Advances in Health Sciences Education

Standards for an Acceptable Manuscript

Advances in Health Sciences Education, like many journals in the field, receives far more manuscripts than we can publish. Number of submissions has doubled since 2010, and we now publish only about 13% of the article submitted. Further, we reject about 60% of manuscripts after an initial screen by an editor, and only 40% are sent out for formal peer review. We try to maintain a review time of 60-80 days to decision, however this has become increasingly difficult. As a result we are implementing a number of initiatives to improve the efficiency of the review process. This detailed specification of requirements for a successful submission is one step in this direction, so authors can see clearly the standards needed for acceptance and use these to guide their decision to submit to AHSE.

It is important to emphasize that the acceptance rate is driven by quality standards and there is no quota. Our perspective is that if an article represents a useful contribution to the field, and has no serious methodological flaws, it should be published. Unfortunately, the evidence cited above indicates that about 87% of submitted manuscripts do not meet these two criteria. Although we have not formally analyzed it, our impression is that about half fail because they do not represent a real contribution and half because of methodological flaws. Furthermore, we have studied the fate of rejected manuscripts, and only about 1/3 are eventually published.

These statistics reveal a disturbing state of affairs. The low acceptance rate of *Advances*, and overall, represents a serious waste of resources. There are likely no easy fixes. However as one attempt to make the quality judgments more explicit, we have created this document to explicitly identify the criteria we use in deciding to accept or reject. We must emphasize that these criteria are not universal, but instead, reflect the particular perspective of *Advances*. As stated in the Aims and Scope on the journal webpage:

> *From the perspective of external validity, it is critical that authors place their study in a theoretical and empirical context. AHSE has no page limit, in order that each paper can be accompanied by a critical review of related research, and the discussion can highlight how the study findings add to knowledge. Authors are encouraged to explore their study from multiple analytical perspectives, to include multiple converging studies if possible, and to specifically state how the study findings add to knowledge in the field.*

> *The editors will not consider studies where the only outcome is a person's opinion or perception of the extent to which they believe they have learned something or improved their skills. The reason is simply that the evidence is consistent that people are not capable of accurate*

self-assessment, so any form of self-assessed improvement cannot be used as an outcome. Self-assessed measures of confidence or competence may well appear to show large differences in response to an educational intervention, but are themselves weak surrogates for actual achievement.

From the perspective of educational importance, studies of a single course or program with weak evidence of effectiveness, such as student ratings, are discouraged as they are unlikely to add to generalizable knowledge, unless the study permits empirical test of theoretical predictions. Further, evaluations of any technology, without consideration of the mechanisms that lead to an observed change, are of limited value. Similarly, proving that some education is better than no education, an educational "placebo-controlled trial;" has very limited value. We will not consider such studies for publication.

We now address specific areas of concern:

1) BACKGROUND AND RELATED LITERATURE

It is critically important that the article provides sufficient background that the reviewer can judge whether the article represents real "value added." There are a number of ways to describe this section – "conceptual framework", "theoretical foundation", but none do justice to the range of possibilities. Not every study question tests a theory, and not every study relies on conceptual frameworks. But unless the author can make it clear that the particular study question has not been previously addressed and is important for the field, we will not consider it for publication.

1a) Specialty and Institutional context
Frequently the study rationale consists of solution to a local problem – either a particular health profession or specialty ("Training of pediatric gerontologists is an important ..."), a country (The Ameraustralasian Task Force on Health Manpower decreed that…."), or a school ("At Slippery Slope U we had a problem with…").

Advances is an international journal, and is read by researchers and educators in a broad range of health disciplines in many countries. It is incumbent on the author to explicitly demonstrate that the study findings are generalizable to other disciplines, educational contexts and countries. If this is not possible, the article does not belong in an international multidisciplinary journal.

1b) Theory development and testing
Far too often, educational interventions are theory-based, not theory-testing or theory-generating. To paraphrase Winifred Castle, a British statistician, "Most educators use theory the way a drunkard uses a lamppost. More for support

than illumination" (Norman 2004). A study that is designed to test the predictions of a theory potentially represents a real contribution to advancing the knowledge of a field, as does one that robustly extends or develops new theoretical perspectives. Conversely, a study that elaborates a theory simply to provide a veneer of scientific respectability adds little.

