omission is a frequent reason for manuscript rejection (Bordage, 2001). The Problem Statement describes the overarching context of the study and conveys how it will advance the literature. In this way, the Problem Statement also helps readers foresee the Statement of Study Intent (McGaghie et al. 2001). Consider the following example:

There is a growing body of research on journal peer review. For example, JAMA has dedicated three complete issues in the past decade . . . to peer review studies and essays, and the Council of Biology Editors . . . also published a book of papers in 1991 from the First International Congress on Peer Review in Biomedical Publishing. However, few studies exist that analyze the content of reviewers' comments when reviewers are recommending rejection or acceptance of a manuscript. (Bordage 2001)

The first two sentences in this excerpt describe the study's context, whereas the last sentence conveys how the study will advance the literature and announces the Statement of Study Intent (see below). Notice that the last sentence becomes a

pivotal point in the introduction as it moves from what is known to what is yet to be (i.e. the current study). Therefore, problem statements often contain transitions of contrast like 'however', 'nevertheless' or 'conversely'.

#### Conceptual framework

We cannot overemphasize the importance of Conceptual Framework. It is a theory, an approach or a model for how things work that situates a research question within the appropriate theoretical context. It guides the selection of study variables (McGaghie et al. 2001) and ultimately provides a means to interpret the study results by allowing for a 'why' or 'because'. Finally, the Conceptual Framework is like glue that unites a body of thought by refining existing theories, developing new theories and providing a basis for further scholarship (Prideaux, 2002; Prideaux & Bligh, 2002; Des Jarlais et al. 2004). Consider this example of a Conceptual Framework stated as a theory:

Gagne *et al.* . . . proposed an information-processing model of learning. This model assumes answering a low-level recall question requires location of information in 'long-term' memory, retrieval into 'working' memory, verification that the information retrieved answers the question, and, finally, answering the question. It is implied that some minimum time is required to complete this process effectively. More complex questions require application of known information to unknown situations and assessment of whether the application of the old information to the new situation is correct . . . Given this model, it is hypothesized that more time is required to answer more complex questions. (Schneider et al. 2004)

Another example illustrates a Conceptual Framework stated as a model or approach:

General medicine services are often comprised of medically complex patients with challenging psychosocial issues, thus allowing general internists the opportunity to model intensive physician-patient and teacher-learner dialogues and the application of a broad knowledge base. Cardiology services, on the other hand, are generally comprised of patients with focused problems (e.g., acute coronary syndromes and arrhythmias), thus allowing cardiologists the opportunity to model the application of a narrow and deep knowledge base. Since general internists and cardiologists model different behaviors and teach different skills, we would also expect assessments of these faculty members to reflect different latent variables (constructs). (Beckman et al. 2006)

Stating a Conceptual Framework is particularly important in the social sciences, psychology and medical education, because research questions in these disciplines often hinge on one of several complementary or contrasting theories. Conversely, in the biomedical sciences underlying theories or

**Table 2.** Quantitative and qualitative approaches to research.

Characteristic	Quantitative	Qualitative
Data	Numbers Answers questions like: 'what percentage of patients are willing to receive a tetanus vaccination?'	Words (e.g. field notes, interviews, focus groups, video tapes) that reflect ordinary events and are grounded in natural settings Answers questions like: 'why do some patients decline disease prevention?'
Basic Tenet	The universe is one reality comprising discoverable facts	The universe contains many, socially constructed truths
Reasoning Process	Hypothetical-deductive	Inductive
Function of the Investigator	Uninvolved observer	Empathetic participant; integrates findings within the study context
Goal of the Study	Identify associations and causal relationships between variables	Understand circumstances from the perspectives of the study subjects
Typical Study Design	Experiment	Grounded Theory: Theories are not identified a priori, but are discovered as the study unfolds
Generalizing Study Findings	Ideally, quantitative research findings generalize to other settings	Generalizability may not be possible, and the user must determine whether results apply to his/her particular setting
Limitations	Numerical data can be perceived as less 'human', and may therefore be less engaging and persuasive. Also, quantitative methods may fail to reveal unanticipated findings	Subjective analysis increases the potential for biased data interpretation

Notes: Increasingly, qualitative and quantitative methods are used in the same study. (For more information see: Miles & Huberman 1994; Greenhalgh & Taylor 1997; Fraenkel & Wallen 2003; Kennedy & Lingard 2006.)

mechanisms may be sufficiently well established as to allow marked abbreviation or even omission. All this may explain our prior finding that Conceptual Frameworks are reported more frequently in medical education journals than in non-education journals (Cook et al. in press).

