# USING QUALITATIVE METHODS TO VALIDATE QUANTITATIVE SURVEY INSTRUMENTS

**John E. Mullens, Policy Studies Associates**
**Daniel Kasprzyk, National Center for Education Statistics**
**John E. Mullens, PSA, 1718 Connecticut Avenue, N.W., #400, Washington, DC 20009**

Key Words: validation, case study, focus group

Surveys are among the most cost-effective and least burdensome methods of collecting data on schools, classrooms, and teachers, but both researchers and respondents know that brief, self-report strategies may not portray a picture of instruction as sufficiently as needed to confidently assess instructional effectiveness. For a project at the National Center for Education Statistics (NCES) designed to investigate techniques and instruments to measure and understand the instructional processes used by eighth to tenth grade mathematics teachers, we used a multi-step pilot-study process to construct, refine, and validate survey instruments. In the process, we honed our knowledge about instruments and methods for collecting accurate, valid, and meaningful information that can be incorporated into future national data collection schemes.

In this paper, we outline the project scope, describe the data collection methods used, and assess their role in evaluating survey responses and improving instruments to provide portraits of relevant classroom processes. A complete description and assessment of this project is reported in Mullens and Leighton (1996).

## The Classroom Instructional Processes Study

The NCES project, "Understanding Classroom Instructional Processes," was designed to (1) develop, pilot, and evaluate methods for collecting data on classroom instructional practices; (2) explore the combined use of questionnaires and related teacher log forms to portray classroom instructional processes; and (3) determine the feasibility of incorporating such methods into future NCES surveys or other data collection efforts. The project piloted focus groups and case studies (using classroom log forms, observations, and artifact collection) to assess the completeness and accuracy of data obtained from questionnaire responses. Through this process, we used qualitative methods to validate responses on quantitative survey instruments.

The results were intended to help NCES make decisions about data collection methods and instruments with which to develop an accurate portrait of eighth to tenth grade mathematics instruction. Having such data would expand NCES's ability to respond to Congress, other offices in the Department of Education; and other federal agencies, state departments of education, associations concerned with elementary and secondary education, and education research organizations. Data from previous surveys on similar topics have been used by all of these sectors, and in recent years there has been interest in expanding the scope of these data.

## Context

Increased use of high stakes student testing as a measure of educational productivity has led to increased interest in determining the precise contribution of schooling to achievement, distinct from, for example, the contributions of prior learning or socioeconomic status. Experts in identifying the correlates of student achievement, such as Porter (1991) and Schmidt (1995) argue that many factors are at work: the content must be presented cogently, using subject-specific instructional techniques appropriate to both the material and the students' prior knowledge, and with emphasis matched to the topic's relative importance among desired outcomes. Valid and reliable assessments of instructional content and practices can contribute to descriptions of educational experiences, help explain achievement outcomes, and inform educational policy development at the local, state, and national levels (Burstein, Oakes, Guiton, 1992; Smith, 1988; Murnane, 1987). Despite this, minimal data on instructional practices are available from a nationally representative sample of U.S. classrooms.

This study builds on a prior review of existing measurement approaches (Mullens, 1995), and focuses on four major dimensions of classroom instruction: the conditions and context that direct or influence a teacher's selection of content and instructional methods; the course content and emphasis on those topics; patterns of classroom pedagogy and how teachers approach the process of teaching; and the resources available and used in the classroom.

## Research Question

The study goal was to produce and evaluate instruments and methods that would provide data on how the instructional processes and content of eighth to tenth grade mathematics classes vary across the country. Within this overall goal, we also expected to advance our understanding about instruments and methods for

capturing accurate and meaningful information about classroom instructional processes: information that could be incorporated into national data collection schemes. Data from the field tests could provide evidence with which we could understand more about the items and instruments themselves. The full study explored three measurement questions, one of which is the focus of this paper:

> Do qualitative data collection instruments and techniques provide validating information with which we can better construct questionnaires?