1c) Invention vs. clarification

We are not particularly interested in studies that demonstrate that some educational intervention or invention "works", whether it is a simulation, a curriculum, an assessment method. This amounts to little more than market research. Instead we want to identify the underlying variables that may contribute to success or failure, and to systematically explore these factors individually and in combination.

This distinction has been described in a number of ways. Cook, Bordage and Schmidt (2008a) distinguish between "justification" and clarification" research, where "justification" shows that a particular innovation "works", and "clarification" attempts to understand why it works. Cook (2005) points out that studies comparing one medium (hi-fi simulation) to another (video) confound a number of factors and are therefore essentially uninterpretable. Prideaux and Bligh (2002) also discuss the critical role of literature in advancing a discipline.

1d) There are clear differences between efficacy and effectiveness studies: "efficacy refers to the beneficial effects of a program or policy under optimal conditions of delivery, whereas effectiveness refers to effects of a program or policy under more real-world conditions." We are interested in both kinds of studies but the type of study should be clearly articulated and its methods, contributions ad implications selected and critiqued accordingly, not least because "efficacy trials significantly overestimate how strong an effect will be when the treatment is used under more usual conditions." (Streiner and Norman, 2009).

1e) Some dead issues

There are some areas in medical education that persist, despite substantial evidence that they are not scientifically defensible. As Roediger (2013) says:

> "The field of education seems particularly susceptible to the allure of plausible but untested ideas and fads (especially ones that are lucrative for their inventors). One could write an interesting history of ideas based on either plausible theory or somewhat flimsy research that have come and gone over the years. And….. once an idea takes hold, it is hard to root out."

Accordingly, we will not consider any original studies or reviews of learning styles or critical thinking skills. The literature on these domains is conclusive. We will consider studies of personality, practical intelligence, emotional intelligence only if they are correlated with measures of behaviour or performance, and NOT with other self-reported measures

2) RESEARCH QUESTION, GOAL, HYPOTHESES

The focus of any scientific paper is the research question. Ideally the literature review should lead naturally to the research question and the question in turn sets the agenda for the research methods and results.

We are not at all concerned whether it is framed as a research question, goal or hypothesis; we view this as a matter of etiquette, not substance. Frequently the study design more naturally lends itself to be framed as a question or hypothesis. For example, describing a reliability or validity study in terms of a hypothesis becomes awkward. Null hypotheses look good in statistics books but are typically awkward in papers.

For qualitative or mixed methods research the authors should clearly state whether the goal is to describe a particular situation (case study), to explain it (what mechanisms are involved), to evaluate it (is it of any use or importance), or to test a particular hypothesis (does this have an impact on that), or whatever combination of goals was involved.

2a) Specificity of the question

What *does* matter is that the question is answerable by the study design. This is what characterizes scientific questions. Far too frequently the question is framed in such general terms that it is difficult to judge what kind of data would answer it. “What are students opinions of technology in nursing?”

2b) Research programs and salami-slicing

Advances was the first journal in the field to abandon a word limit. In doing so, we emulated journals in experimental psychology, where a single article may have as many as 10 or 12 studies in a carefully designed research program, so that, by the end, the phenomenon is well understood.

This is unlikely to arise in education for a number of reasons, which are not relevant here. But we retain the lack of word limit to encourage authors to publish results in a single, comprehensive paper. We abhor the practice of taking a single study and spinning it into several short papers, each of which is just a

brief snapshot of the whole study. The whole is usually more than the sum of the parts.

There are exceptions of course. Large databases frequently yield valuable insights into multiple questions. Research programs may result in multiple studies that provide more insight, while each study still stands on its own as a contribution. There is a grey zone between multiple publications that are too repetitive and a legitimate research program.

We now routinely ask authors to disclose related publications and will not hesitate to reject manuscripts that have a "me-too" character.

Directly copying any portion of a previous manuscript, even a single paragraph, without proper attribution and, where required, permission from the publisher (e.g. figures or tables) constitutes copyright infringement (the publisher owns the copyright, not the author) and is absolutely forbidden.

3) METHODOLOGY

3A) Intervention studies

One mainstay of educational research is the experimental intervention study. This may arise in validations a new curriculum approach like Team Based Learning, a simulator for instruction in motor skills, different processes of learning (e.g. student-led vs. instructor led tutorials) and so on. We are frequently criticized because we do not do enough randomized controlled trials (Torgerson, 2002).