### Statement of study intent

The Statement of Study Intent, which is usually placed last in the Introduction (McGaghie et al. 2001) or first in the Methods, may be worded as a scholarly question, hypothesis, or statement of goal, purpose or aim. Consider these examples.

- The scholarly question: 'How do PBL [Problem Based Learning] and non-PBL students compare in terms of performance in an anatomy test consisting of non-contextual fact-oriented items and clinically contextualised items?' (Prince et al. 2003)
- The hypothesis: 'This study investigates the hypothesis that the feedback provided by SAIL can improve the quality of hospital doctors' written communication.' (Fox et al. 2004)
- The goal statement: 'the goal of this study was to better understand the nature of strengths and weaknesses in medical education reports by analyzing the ratings and comments made by external reviewers'. (Bordage 2001)

The focal point of any study is the Statement of Study Intent. Morrison (2002) asserts, 'It is more important to understand the question than to find the answer'. Others rejoin, 'The single most important component of a study is the research [scholarly] question' (Marks et al. 1988; Bordage & Dawson 2003). The Statement of Study Intent consolidates the literature review, conceptual framework and problem statement, and clarifies the study variables.

Bordage and Dawson (2003) identified important steps in formulating a scholarly question. First consider the topic

of study. In this case, the FINER mnemonic proves useful (Hulley & Cummings 1988). The topic should be Feasible, Interesting, Novel, Ethical and Relevant. Second, review the literature to confirm that your scholarly question has not been answered previously. Third, identify both the intervention (independent variable) and the outcome (dependent variable). Note that operationally defining study variables within the scholarly question facilitates study design. Operational definitions describe not only the variable (e.g. 'active-reflective learning style'), but also the way that it will be identified (e.g. 'assessed using Kolb's Learning Style Inventory'). Fourth, determine whether you seek a difference or an association. For example, 'Is cognitive structuring of knowledge better among learners in a Problem Based Learning (PBL) program than those in a traditional program?' looks for a difference. Conversely, 'Is there a correlation between attendance at PBL sessions and test scores?' looks for an association. Fifth, identify the target population to which you will generalize the results. Researchers utilize a sample, which belongs to a sample population, which in turn belongs to a target population. The generalizability of study findings will be determined by the degree of similarity between the sample and the population, and by the study methods (e.g. findings from multi-institutional studies are more generalizable than from single institution studies). Finally, always hypothesize what the results will be.

### Step 2: Identify study designs and methods

A major reason for manuscript acceptance is sound study design (Bordage 2001) and explicit design statements are under-reported in published education studies (Cook et al. in press). Hence, contemplating study design early in the development of any education study, and explicitly stating the design in the corresponding manuscript, is crucial. We discuss

**Table 3.** Experimental study designs.

Study Design	Group 1	Group 2	Comment
True Experimental			Random assignment
Post-test only	R X1 O	R X2 O	
Pre-test–Post-test	R O X1 O	R O X2 O	
Solomon 4-Group	R O X1 O R X1 O	R O X2 O R X2 O	Controls for the effect of a pre-test by giving one half of the groups the pre-test and post-test and the other half the post-test only
Quasi-Experimental			Non-random assignment (e.g. self-selection or instructor assignment)
Post-test only	NR X1 O	NR X2 O	
Pre-test–Post-test	NRO X1 O	NRO X2 O	
Static Group Comparison			Uses groups that exist at the start of the study (e.g. different academic years, different classes)
Post-test Only	X1 O	X2 O	
Pre-test–Post-test	O X1 O	O X2 O	
Single Group			No comparison group
One Shot Case Study	X O		
One Group Pre-test–Post-test	O X O		

Notes: R = random assignment, NR = non-random assignment, O = Observation (e.g. pre- or post-test), X = exposure of a group (X1 is 'group 1' and X2 is 'group 2') to an intervention or alternate intervention. (For more information see: Campbell & Stanley, 1966; Fraenkel & Wallen, 2003.)

below the study designs and methods that seem to occur commonly in the medical education literature, although we realize this is not an exhaustive list. Additionally, we realize that not all education scholarship is research, and research can be both qualitative and quantitative (Table 2). Nevertheless, one should identify and understand the category of their scholarly approach and proceed in a thoughtful and systematic way.