**Study Design**

We selected mathematics as the focus of this study because it is a core subject of great interest to policy makers and thus one in which early exploratory work already provided a sound research foundation for further study. Within this content area, the study concentrated on eighth to tenth grade mathematics courses, covering topics in pre-algebra, algebra, and geometry: the math courses designed to serve as a "bridge" to more advanced math courses, and to offer students a conceptual understanding of mathematics with broad applications in life.

From preliminary research, we decided to validate the teacher questionnaire through focus groups and case studies. Focus groups are roundtable open discussions with small numbers of respondents who had already completed the questionnaire, to examine each item for unfamiliar or inexact terminology and how well the items and their responses represented their own teaching.

Case studies of classroom teachers included observing their teaching, having them maintain daily logs, and collecting artifacts of their instruction. We observed classroom instruction to evaluate the completeness of the instrument, looking especially for conceptual gaps in our understanding of how instruction occurred, and in how it was represented on the instruments. Daily logs, or diaries, were records documenting learning objectives, teachers' actions, students' activities, and the materials used during a single class. Four weeks of data enabled us to evaluate the consistency between teacher's questionnaire responses and her daily recordings of activities. Examining the instructional materials or artifacts used by teachers during that same period of time were intended to provide information on the same events from a different slant.

Elements of these processes to validate questionnaire items have been explored and improved in several recent studies in this field, including Reform Up Close (Porter, Kirst, Osthoff, Smithson, & Schneider, 1993), Third International Study of Mathematics and Science (1991), and Validating National Curriculum Indicators (Burstein, McDonnell, Van Winkle, Ormseth, Mirocha, & Guiton, 1995).

We piloted our instruments and process in two school districts, revised them, and obtained OMB clearance. We fieldtested the instruments and process in three school districts: one was a large, independent, urban district on the West Coast; the second was a large city/county urban district in the Southeast; and the third was a smaller, suburban/rural, county district in the Mid-Atlantic region. In all, 111 teachers completed questionnaires, and seven teachers, one or more from every field test site, participated in the case study.

**Data Collection Methods**

We used focus groups and case studies to validate responses on the teacher questionnaire.

**Focus Groups**

At each of the three sites, all teachers of eighth, ninth, and tenth grade mathematics received a letter of invitation that explained the study and requested their participation. Attendance at the seven voluntary focus group meetings ranged from one to 12 teachers and totaled 38. Teachers commented on their understanding of the item's intent and the appropriateness of the response format.

The greatest teacher concern across all sites was not (as we suspected) that the teacher questionnaire would not adequately portray their teaching, but that the particular class they were asked to describe (the first instructional period of the day) was not representative of their whole teaching load. Specific characteristics of the students in that class, according to most teachers, caused them to teach in some manner they felt was not representative of their overall efforts. Despite this concern, most focus groups came to the conclusion that while there was no single period that would catch each of them at their most representative, the combined results of all sampled teachers would indeed represent the overall collection of the activities of all teachers throughout the day.

## Case Studies

Volunteer case-study teachers were observed by a project researcher, kept a daily log of classroom instructional activities and those of the students in their designated class during a four-week period, and collected instructional artifacts. The project had seven case studies.

***Classroom logs.*** Information from classroom logs was used to assess the consistency of teachers' daily recordings of classroom practice with their one-time account of practice from the teacher questionnaire. The picture of classroom practice obtained from multiple weeks of log form data is a finer grained view of the enacted pedagogy than that provided by a teacher's questionnaire responses summarizing a semester of practice. For both practical and perceptive reasons, completing logs daily (or at most, weekly) is likely to result in more accurate data than a one-time retrospective survey. Because logs rely on teachers' short-term memory rather than their long-term memory and the summative ability needed for the questionnaire, the resulting data can be presumed to be more accurate. Furthermore, teachers may be more inclined toward honest accounts on a daily rendering since a single daily log, unlike teacher questionnaire responses, becomes one of many depictions of their practice. Classroom logs were also intended to be used by researchers to record events and activities as they observed in classrooms.