Unlike many clinicians, we do not view the RCT as the only or the most legitimate design. It is useful for studying interventions, but interventions are only one class of educational research. Moreover, educational research has specific affordances and constraints that must be recognized in designing experiments.

3Ai) Research designs

a) One group – pretest-posttest
   Undoubtedly the most common design to examine interventions is the one-group pretest – posttest design. You measure something, do an intervention and then measure it again. Fifty years ago, Campbell and Stanley (1963) called this a "pre-experimental design" and did not view it as interpretable. Simply put, it has no control for all the other things that may have happened to the subjects during the intervention, from maturation to Google. It is not possible to draw any conclusion about effectiveness from such a design.

b) Placebo controlled

A better design is a two group design, the classic RCT. However, although in clinical medicine, with many diseases and small effects, it may make sense to conduct a "placebo-controlled" trial, this is no longer defensible in education. We have ample evidence that time spent learning something will result in more learning than no time spent learning something (Cook, 2012).

c) Curriculum – level interventions

In our view, interventions designed to determine whether an entire curriculum was effective are of minimal value, and fall under the category of "inventions" and "justification" research described earlier. They are unlikely to lead to any generalizable knowledge (Cook & Beckman, 2010).

3aii) The role of pretest

Although many researchers believe that it is essential to conduct a pretest to ensure that the "randomization" worked, we believe this is flawed logic. True randomization always works; the finding that there is a statistically significant difference on some baseline variable is not evidence that it failed. Five percent of all randomizations will "fail" by this criterion. In any case, there is no need to determine the probability that a baseline difference arose by chance. It is 100%, because it **did** arise by chance.

Practically, pretests are not learning neutral. They provide the participants with foreknowledge of the posttest, so are part if the intervention and may be as powerful as anything that is manipulated experimentally (Cook, 2009).

3A iii) To randomize or not

In medical studies, a preoccupation with randomization is likely well-placed. Effects are small and vested interests are sometimes large. In education, the opposite is more likely true. The average effect size of educational interventions is 0.5 (Lipsey & Wilson, 1998) and has been shown to be independent of whether or not people were randomized.

Randomization is a means to an end. If students end up in groups by some mechanism (e.g. clerkship assignment to various hospitals) unrelated to the intervention and outcome (e.g. learning auscultation), that is good enough.

3A iv) The outcome measure

The outcome measure should be chosen with respect to 2 conflicting goals. It should be sufficiently proximal to the intervention that it is likely to be sensitive to the differences between interventions. On the other hand, it should be sufficiently related to performance to have credibility.

Practically, this effectively rules out any outcome based on some satisfaction, confidence, or self-rated ability measure ("I am confident I can tie knots now"). As discussed earlier, self-assessment is not valid, so an outcome based on self-assessment is not credible. On the other hand, despite exhortations to look at patient outcomes, the reality is that they are relatively insensitive to most medical therapies, so are very unlikely to be sensitive to some educational intervention on students who are themselves at arm's length from patient care (Cook & West, 2013).

3B) Psychometric / assessment studies

Studies of assessment methods are the most common type of research in health sciences education. To some extent these re more straightforward than other areas, in that there are well-defined terminologies described in the APA manual and other sources. Nevertheless, some practices are unacceptable.

As a preamble, we remind the reader that self-report measures are highly suspect and cannot be accepted as the only outcome measure. Second, we recommend that you seek information from a book like Health Measurement Scales (Streiner & Norman, 2014).

Below are some serious methodological flaws associated with reliability and validity studies

3Bi) Reliability

There is one standard approach to computing the reliability coefficient – the intraclass correlation (Streiner & Norman, 2014). Cohen's Kappa is mathematically equivalent but is restricted to pairwise observations. Pearson correlation is similar, but also restricted to pairs.

Caution should be taken to ensure that the study sample is similar to the population to which the instrument will be applied. As a counterexample administering a set of diagnostic cases to samples of first and 4^{th} year residents says nothing about its application in measuring competence of first years.

Generalizability theory, an extension of classical test theory, is a very powerful alternative to classical reliability.