### Experimental and observational studies

When designing clinical research the fundamental question is whether or not to alter the events under study. If the answer is yes, then the design is *experimental*, and if the answer is no, then the design is *observational* (Hulley & Cummings, 1988). For observational studies it is further determined whether to take measurements on one occasion (e.g. a cross-sectional study) or on several occasions (e.g. a longitudinal study such as a cohort study). Recall that not all studies are created equal. At the top of the evidence pyramid are randomized controlled trials, followed by cohort studies, case control studies, case series, and at the bottom of the pyramid is expert opinion (Oxford Centre for Evidence Based Medicine, 2006). Experts have called for the application of these traditional study designs to education scholarship (Carney et al. 2004), but we have observed that even randomized controlled trials do not account for all sources of study variance, such as learning outside the curriculum and variation in teaching quality and style (Beckman & Cook 2004).

Experimental studies manipulate an independent variable and assess its effect on a dependent variable (Fraenkel & Wallen 2003). Experiments are subdivided into single-group,

static group, quasi-experimental and true experimental designs (Table 3) (Campbell & Stanley 1966; Fraenkel & Wallen 2003). The weakest study designs use one group of subjects (no comparison or control). Static group designs use unassigned groups of participants that existed before the study began (e.g. different classes or grades). In quasi-experimental designs, subjects are assigned to groups by the investigator; however, the assignment is non-random. In the most rigorous design—true experimental—participants are randomly assigned to groups. Note that while all experimental studies might demonstrate causal relationships, randomized study designs provide the strongest causal argument. Yet even randomized education studies can be hopelessly flawed (Norman 2003; Beckman & Cook 2004). Certainly, a carefully designed quasi-experimental study could be more rigorous than a poorly designed randomized control trial. Therefore, randomization is not always feasible or desirable in educational research.

Observational studies determine relationships between variables that are not manipulated. While such studies cannot establish causality, they suggest causal relationships and generate interesting hypotheses. Two categories of observational studies commonly used in education are *correlational* and *causal-comparative* (Fraenkel & Wallen 2003). Correlational studies determine usually occurring relationships between variables; correlations between these variables can be positive (convergent) or negative (divergent). Causal-comparative studies explore relationships between groups that already exist separately in nature. Causal-comparative studies can thus be useful in identifying differences between groups such as physicians and surgeons, doctors and nurses, medical students and dental students, etc.

**Table 4.** Validity: Sources of evidence, definitions, and examples.

Evidence source	Definition	Examples
Content	The relationship between a test's content and the construct it is intended to measure. Refers to themes, wording, and format of items on an assessment instrument. Includes analyses by experts regarding how adequately items represent the content domain. Also includes development strategies to ensure appropriate content representation	Surveying experienced teachers regarding the adequacy and representativeness of proposed instrument items. Choosing items previously utilized in similar settings. Developing instruments based on existing literature and established educational theories
Response process	Analyses of responses, including the actions, strategies and thought processes of individual respondents or observers. Differences in response processes may reveal sources of variance that are irrelevant to the construct being measured. Also includes instrument security, scoring and reporting of results	Interviewing and studying learners regarding factors that influence the ratings they assign to teachers. Analyzing varying response patterns among different categories/levels of learners
Internal structure	The degree to which individual items within the instrument fit the underlying constructs. Items measuring a unidimensional construct should be homogeneous, while items measuring complex constructs should not. Most often reported as measures of internal consistency reliability and factor analysis	Using factor analysis to determine the dimensional structure of an instrument's scores, and determining the reliability of scores. Studying differential functioning of items among a homogeneous group of evaluators
Relations to other variables	The relationship between scores and other variables relevant to the construct being measured. Relationships may be positive (convergent or predictive) or negative (divergent or discriminant) depending on the constructs being measured	How well do teachers' assessment scores predict learners' performance on high-stakes examinations, or their choice of a medical specialty? Do scores correlate with other measures of the same construct? Can results of an evaluation be generalized from one setting to another, similar setting?
Consequences	Assessments are intended to have some desired effect (e.g. improve teaching and learning performance), but they also have unintended effects. Evaluating such consequences can support or challenge the validity of score interpretations	Does an assessment of teaching accompanied by feedback improve overall course evaluations (or students' scores on high-stakes examinations)? Do equally qualified teachers' performances on clinical teaching assessments correlate with factors that are not being measured, such as gender or ethnicity?