A weakness of prior research had been the inability to use data from classroom logs to estimate the reliability of questionnaire items (Porter, 1993). This difficulty stemmed from using different items of the log than were included on the questionnaire. We designed the log to be completed by case-study teachers as a record of the classroom instruction occurring during a single class period so that the data from four weeks of logs could be used to evaluate the validity of the teacher's responses on the questionnaire. To make this direct link possible and to build on the knowledge gained from prior research, we constructed the log by directly copying specific items and activities from those on the questionnaire; frequency response options covering a semester were replaced with time per use response options covering a single period. Sharing identical items between the two instruments was intended to facilitate the later comparison of the teacher's daily logs with her responses on the survey.

***Classroom observations.*** Researchers observed case-study volunteers to help them understand the function of the classroom log and the process of using it. After observing the teacher instructing the targeted class, both the researcher and the case-study teacher completed a log form. Teacher and researcher then compared observations, discussing differences in coding. For all but two teachers, those differences were slight. Because these teachers had participated in the focus group discussions of the items, most already understood nuances of meanings that might make a difference in how they recorded their instruction. Researchers had enough concerns about the coding patterns of one teacher, however, to repeat the calibration process a second time.

***Artifact collection.*** To provide further detail about their lessons (and reduce the need for written explanations), case-study teachers were asked to submit certain instructional items figuring prominently in lessons for the designated class. Such instructional items included copies of homework and in-class assignments; directions for papers, reports, or projects; copies of tests and quizzes; and any other written assignments. These artifacts were intended to provide another avenue through which researchers could interpret the teacher log data for each lesson.

## Assessing the Methods

While each qualitative method helped validate the quantitative data obtained from the teacher survey, some contributed more information than others to our analysis.

## Focus Groups

The purpose of focus groups was to provide respondent feedback on the survey instrument. They served that purpose well, and, unexpectedly, proved to be a major source of case-study volunteers. Especially at the beginning, focus groups allowed researchers to directly hear respondents' comments and probe their exact meanings. Such exchanges allowed both researchers and respondents to raise and explore many issues usefully, to validate the relevance of certain items across sites, to hone wording, and to generate additional ideas of emerging instructional practices. For example, an earlier version of the questionnaire included "calculator" on the list of instructional materials. Teachers were asked to indicate if calculators were available for use by students. In one focus group, teachers suggested that was not the issue. They had plenty of calculators available and even sufficient batteries. But the calculators were not sophisticated enough to do the kinds of operations the teachers wanted to teach their students. Because of that

discussion, the item was changed to "appropriate calculator".

Teachers at another site complained that the questionnaire afforded them no opportunity to indicate they structured their classes around cooperative learning strategies. They explained that cooperative learning was a major effort in the school district's instructional program, yet there was no place on the questionnaire to indicate the use of that strategy in the classroom. When that same comment was heard in another location, it was added to the list of teacher activities and student activities.

As the project continued and researchers held additional focus groups in new fieldtest sites, however, the utility of additional information new to researchers substantially decreased. We obtained little new information from the later focus groups.

## Case Studies

The case studies provided substantial information with which to assess the construction of the teacher questionnaire.

***Classroom logs.*** The project was not funded to design a process to validate the reliability of questionnaire items, but to fully understand the benefits and limitations of the information that we could obtain from logkeeping, we measured the consistency between teachers' daily reports of instructional activities and their semester report of the same activities. Assuming that the daily log reports had a higher level of teacher reporting accuracy than the questionnaire responses for reasons stated above, we used log responses to assess the reliability of the questionnaire responses. For each case-study teacher, we compared the sum of log-reported activities across a representative four week period with that teacher's questionnaire-reported activities over the semester. For example, if the questionnaire responses indicated that the teacher stimulated student discussions of multiple approaches more than once a week, we expected to see confirming entries of such discussions on the teacher's daily log. This provided a measure of the reporting reliability of individual questionnaire items.

Our sample of seven case-study teachers was purposive and too small to generalize to the larger sample of all survey respondents; nonetheless, Table 1 illustrates the type of information we might obtain using this process with a larger and appropriately random subsample of case-study teachers.