3Bii) Internal consistency – the reliability of the total score averaged across items is a useful statistic for some measures like high-stakes written multiple choice tests. However, with rating scales it is frequently of minimal usefulness. If you have designed a scale to assess, for example CanMEDS roles, you should expect low correlations across different roles, but you will commonly get alpha coefficients of

0.7-0.8 for these rating scales. This is likely too high, but no one says what it should be.

3Biii) Construct validity and differences among groups

Showing that an instrument gets higher scores with 4th year than 1st year students is of little value when the goal is to distinguish among 4th years. It also provides minimal information about validity, as 4th years are different from first years in all sorts of ways including debts, grey hairs, and likely of car ownership. Constructs should be much more specific than this.

3C) SURVEYS

When surveys are used in education, they tend to be "purpose – built" to address a particular question of the researcher. A consequence is that they frequently have minimal evidence of reliability and validity.

3Ci) Survey design

There are well described principles of questionnaire design (Streiner and Norman, 2014). Response scales should use minimum of 5 to 7 steps. Appropriate methods such as focus groups should be used to obtain questions.

3Cii) Psychometrics

Issues described above under "Psychometric studies" are relevant to surveys as well. Commonly, only internal consistency reliability is reported, since this can be obtained from a single administration. This is rarely informative. Some attempt to look at other areas such as test-retest reliability Is desirable.

Some attempt to establish validity of the survey, beyond simple face and content validity, is desirable.

3Ciii) Analysis

Generally surveys should, wherever possible summarize individual items into scores and subscores to improve validity and to minimize the number of possible analyses. Analysis at an item level is discouraged, unless specific hypotheses are identified a priori, and researcher takes steps to minimize Type I errors (See 4 A iii below).

3D) QUALITATIVE STUDIES

There are many branches to qualitative research, each of which has particular methodological and reporting standards. Studies should clearly articulate their theoretical stance and the basis for their work including appropriate methods and data collection. It is insufficient to simply declare that a study is qualitative. For instance, there can be a significant difference between adopting a particular methodological stance, such as grounded theory, and using some of its techniques (Kennedy and Lingard, 2006).

Patton (1999) identifies three key requirements for high quality qualitative research. Firstly, there needs to be a clear description of what was done, step by step, including who did it and the basis of these actions in established qualitative research methods with particular "attention to issues of validity, reliability, and triangulation". Secondly a clear articulation of the researchers' backgrounds and skills, and thirdly a clear articulation of the philosophical and theoretical bases of the study and how they translate into the methods used. If new methods have been developed then they need to be robustly described and grounded in theory and related approaches.

3E) MIXED AND MULTIPLE METHODS STUDIES

Not only should quantitative and qualitative components follow good practice in their respective domains, methods should be selected and pursued in ways that address the question or topic of the study and yield meaningful data that can be correlated, triangulated or otherwise integrated to create a meaningful whole.

4) ANALYSIS AND STATISTICS

4A) QUANTITATIVE STUDIES

4Ai) Descriptive data

Typically results sections fail on one simple criterion – they do NOT provide sufficient information for the reader to understand what the data actually look like. A p-value should NEVER appear in a text without the data (means, SDs, frequencies) on which it is calculated. All it says is that a difference is unlikely to arise by chance – no more, no less. And in large studies small effects can be significant; in small studies large effects can be non-significant.

It is a MINIMUM expectation that the author will provide the appropriate descriptive statistics – means, standard deviations, or frequencies. That does not mean providing all the raw data however; the goal is transparency. This can be provided in tabular or graph form, as long as meaning is clear. We do not insist on effect sizes; with adequate descriptive data this is an easy exercise in mental arithmetic. We are also indifferent to p-values or confidence intervals – again give me one, I'll give you the other.

4A ii) Parametric and non-parametric statistics

With the exception of frequencies in categories (male vs. female) it is rarely necessary to revert to non-parametric statistics. In particular, some urban myths should be dispelled:

a) Likert scales are ordinal, but can and should be analyzed with parametric stats
b) The data need not be normally distributed to use parametric stats, since methods like ANOVA look at the distribution of means, and the central limit theorem guarantees that means are normally distributed for sample size greater than 10 or so. In any case parametric stats are very robust.
c) Why parametric? They are much more universally understood, and are much more powerful. There is no non-parametric equivalent of a 3 way repeated measures ANOVA.