Source: Beckman et al. (2005). Adapted with permission.

## Validity studies

Scores from education instruments such as teaching assessments, survey questionnaires and knowledge tests require valid interpretations to be meaningful (Downing 2003; Beckman et al. 2005; Cook & Beckman 2006). Scores are valid to the extent that they realistically portray the phenomenon of interest. The purpose of a validity study is to collect evidence supporting the interpretations of an instrument's scores.

Validity is defined as the 'degree to which evidence and theory support the interpretations of test [and assessment] scores' (American Education Research Association et al. 1999). This definition underscores several important points. First, validity is not present or absent, but occurs by degree. Second, valid interpretations are always grounded in evidence and theory. Third, validity refers not to tests and instruments, but to the interpretation of test and instrument *scores* (Messick 1993; Downing 2003; Cook & Beckman 2006). Furthermore, validity is considered a hypothesis (Messick 1993; Downing 2003; Cook & Beckman 2006) and the decision to accept (or reject) this hypothesis will depend on the weight of evidence for (or against) the hypothesis. The validity hypothesis is continually enhanced or contravened by new evidence, resulting in a never-ending cycle of hypothesis revision and reformulation (Messick 1993).

When applied to the interpretation of a study, validity is either internal (i.e. can the study results be trusted?), or

external (i.e. do the results generalize to other settings?). But validity has a different meaning when applied to the interpretation of an instrument's scores. In this case, all validity is construct validity, and evidence is collected to support the intended construct from five sources: content, response process, internal structure, relations to other variables and consequences (Table 4) (Messick 1993; American Education Research Association et al. 1999; Downing 2003; Beckman et al. 2005; Cook & Beckman 2006). Realize that not every study will require all sources of validity evidence, and it is likely that some sources of evidence will be more important than others for a given study purpose. Nonetheless, as for any hypothesis, convincing evidence from a variety of sources will strengthen the validity argument.

## Qualitative research

The term 'qualitative research' relates to features of both study design and method. Qualitative research has matured substantially over recent decades and, increasingly, investigators are combining qualitative and quantitative methods (Rossman & Wilson 1985; Miles & Huberman 1994). A fundamental distinction between quantitative and qualitative research is that data for quantitative research are numbers and data for qualitative research are words (Miles & Huberman 1994; Greenhalgh & Taylor 1997; Fraenkel & Wallen 2003). Other striking differences exist (see Table 2). For example,

in Grounded Theory, new theories are identified as data are collected, whereas in the hypothetical-deductive model common to quantitative research, data are collected to verify or challenge an existing theory (Glaser & Strauss 1995; Kennedy & Lingard 2006). Although all qualitative research methods share similar features, there are dozens of qualitative methods (Miles & Huberman 1994).

### Surveys

Surveys ask questions to better understand subject characteristics like behavior, attitude and knowledge. Surveys are administered to a selected sample, responses are gathered and interpreted, and then inferences are made regarding characteristics of the intended population. Surveys can be either cross-sectional (administered at one point in time), or longitudinal (administered at different points in time) (Fraenkel & Wallen 2003). For surveys to be meaningful, the questions should be carefully constructed, the sample should represent the intended population, and the response rate should be adequate—generally >60% for mail and Internet surveys. Note that surveys overlap with other study designs and methods (Fraenkel & Wallen 2003). For example, determining the internal consistency and interrater reliability of survey instrument scores might constitute a validity study, while correlating survey data with other variables of interest would be association research.