Table 1: Examples of consistency between (a) teachers' survey responses describing a semester of classes and (b) their class log entries over a four-week period, on the same teacher activity (nonrandom sample, N=7).

| Teacher Activities | Percent direct agreement | Percent agreement within one survey response value category |
|---|---|---|
| Provide individual or small group tutoring as needed during individual seatwork or small group activities involving everyone | 100 | NA |
| Lecture, perhaps occasionally using the board or overhead projector to highlight a key term or present an outline | 71 | 86 |
| Demonstrate a concept, using two-dimensional graphics such as drawings on the board, overhead projector, or computer | 71 | 86 |
| Provide supplemental--remedial or enriching--instruction to a pull-out group while the rest of the class works in assignments | 71 | 86 |
| Administer a test or a quiz | 57 | 86 |
| Demonstrate a concept, using three-dimensional tools such as manipulatives, models, or other objects | 57 | 71 |
| Lead students in discussion, recitation, drills, or question-and-answer sessions | 43 | 100 |
| Observe or monitor student-led discussions | 43 | 57 |
| Work on administrative tasks while students work on assignments | 29 | 57 |

Table reads: In a nonrandom sample of seven teachers over four weeks, teachers' responses on a survey item about tutoring were consistent in 100 percent of cases with their responses on the log item on the same topic. Teachers' responses on the survey item about lecturing were consistent with their log responses in 71 percent of cases and within one response value in 86 percent of cases.

These data, reporting consistencies between the teacher questionnaire and the log form for teacher activities from item 13 of the questionnaire, suggest that teachers' recollection of their instructional behavior varies according to the activity being reported. Examining the extreme cases, for example, there was one hundred percent consistency between teachers' reports on the questionnaire and on their daily logs about the frequency with which they provided individual or small group tutoring. Teachers apparently remember that type of instruction well. There was far less agreement (29 percent) between questionnaires and logs on how often the teacher works on administrative tasks while students work on assignments. Temporarily relaxing stringency to assess agreement within one survey response category increases the level of consistency between logs and questionnaires on this same item but only to 57 percent.

For four items, the rate of agreement between responses on the two instruments is above 70 percent, while in five it is 57 percent or less. To better understand these differences, we examined the direction of the mismatch of the five items with low agreement for possible evidence of socially desirable questionnaire responses. For three such items, data from the daily logs reported that the following activities actually occurred *more frequently* than teachers' questionnaire responses would indicate:

- administrative tasks
- drill and recitation
- student-led discussions

In one light, these responses may show evidence of social desirability, since the first two activities could be considered old fashioned or less than desirable in a climate of reform. Such an argument would suggest that subtle pressures may have influenced teachers' questionnaire responses. That student-led discussions appear to have actually happened more often than teachers indicated they did does not seem to follow that same explanation.

For two other items, data from the daily logs suggested that the following activities occurred *less frequently* than teachers' questionnaire responses would indicate:

- demonstrating a concept with three-dimensional tools
- administering a test or quiz

These differences suggest that teachers like to think they use three-dimensional manipulatives more than they actually do and that teachers administer fewer tests and quizzes than they might think.

Recalling again that this is only a demonstration analysis based on seven nonrandom sets of logs, we suggest no generalization of results beyond these seven teachers. The process, however, seems to show promise. Specific results from a larger and representative validation study might be different, but would likely be no less interesting.

***Classroom observations.*** We designed the classroom observations of case-study teachers and the later discussion about completed log forms to provide those teachers with an experiential-based understanding of the meanings of the log form terms, and with practice in completing the form. Discussing specific events occurring within a particular class and how they translated into log form responses established common understandings of log form terms more directly than would have resulted from an abstract discussion only. Conducting multiple classroom observations across different sites and the resulting observation data also provided researchers with evidence with which to assess (1) the ability of the survey instrument to portray classroom processes accurately and (2) the match between actual classroom practice and survey scope, individual items, and response formats. In addition, the nonjudgmental research approach to discussing the observed instructional activities proved to be an unanticipated and effective method for cementing teacher cooperation and building confidence in the process.