For more information on these issues, see Norman(2010)

4A iii) Multiple testing

It is unfortunately common, particularly with survey research, to create a table of correlations and then build a post-hoc story about the 3 correlations out of 100 that had a p-value less than .05.

When you are doing multiple testing like this – involving multiple questions in a survey, subscales on a test, different outcome measures, stations in an OSCE, it is an absolute requirement that you first do a Bonferroni correction – dividing the p – value (.05) by the number of tests. So if there are 20 correlations you are going to look at, you must use a p-value of .05/20 = .0025.

4A iv) Sample size and power calculations

Sample size calculations have a useful role in the design of a study, but, in our view, are not needed in reporting a study. If a result was statistically significant, then the sample size was large enough. If the result was not significant (and there was an expectation that it should be) then it may be appropriate to do a power calculation using a plausible estimate of the expected difference.

4B) QUALITATIVE STUDIES

Again indicators of quality should refer to the particular qualitative paradigm or tradition that is being employed (and should be clearly articulated as such). Patton's criteria also apply here. Each step in the analysis should be described and grounded and there should be a clear articulation of how findings were derived, the possibility of alternatives and the means by which trustworthiness and rigour were established and maintained.

4C) MIXED AND MULTIPLE METHODS STUDIES

In addition to the comments under the methods section, there is a difference between mixed and multiple methods (e.g. data combination or parallel processing) which should be clearly articulated and pursued. The means by which correlation, triangulation, or some other kind of synthesis is used should be described and grounded.

5) REVIEW ARTICLES – CRITICAL AND SYSTEMATIC

5A) CRITICAL/SCOPING REVIEWS

AHSE publishes many critical reviews of issues in education. There is good reason for this: we are convinced that the critical review by someone steeped in the field can offer real insight into the area – what is known, where it is heading, and what are the unanswered questions. Such a review can only emerge from deep understanding of the field.

Of course the possibility exits for author bias to emerge, as the reviewer inevitably has some personal investment in the way the area is portrayed. We believe this is of relatively little concern as the peer review process is hopefully capable of sorting this out. Nevertheless, as a minimum, the critical reviewer must specify the search strategy to some degree, although it many not be as systematic or exhaustive as those of systematic reviews.

We are more interested in the quality of the synthesis than the exhaustiveness of the search. We expect real synthesis, not a recounting of "this study did this and found that. The next study did that and found the other thing" It is on the synthesis that the critical review stands or falls.

The difference between a critical and a scoping review is that the former synthesizes and critiques understanding of an established subject or issue while the latter explores or sets out new ground for subsequent inquiry. The purpose of a review should be clearly stated and its execution reflect this purpose.

5B) SYSTEMATIC REVIEWS +/- META ANALYSIS

By contrast, we publish relatively few systematic reviews. We have seen too many examples of systematic reviews that had very limited value. There are several reasons for this, which should be borne in mind by authors.

a) Systematic reviews of quantitative research are most useful for studies of effectiveness, and less useful for other studies. Historically, Glass did the first

systematic meta-analysis of psychotherapy effectiveness. However, systematic reviews are now commonplace in medicine. There are good reasons for this: in medical interventions , the population is relatively homogeneous (Patients with MS); the therapy can be standardized (300 mg. t.i.d.) and the outcome can be standard and objective (death). Such circumstances are relatively absent in education, although there are certainly some useful systematic reviews in our field (e.g. Cook, 2011; Cook, 2008b; Issenberg, 2005). All of these were able to identify a minimum of over 100 studies on which to conduct an analysis.

b) If questions are well chosen, systematic reviews can be informative (e.g. predictive validity of medical licensing examinations, effectiveness of technology enhanced simulations). However, because of the lack of standardization of questions, therapies and outcomes, far more systematic reviews are inconclusive (e.g. does interprofessional education work? What is interprofessional education? For whom? How do you define "work?).

c) Moreover, because the outcomes are so far ranging and heterogeneous, many quantitative reviews abandon any attempt at meta analysis and end up bean counting ("12 out of 15 studies looked at self report"), which is not helpful.