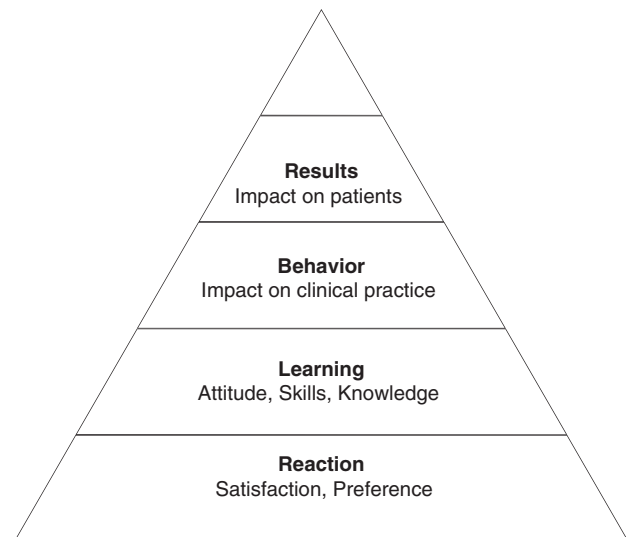
### Systematic reviews and meta-analyses

Systematic reviews utilize explicit methods to identify previously published studies to answer a specific scholarly question. A meta-analysis is a subset of systematic reviews that statistically combines results from various studies to provide an overall estimate of the effect (Moher et al. 1999). Systematic reviews on medical education topics are challenging because there are few standardized interventions and assessments and research reporting is often poor. Nonetheless, the quality of systematic reviews on medical education topics can be optimized by identifying a focused scholarly question; being systematic in all reporting elements including literature search, article inclusion and reviewing methods; and clearly answering the original question in the discussion and applying the same framework to each article discussed.

## Step 3: Select outcomes

The study intervention is the independent (predictor) variable, whereas the study *outcome* is the dependent variable. It is important to distinguish outcomes, which are conceptual, from methods (general approaches to assessing a given outcome) and instruments (specific devices for systematically collecting data). For example, to demonstrate learner knowledge (outcome) one could use a multiple-choice test (method), which could be measured using the USMLE Step II (instrument). Notably, there are usually several potential methods and instruments for measuring a given outcome.

Kirkpatrick (1996) observed that outcomes can be separated into four levels (see Figure 1). At the lowest level



**Figure 1.** Kirkpatrick's Outcomes Hierarchy.

*Note:* Outcomes become progressively more meaningful, yet more challenging to demonstrate, when progressing from the bottom to the top of the pyramid. (For further details see: Kirkpatrick 1996; Hutchinson 1999.)

is reaction, followed by learning, then behavior, and results at the top (Kirkpatrick 1996; Hutchinson 1999). Using a hypothetical curriculum for teaching the management of patients with diabetes, we can develop examples for each of Kirkpatrick's levels. Surveying students as to how enjoyable the curriculum is would be a *Reaction* outcome. Using a multiple-choice test to measure student knowledge of diabetes management would be a *Learning* outcome. Reviewing charts to determine whether students actually ordered tests consistent with diabetes management guidelines would be a *Behavior* outcome. Finally, determining whether patients managed by students completing the curriculum experienced improvements in glycemic control (compared with baseline values or patients managed by a control/comparison group) would be a hard *Result*.

Notice that outcomes become more meaningful when progressing from reaction to results (Kirkpatrick 1996). Unfortunately, outcomes also become more difficult to measure when progressing from reaction to results, because the higher the level, the more abundant the confounders that could be contributing to the outcome, and the smaller the effect size (and hence the larger the sample size needed to show a significant effect). It was thus expressed by Shea (2001) that the outcome level which often strikes an appropriate balance between feasibility and meaningfulness is behavior.

So consider the following steps when selecting outcomes for educational projects. First, choose the desired outcome, balancing feasibility with meaningfulness. Second, choose a method for measuring the outcome. Third, choose an instrument appropriate for the chosen method. Also consider whether the instrument has prior evidence of score reliability and validity. If not, then consider conducting a validity study prior to your initially planned study. Last, for

every test or assessment, remember to sample the content domain adequately, since score reliability is proportional to the number of observations and instrument items (DeVillis 1991).

## Conclusions

Boyer's definition of scholarship embraces all dimensions of education. Glassick's standards for assessing scholarship are widely accepted, but research on the quality of medical education studies indicates that published articles often neglect these standards. Mindful of Glassick's criteria and demonstrated shortcomings in the literature, we proposed a three-step approach to developing scholarly education projects. Refining the statement of study intent (Step 1) requires creating a conceptual framework, reviewing the literature and finding existing gaps in understanding. Understanding and identifying appropriate study designs and methods (Step 2) will substantially improve the quality of scholarly projects. Selecting outcomes (Step 3)—whether at the level of reaction, learning, behavior or results—completes the cycle of scholarship by formalizing the dependent variables as defined in the statement of study intent. We anticipate that applying this systematic approach will improve the understanding and communication of scholarly medical education projects.

