Strategies for Identifying a Research Topic in Educational Measurement

Jason Millman
Cornell University

How might a novice researcher go about setting a research agenda? What are the most productive avenues for generating research questions? What areas of educational measurement currently seem most promising?

Editor’s note: At the 1997 NCME Annual Meeting, Jason Millman made an invited presentation containing advice to beginning researchers. A revised version of that talk is offered here to let a broader audience benefit from Jason’s wisdom and friendly counsel. We will miss him.

By way of introduction, here is a paraphrase of an anecdote found in Robert Mager’s (1962, p. IX) classic book on behavioral objectives:

A shark and a sea horse chanced to meet. The shark asked, “Where are you going?”

“I’m going to find my fortune,” replied the sea horse.

“Do you know where that might be?” asked the shark.

“Well, not exactly.”

“You’re in luck,” grinned the shark. “If you’ll just follow this path,” he said pointing to his open mouth, “you’ll get there in no time.”

“Gee thanks!” said the sea horse and swam off to her doom.

The moral of the story is if you don’t know where you are going, you might not like it when you get there.

We all know young researchers who are floundering in shark-infested waters. It has been my experience that the single biggest difficulty they face is generating a research topic. That task alone can sometimes take an inordinate amount of time and, if not clearly formulated, can defeat the researcher. Pity the doomed researcher who leaves the safety of supportive waters and ventures into real-job seas without a clear idea about his or her research direction. Chances are the researcher will be lost at sea.

Following the lead of David Letterman, and our own Ron Berk, I give a list of the nine most productive ways of identifying a research topic and illustrate by example, typically from the field of educational measurement.

Number 8: You Can See a Lot by Looking

Ideas for research can come from anomalies, puzzling observations, a series of contradictory facts, or similarities among distant topics. As Cronbach and Suppes (1969) put it:

As investigators discuss their results with colleagues, a restlessness emerges. Findings from different studies seem to conflict. Phenomena are noticed that cannot be well summarized in the available language. Similarities are observed among findings on what were previously considered to be distinct topics. (p. 128)

Provocative statements, uncharted areas, unverified findings, and interesting ideas can also serve as sources for identifying research problems. For example, Josephson’s research took off from Allison Davis’s claim that “Students in slum schools may well find it more rewarding to be considered academic failures than to be successes in school” (cited in Millman & Gowen, 1974). A more current example is President Clinton’s perplexing, if not provocative, statement that his

Jason Millman was a Professor of Educational Research Methodology, Department of Education, Kennedy Hall, Cornell University, Ithaca, NY 14853. His specialties were evaluation, research methodology, and statistics.

Summer 1998
goal is for every child to read by the third grade.

Once-in-a-lifetime events can also serve to provide fodder for research ideas. The initial truth-in-testing legislation and the trial in federal court of the Florida graduation test are two instances that profoundly affected the measurement community and led to many research papers. More contemporarily, the redesign of the National Assessment of Educational Progress will be a significant event and begs for the chronicling of its consequences.

Number 7: Give a Kid a Hammer and the Kid Will Hit Everything in Sight

According to the rule, give psychological researchers a new analysis technique or measurement device, and they will discover all their research can profit from its use. When factor analysis became feasible, researchers factor analyzed every data set they could get their hands on. A few years later, locus of control instruments edged their way into scores of studies. These tools have served and are serving us well.

My contemporary nominee for the hammer is sensitivity analysis. By sensitivity analysis, I mean assessing the effect of the input variables on the results. In the days before the widespread use of computers, one picked a statistical model and lived with the consequences. Now that is no longer necessary. Multiple analyses can be performed in fractions of a second. Researchers no longer have to make a choice. They can show how the difference in their initial choices affects the final outcome. A range of input values and/or models and the corresponding results would replace the winner-take-all approaches now in use. Welcomed, at least by me, would be research that describes how sensitivity analysis is used in fields other than education and illustrates how it might look applied in an educational setting.

As an example of the influence sensitivity analysis can have on our thinking, a recent graduate student of mine was faced with analyzing data on course grades in which F (failure) = 0 and grades D through A+ received scores of 60 to 100. Should she continue to fail a grade zero, remaining true to the grading scale, or should she set a grade of failure equal to 50 and more likely be true to the underlying interval scale of the course achievement? It is not an either-or decision. My graduate student could do her analysis both ways, display her results, and let the readers make whatever interpretation they wished.

Number 6: Dig for Gold in Them Thar Hills

There is now an abundance of riches; gold mines of data are available to be worked. I refer to data available from large projects such as the National Assessment of Educational Progress, High School and Beyond, and the National Longitudinal Study. Using these data as a source of information can be tremendously efficient. Individuals could collect on their own even a small fraction of the data in these banks. Besides, the archivists of these collections usually go out of their way to facilitate the use of the data. The downside of using existing pools of data is that often the question you want to address in your research cannot be answered with the information available. The possibility of changing the question to fit the data should not be overlooked.

Number 5: Look a Gift Horse in the Mouth

Occasionally we are blessed with an entire area ripe with possible research topics. Such was the case when Lee Shulman called attention to the systematic ignoring of teachers' knowledge of subject matter as a central construct in the study of teaching. Shulman laid out a number of research questions that his central thesis suggested. One measurement question still to be answered is whether the ability to teach a particular subject is anything more than can be accounted for by general teaching ability plus subject matter knowledge.