In short, a credible systematic review must have:

a) A well defined question, of broad interest
b) Sufficient numbers of studies on which to base an analysis
c) Sufficient richness of data on which to draw quantitative conclusions about what is and is not effective.

d) Systematic reviews may also pursue qualitative evidence and employ qualitative methods or combine qualitative and quantitative materials and methods. Although the nature of synthesis may differ from purely quantitative reviews there are standards for qualitative reviews (such as RAMESES for realist reviews - Wong et al. 2013) and these should be followed where at all possible. To be credible qualitative systematic reviews must have:
    a) A well defined question, of broad interest, with a theoretical grounding and methods that match the question
    b) A systematic execution of the review
    c) Sufficient numbers of studies on which to base an analysis
    d) Sufficient richness of data on which to respond to the review question

References

Campbell, D. T., Stanley, J. C., & Gage, N. L. (1963). *Experimental and quasi-experimental designs for research* (pp. 171-246). Boston: Houghton Mifflin.

Cook, D. A., Bordage, G., & Schmidt, H. G. (2008a). Description, justification and clarification: a framework for classifying the purposes of research in medical education. *Medical Education*, *42*(2), 128-133.

Cook, D. A. (2005). The research we still are not doing: an agenda for the study of computer-based learning. *Academic Medicine*, *80*(6), 541-548.

Cook, D. A. (2012). If you teach them, they will learn: why medical education needs comparative effectiveness research. *Advances in health sciences education*, 1-6.

Cook, D. A., & Beckman, T. J. (2010). Reflections on experimental research in medical education. *Advances in health sciences education*, *15*(3), 455-464.

Cook, D. A., & West, C. P. (2013). Perspective: Reconsidering the focus on "outcomes research" in medical education: A cautionary note. *Academic Medicine*, *88*(2), 162-167.

Cook, D. A., Hatala, R., Brydges, R., Zendejas, B., Szostek, J. H., Wang, A. T., ... & Hamstra, S. J. (2011). Technology-enhanced simulation for health professions education: a systematic review and meta-analysis. *Jama*, *306*(9), 978-988.

Cook, D. A., Levinson, A. J., Garside, S., Dupras, D. M., Erwin, P. J., & Montori, V. M. (2008b). Internet-based learning in the health professions: a meta-analysis. *Jama*, *300*(10), 1181-1196.

Flay BR, Biglan A, Boruch RF, Castro FG, Gottfredson D, Kellam S , Moscicki EK, Schinke S, Valentine JC, Ji P. (2005) Standards of Evidence: Criteria for Efficacy, Effectiveness and Dissemination. Prevention Science; 6(3):151-175.

Issenberg, S.B., McGaghie, W. C., Petrusa, E. R., Lee Gordon, D., & Scalese, R. J. (2005). Features and uses of high-fidelity medical simulations that lead to effective learning: a BEME systematic review*. *Medical teacher*, *27*(1), 10-28.

Kennedy TJT, Lingard LA. (2006) Making sense of grounded theory in medical education. Medical Education 2006; 40: 101–108

Norman, G. (2004). Editorial–Theory Testing Research Versus Theory-Based Research. *Advances in Health Sciences Education*, *9*(3), 175-178.

Norman, G. (2010). Likert scales, levels of measurement and the "laws" of statistics. *Advances in health sciences education*, *15*(5), 625-632.

Patton MQ. (1999) Enhancing the quality and credibility of qualitative analysis. Health Serv Res. Dec 1999; 34(5 Pt 2): 1189–1208.

Prideaux, D., & Bligh, J. (2002). Research in medical education: asking the right questions. *Medical Education*, *36*(12), 1114-1115.

Roediger, H. L. (2013). Applying Cognitive Psychology to Education Translational Educational Science. *Psychological Science in the Public Interest*, *14*(1), 1-3.

Streiner, D. L., & Norman, G. R. (2014). *Health measurement scales: a practical guide to their development and use*. Oxford university press.

Streiner DL, Norman GR. (2009) Efficacy and effectiveness trials. Community Oncology; 6(10):472–474

Torgerson, C.J. (2002) Educational research and randomized trials. Medical Education, 36, 1002-1003.

Wong G, Greenhalgh T, Westhorp G, Buckingham J, Pawson R. (2013) RAMESES publication standards: realist syntheses. BMC Medicine 2013, 11:21