## Notes on contributors

THOMAS BECKMAN, MD, is Assistant Professor in the Mayo Clinic College of Medicine, Chairman of the Scholarship in Medical Education Group for the Division of General Internal Medicine, and Consultant in the Department of Internal Medicine, Mayo Clinic, Rochester, Minnesota.

DAVID COOK, MD MHP, is Assistant Professor, Chairman of the Medical Education Research Group at the Mayo Clinic College of Medicine, and Senior Associate Consultant in the Department of Internal Medicine, Mayo Clinic, Rochester, Minnesota.

## References

- American Education Research Association, American Psychological Association, & Council on Measurement in Education. 1999. Standards for educational and psychological testing. Washington, DC: American Education Research Association.
- Beattie DS. 2000. Expanding the view of scholarship: Introduction. *Acad Med* 75:871–876.
- Beckman TJ, Cook DA. 2004. Educational epidemiology. *JAMA* 292:2969.
- Beckman TJ, Cook DA, Mandrekar JN. 2005. What is the validity evidence for assessments of clinical teaching? *J Gen Intern Med* 20:1159–1164.
- Beckman TJ, Cook DA, Mandrekar JN. 2006. Factor instability of clinical teaching assessment scores among general internists and cardiologists. *Med Educ* 40:1209–1216.
- Bordage G. 2001. Reasons reviewers reject and accept manuscripts: the strengths and weaknesses in medical education reports. *Acad Med* 76:889–896.
- Bordage G, Dawson B. 2003. Experimental study design and grant writing in eight steps and 28 questions. *Med Educ* 37:376–385.
- Boyer EL. 1990. *Scholarship reconsidered: priorities of the professoria* (New York, Wiley).
- Campbell DT, Stanley JC. 1966. *Experimental and quasi-experimental designs for research* (Chicago, Rand McNally).
- Carney PA, Neirenborg DW, Pipas CF, Brooks WB, Stukel TA, Keller AM. 2004. Educational epidemiology: applying population-based design and analytic approaches to study medical education. *JAMA* 292:1044–1050.
- Chubin DE, Hackett EJ. 1990. Peer review and the printed word, in: DE Chubin & EJ Hackett (Eds), *Peerless science: peer review and U.S. science policy*, pp. 83–122 (Albany, State University of New York Press).
- Chubin DE, Hackett EJ. 1990. Peer review and the printed word, *Peerless science: peer review and U.S. science policy* & pp. 83–122 (Albany, State University of New York Press).
- Cook DA, Beckman TJ. 2006. Current concepts in validity and reliability for psychometric instruments: theory and application. *Am J Med* 119:166e7–166e16.
- Cook DA, Beckman TJ, Bordage G. Quality of reporting experimental studies in medical education: A systematic review. *Med Educ*. In press.
- Crandall SJ, Steinecke A. 2001. Reference to the literature and documentation. *Acad Med* 76:925–927.
- Des Jarlais DC, Lyles C, Crepaz N. 2004. Improving the reporting quality of nonrandomized evaluations of behavioral and public health interventions: the TREND statement. *Am J Public Health* 94:361–366.
- DeVillis RF. 1991. *Scale development: theory and applications* (London, Sage Publications).
- Downing SM. 2003. Validity: on the meaningful interpretation of assessment data. *Med Educ* 37:830–837.
- Fincher RE, Work JA. 2006. Perspectives on the scholarship of teaching. *Med Educ* 40:293–295.
- Fox AT, Palmer RD, Crossley JGM, Sekaran D, Trewavas ES, Davies HA. 2004. Improving the quality of outpatient clinic letters using the Sheffield Assessment Instrument for Letters SAIL. *Med Educ* 38:852–858.
- Fraenkel JR, Wallen NE. 2003. *How to design and evaluate research in education* (New York, McGraw-Hill).
- Glaser BG, Strauss AL. 1995. *The discovery of grounded theory: strategies for qualitative research* (New York, Aldine De Gruyter).
- Glassick CE. 2000. Boyer's expanded definitions of scholarship, the standards of assessing scholarship, and the elusiveness of the scholarship of teaching. *Acad Med* 75:877–880.
- Glassick CE, Huber MT, Maeroff GI. 1997. *Scholarship assessed: evaluation of the professoriate* (San Francisco, Jossey-Bass).
- Greenhalgh T, Taylor R. 1997. How to read a paper: papers that go beyond numbers qualitative research. *BMJ* 315:740–743.
- Haig A, Dozier M. 2003. BEME Guide No 3: Systematic searching for evidence in medical education—part 1: Sources of information. *Med Teach* 25:352–363.
- Hulley SB, Cummings SR. 1988. *Designing Clinical Research*. Baltimore (MD, Williams & Wilkins).
- Hutchings P, Schulman L. 1999. The scholarship of teaching. *Change* 31:11–15.
- Hutchinson L. 1999. Evaluating and researching the effectiveness of educational interventions. *BMJ* 318:1267–1269.
- Kennedy TJ, Lingard LA. 2006. Making sense of grounded theory in medical education. *Med Educ* 40:101–108.
- Kirkpatrick D. 1996. Revisiting Kirkpatrick's four-level model. *Training and Development* 50:54–59.
- Marks RG, Dawson-Saunders EK, Bailar JC, Dann BB, Verran JA. 1988. Interactions between statisticians and biomedical journal editors. *Stat Med* 7:1003–1011.
- McGaghie WC, Bordage G, Shea JA. 2001. Manuscript introduction: Problem statement, conceptual framework, and research question. *Acad Med* 76:923–924.
- Messick S. 1993. Validity, in: RL Linn (Ed.) *Educational measurement*, 3rd edn, (Phoenix, AZ, Oryx Press).
- Miles MB, Huberman AM. 1994. *Qualitative data analysis*. Thousand Oaks (CA, Sage Publications).
- Moher D, Cook DJ, Eastwood S, Olkin I, Rennie D, Stroup DF, et al. 1999. Improving the quality of reports of meta-analyses of randomized controlled trials: The QUOROM statement. *Lancet* 354:1896–1900.
- Moher D, Schulz KF, Altman D, the CONSORT Group. 2001. The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomized trials. *JAMA* 285:1987–1991.
- Developing research questions in medical education: the science and the art. *Med Educ* 36:596–597.

