VALIDATING ITEM RESPONSES ON SELF-REPORT TEACHER SURVEYS

John E. Mullens, Mathematica Policy Research, Inc.
Daniel Kasprzyk, National Center for Education Statistics
John E. Mullens, Mathematica, Inc., 600 Maryland Ave, SW, Suite 550, Washington, DC 20024

Key Words: instructional processes, self report, pretesting, validity, reliability

This paper describes survey development work being done by the Elementary and Secondary Sample Survey Studies Program of the National Center for Education Statistics (NCES) to develop items and processes to collect accurate, valid, and meaningful data on the instructional processes teachers use with students. Instructional processes (IP) data can illuminate how teachers and students work together to approach classroom educational objectives: the emphasis teachers put on important topics within the curriculum, the learning objectives they have for their students, the activities in which students and teachers engage, and the ways in which teachers assess student learning. This information is important for national, state, and local policymakers and others interested in how school reform affects classroom practice. This article reports on the NCES Classroom Instructional Processes Study, conducted in 1997-98 and more completely described in Mullens & Gayler (1999). That work is one of a number of activities through which NCES is both collecting IP information and examining, refining, and improving the quality of data collection methods and instruments they use in national data collection programs.

NCES became interested in this line of data collection in 1994 when Commissioner Emerson Elliot authorized a comprehensive review of then-current research efforts (Leighton, Mullens, Turnbull, Weiner, & Williams, 1995), an analysis of measurement approaches (Mullens, 1995), and development of a module of items to measure IP for the Current Teacher's Questionnaire of the SASS 1994-95 Teacher Followup Survey. Following that data collection, NCES continued to fund IP item development and refinement.

Why Measure Instructional Processes

Societal demands on schools and teachers and the resulting close scrutiny of educational outcomes continue to heighten interest in how schools and teachers can better do their jobs. The desire to understand variation in student outcomes leads policymakers and researchers to seek a better understanding of how teachers and students approach the math curriculum. How do teachers approach math instruction? Do teachers use different techniques when they emphasize broad concepts compared to specific facts or processes? To what extent do teachers use newly-recommended instructional techniques? Do they integrate new techniques with their "tried-and-true" methods? To the degree that differences in teachers' instructional practices directly affect the quality of learning in the classroom, answers to such questions will inform our understanding of effective approaches to student learning.

When IP data are combined with information on student learning, policymakers, teacher trainers, and professional developers have the means to guide instructional techniques toward those that are most effective in creating desired student outcomes. Understanding how variation in student learning relates to variation in instructional methods could inform local, state, and national education policy (Burstein, Oakes, Guiton, 1992; Smith, 1988; Murnane, 1987).

Stodolsky (1996) summarized the rationale for generating a broadly representative, yet finely-textured data base of information about classroom-level IP:

> "If we are to understand, monitor, and improve our nation's schools, accurate and timely empirical, descriptive data about how schools' work must be available. The activities that take place in classrooms to engender student learning and development are the heart of any school's education efforts. It is in the transactions between and among teachers, students, materials, and tasks that deliberate efforts to educate occur. Descriptive information about how teaching and learning occur in classrooms and about what is taught provides the basis for monitoring the status of instruction in a large number of settings. Such information can provide periodic assessments of stability and change in instruction, particularly as changes relate to deliberate efforts to reform or alter curriculum and instruction."
>
> (Stodolsky, 1996)

Survey data are likely to be the major source of nationally-representative information about instructional content and practices, but there are questions about the quality of such data.

Potential Threats to the Reliability and Validity of Self-reported IP Data

While well-designed focused surveys can be cost-effective for administrators and place only limited burden on respondents, the accuracy of self-reported responses sometimes calls into question the reliability and validity of the resulting data. There are at least three reasons why this might be so (Mayer, 1999). First, teaching and learning in any context is a complex human endeavor that cannot yet be adequately represented by responses to a single survey. Second, some survey items may contain unknown phrases or ambiguous concepts that make an appropriate response difficult. Finally, for reasons beyond a survey's scope, some teachers may be sensitive to particular questions and/or the concepts they represent and therefore feel pressured to provide (perhaps socially desirable) responses that are less than accurate. These and possibly other equally serious concerns pose serious threats to using surveys to accurately portray instructional practices. Therefore, the quality of the survey items needs to be initially validated and periodically confirmed (Burstein, McDonnell, Winkle, Ormseth, Mirocha, Guiton, 1995).

The Fieldtest

To explore these possible threats to the reliability and validity of the self-reported data, this fieldtest set out to determine the accuracy of teachers' descriptions of classroom instruction when recorded on a daily basis and over one semester. It included a mail questionnaire sent to approximately 400 math teachers of eighth to twelfth grade students and a case study of 41 teachers in similar settings. Case study teachers were volunteers and received no financial incentive to participate in this study. Mail survey respondents described their instruction in one designated math course over the previous semester; case study teachers responded to the same questionnaire about a designated math course, were observed teaching, and kept logs of daily instructional activities in that course over a four week period. Mail responses were used to assess the adequacy and scope of items and response options; case study data were used to examine the reliability and validity of those same teachers' questionnaire responses. For reasons of space, this article discusses the case study fieldtest data only.