Another example of a broad research area suggesting specific research studies is computer-assisted testing. Specifically, Randy Bennett and his colleagues at Educational Testing Service posed a set of questions concerned with bringing multimedia to existing tests (Bennett, Goodman, Hessinger, Liggett, Marshall, Kahn, & Zack, 1996). Their list of questions divides into four groups: measurement, test development, delivery, and finance.

Number 4: Milk It for All Its Worth

The previous suggestion for identifying a research topic recommended looking toward an area of research. This suggestion involves looking for a research topic from an outstanding piece of existing research. As one wise poet said: "Problems worthy of attack/Show their worth by hitting back."

Clearly, a piece of scholarship that would qualify as a rescuer of distraught researchers because it opens up a number of investigations is Messick's work (see, e.g., Messick, 1989) on consequential validity. Leaving aside the question of whether consequential validity is validity at all, plenty of work remains. What consequences are we talking about? If only intentional consequences then whose intentions? The users? Which users? The test sponsors? The test developers? What should a consequential validity study look like? What test standards apply, and what new ones need to be written?

Number 3: Twenty (Plenty) Questions

Any study has unfinished business, which can be exposed most easily by what-if questions that challenge the external validity of the study. For its usefulness in suggesting ideas for research, the study itself need not be of special value, except perhaps that it is allied with a topic of interest to the researcher.

For example, consider a study by Zirkel and Moses (1971) which investigated relationships between self-concept and ethnic group membership and whether the ethnic group was in the majority or not in that school. Zirkel and Moses found that, compared to Puerto Rican students, Black and White students had higher self-concept scores and that these differences were greater when their ethnic group was in the majority of students at the school. Before accepting the generality of those findings, one might want to conduct additional research which addresses such questions as: Does the same pattern appear for other measures of self-concept? Does self-
concept in school differ from self-concept out of school? Is the pattern similar for traits other than self-concept? Are differences based more on language differences than ethnic differences?

**Number 2: Necessity Is the Mother of Invention**

Statistical consultants have long known that a valuable source of research topics can be found in the consulting problems that are presented to them. For example, a recent graduate student of mine had to evaluate the quality of an examination in which the test takers had a choice of which questions to answer. Computing the reliability of such examinations could be a study of its own. Another example of how a real-world problem can suggest needed research is the experience of many test directors. They often are frustrated at the half-hearted efforts of their senior high students. The performance of second semester seniors is particularly suspect. The researcher who can devise a feasible scheme for motivating such students to do well would probably have a more positive future.

**Number 1: Read Much About Little**

The suggestion for identifying a research topic that is ranked first is to become an expert in some aspect of your field. It is generally true that the more you know about a subject, the more you don’t know, and the “don’t know” becomes candidates for research topics. To get to this level of expertise, investigators must read more than textbooks. They must become familiar with the research literature and fugitive materials relevant to the subject.

**Two Examples of Promising Areas of Research**

Let me conclude this piece with two examples of research topics that I consider to be particularly promising. The first proposed topic is reliability of group processes. Many components of test development and score interpretation depend on group work. Test specification is often the result of compromises among the members of a sometimes large and allegedly representative group. Which items should appear on a test is often a group decision. Setting a passing score often involves the give and take of a cut-score committee. In computing the reliability of the decision (e.g., the passing score that was set), we usually behave as though the ratings of the judges were independent. Rarely do you find multiple committees at work at the identical task so that the reliability of the groups could be assessed more directly. What difference does this make? Does the consistency depend on the magnitude of the interactions among group members and on the extent to which the raw data are referenced?

The second research idea is to conduct an evaluation of a scoring scheme that is hypothesized to increase scorer reliability and reduce scoring costs over existing methods. The scheme only works well when the answers of a knowledgeable examinee cannot be anticipated ahead of time. It is appropriate when the skill of writing itself is being assessed or, more generally, when any number of answers can be correct or can demonstrate the skill. This scheme works less well when quality is determined more objectively by seeing if an examinee gives anticipated, specific responses.

Analytical and holistic scoring procedures have been proposed to deal with items where a range of acceptable, but unanticipated, responses occur. With either scoring procedure, the response is graded on one or more scales. The score categories (e.g., 1, 2, 3, 4) are defined either in general terms that can be applied to several constructed items on a test or in terms specific to a particular question.

Regardless of whether the criteria are general or specific, scorers are typically given one or more examples of model responses at each score point. From the scorer’s perspective, the task is to assign an answer to a particular score category. That task is difficult because few responses exactly mirror any given category description or model response. The typical instance of unreliability is where two scorers disagree by selecting adjacent scoring categories. This disagreement represents a difference of opinion about which category definition or model response the examinee’s response was closer to.

In contrast, the scheme proposed here is to give model responses that lie halfway between two score points. For example, instead of having model responses for Points 3 and 4, the model response would be for answers that would be a toss-up between a high 3 and a low 4. The scorer’s task now becomes a judgment of whether a particular response is better or worse than the model response. It is my belief that judgments of dominance will be more reliable than judgments of proximity.

**Reprise**

A shark and a sea horse chanced to meet. The shark asked, “Pst, where are you going?”

“I’m going to find my fortune,” replied the sea horse.

“Do you know where might that be?” asked the shark.

“Well, not exactly.”

“You’re in luck,” grinned the shark. “If you’ll just follow this path,” he said, pointing to his open mouth, “you’ll get there in no time.”

“No, thanks!” said the sea horse and deftly avoided the scissor-sharp jaws.

With a research agenda, the beginning investigator leaves the supportable waters, enters the stimulating real-job seas, and has many successful voyages.

**References**