- Norman, G. 2003. RCT = results confounded and trivial: the perils of grand educational experiments. *Med Edu* 37:582–584.
- Oxford Centre for Evidence Based Medicine. Available online at: [http://www.cebm.net/levels\\_of\\_evidence.asp](http://www.cebm.net/levels_of_evidence.asp) (accessed March 2006).
- Prideaux D. 2002. Researching the outcomes of educational interventions: A matter of design. *British Medical Journal* 324:126–127.
- Prideaux D, Bligh J. 2002. Research in medical education: asking the right questions. *Med Educ* 36:1114–1115.
- Prince KJAH, van Mameren H, Hylkema N, Drukker J, Scherpbier AJJA, van der Vleuten CPM. 2003. Does problem-based learning lead to deficiencies in basic science knowledge? An empirical case on anatomy. *Med Educ* 37:15–21.
- Reed D, Price EG, Windish DM, Wright, Gozu A, Hsu EB, *et al.* 2005. Challenges in systematic reviews of educational intervention studies. *Ann Intern Med* 142:1080–1089.
- Rossmann GB, Wilson BL. 1985. Numbers and words: combining quantitative and qualitative methods in a single large-scale evaluation study. *Evaluation Revi* 9:627–643.
- Schneider JR, Sherman HB, Prystowsky JB, Schindler N, Darosa DA. 2004. Questioning skills: the effect of wait time on accuracy of medical student responses to oral and written questions. *Acad Med* 79:s28–s30.
- Shea JA. 2001. Mind the gap: some reasons why medical education research is different from health services research. *Med Educ* 35:319–320.

Copyright of *Medical Teacher* is the property of Taylor & Francis Ltd and its content may not be copied or emailed to multiple sites or posted to a listserv without the copyright holder's express written permission. However, users may print, download, or email articles for individual use.