During the fieldtests, researchers used their copy of the observation form to record classroom activities as they occurred, creating a real-time log of the instructional processes occurring during a single class session. Having the researcher use a more structured classroom observation instrument with which to initially record teacher and student activities and elapsed time may improve researchers' understanding of how teachers record their instructional processes on the log form, and may result in a more accurate recording of the duration of instructional elements occurring during instruction and the order in which they occurred.

***Artifacts.*** We collected artifacts from case-study teachers to investigate the potential of such documents to more completely describe or illuminate classroom instructional processes. Although this process was

inexpensive, it afforded little analytical benefit to this study. The artifacts collected were primarily assignment sheets and examples of student work. We know in some cases, and suspect in others, that participating teachers sent incomplete records of the mathematic textbooks they used. Textbook pages or items used during instruction were the most notable void. Those artifacts we did have were difficult to assess. We can identify, for example, what was used during class (e.g., a practice sheet) and evaluate some elements of its content (e.g., estimation) but can tell little from the artifact itself about the instructional objective being addressed, how the artifact was used, or the amount of emphasis given to each element of the artifact. In further experiments with artifacts, we would investigate developing (1) a teacher checklist of contextual data surrounding the artifact's use within the lesson; and (2) a specific protocol for assessing important features of the artifact (such as instructional objective) and its use. Such protocols may be time consuming (and therefore expensive) to implement, effectively negating the original low cost of collecting the artifacts. So although artifact analysis may have great potential to add substance to self-reports, the process needs further attention.

**Summary**

With this task, we evaluated the usefulness of supplementary data collection in the form of focus groups and case studies in contributing to an understanding of how well our instrument measured the domains of interest, and how well the survey responses represented what teachers actually do. The focus group discussions provided excellent feedback on the survey, but are limited in the amount of new information provided by multiple focus groups. The case-study process, and classroom logs in particular, provide a valuable estimation of the consistency between responses on teacher questionnaires and on class logs. Classroom observation is beneficial to the researcher's understanding of the phenomenon under investigation and to the process of gaining trust for later segments of data collection, but we were disappointed in the results of our attempts to use artifacts to expand our understanding of classroom instructional processes, and see further experimentation as the key to greater benefits.

Based on the results of the research described above, we think certain qualitative methods can expand the ways in which we validate quantitative survey instruments. We are currently embarking on a project to survey a sample of 400 teachers of eighth to twelfth grade mathematics, engaging a subset of 60 in case studies. We will use the results reported here to expand our use of classroom observations and logs to validate the quantitative survey instruments.

**References**


Burstein, L., Oakes, J., & Guiton, G. (1992). Education indicators. In M.C. Alkin (Ed.), Encyclopedia of educational research (5th ed., pp. 409-418). New York: MacMillan.

Burstein, L., McDonnell, L., Van Winkle, J., Ormseth, T., Mirocha, J., & Guiton, G. (1995). Validating national curriculum indicators. Santa Monica, CA: RAND.

Mullens, J. (1995, April). Classroom instructional processes: A review of existing measurement approaches and their applicability for the Teacher Follow-up Survey (Working Paper No. 95-15). Washington, DC: National Center for Education Statistics.

Mullens, J., & Leighton, M. (1996). Understanding classroom instructional processes (draft). Washington, DC: Policy Studies Associates.

Murnane, R. (1987). Improving education indicators and economic indicators. Educational Evaluation and Policy Analysis, 9(2), 101-116.

Porter, A. (1991). Creating a system of school process indicators. Educational Evaluation and Policy Analysis, 13(1), 13-29.

Porter, A. (1993). Defining and measuring opportunity to learn. The debate on opportunity-to-learn standards: Supporting works. Washington, DC: National Governors' Association.

Porter, A., Kirst, M., Osthoff, E., Smithson, J., & Schneider, S. (1993). Reform up close: An analysis of high school mathematics and science classrooms. Madison, WI: Wisconsin Center for Education Research.

Schmidt, W. (1995, June). Presentation made at the Council of Chief State School Officers National Conference on Large Scale Assessment, Phoenix, AZ.

Smith, M. (1988, March). Educational indicators. Phi Delta Kappan, 487-491.

Third International Mathematics and Science Study. (1991). Project overview. East Lansing, MI: Author.