Building on previous work. The fieldtest built on the findings and recommendations from a previous NCES pilot project and on other earlier studies including the Third International Math and Science Study (1998, 1996), the UCLA/RAND Validating National Curriculum Indicators project (1995), and Reform Up Close (1993). The previous pilot project fieldtested a draft questionnaire with 111 eighth to tenth grade teachers in three districts (Mullens & Kasprzyk, 1996a). Results from that project and from subsequent experience with instructional practice items on the 1995 Teacher Followup Survey guided our questionnaire revisions and planning for this fieldtest. When refining items, we also built on the early TIMSS (1996) work developing items on IP and on Andrew Porter's (1993) work identifying effects of increased enrollments on the content and pedagogy of high school math and science courses. When designing the fieldtest, we drew heavily on the prior work of Burstein, McDonnell, et. al. (1995) developing validation procedures to improve the quality of national indicators of curriculum.

Fieldtest goal. The goal of the project was to collect information about the accuracy and reliability of self-reported data on the instructional practices of secondary math teachers and the contexts within which they occur. The items collected information on four areas of instructional practice: a) conditions for teaching and learning in the school and classroom, b) course content and emphasis, c) instructional activities, and d) the availability and use of instructional resources.

Fieldtest design. We conducted case studies during April and May 1997 in six geographic areas designed to attain some measure of dispersion yet limit travel costs: Baltimore City, Frederick, and Hagerstown, Maryland; Austin, Texas; Charleston, South Carolina; Milwaukee, Wisconsin; and Aberdeen, Bremerton, and Olympia, Washington. Fifty teachers identified one course (the "designated class") for which they were willing to be observed and to record classroom activities daily for four weeks. Together the courses covered the curriculum spectrum from eighth grade mathematics to Calculus. Forty-one of the 50 volunteers ultimately completed the case studies.

The case studies had five parts: a mail questionnaire, classroom observation, teacher interview, daily classroom logs, and a second administration of the questionnaire. At the beginning of the case study process, participating teachers completed the IP survey about the most recent semester. A researcher observed a class period in each teacher's designated class, recording on a log form the instructional objectives, classroom activities of the teacher and student, and the use and availability of instructional materials. Teachers completed a classroom log form about the same class and discussed the class and their questionnaire responses during a subsequent interview. Every day for four weeks, classroom teachers recorded their activities and those of their students. At

the conclusion of the case study period, teachers completed a second questionnaire, identical to the first.

Items on instructional techniques were the core of the questionnaire. Those items asked teachers to indicate the frequency and duration with which they used various instructional methods in a single targeted class. Activities included those commonly associated with traditional teaching (such as lecture and student recitation or drill), those reflecting reform recommendations (such as student discussions of problem solving approaches), and some common to a range of styles (such as giving tests). Other items asked teachers to describe their use of student activities, which were similarly distributed among instructional approaches.

The fieldtest had two limitations. First, while the questionnaire collected information covering a full semester of instruction, the design of the case study portion included data collection on only four weeks of that semester. Ideally, the two time periods would have been identical and we could have used a semester's worth of log data with which to validate questionnaire responses. The decision to collect only four weeks of log data reflected project funding limitations. Additionally, although 41 case study teachers completed four weeks of daily logs, only 20 completed the second questionnaire. Thus the analysis of teachers' responses on the two surveys was limited to those 20 sets. We think this low response was caused by the lateness in the school year. We have no reason to believe that the teachers who returned the second questionnaire were different from the non-responding teachers in some systematic way that might bias our interpretation of their responses.

Analysis of Fieldtest Data

At the conclusion of the case studies, we used the two questionnaires, the teacher logs, and the researchers' logs to investigate the reliability and validity of the questionnaire items. Among other analyses, we examined:

- percent teacher and researcher agreement on the occurrence of student learning objectives and instructional activities to understand the extent to which teachers and researchers shared a similar understanding of the concepts in question;
- percent teacher and researcher agreement on the length of time the objective or activity occurred to understand the extent to which teachers and researchers shared a similar conception of elapsed time;
- percent agreement between case study teachers' responses to the first and second questionnaires on the frequency and duration of classroom instructional practices to understand the extent to which survey responses completed six weeks apart are the same.

Fieldtest Results

With few exceptions, fieldtest data suggest that the case study teachers interpreted key words describing instructional processes in ways that were consistent with the independent observers. Teachers also had the same sense of the passage of time as observers when recording that information. Where low rates of agreement occurred, they reflected differences of opinion between teachers and observers about what constituted "whole class discussions," "practice or drill," and "several appropriate answers or approaches."

Determining the validity of teachers' daily descriptions of classroom instruction. To assess the accuracy with which teachers described on the log form the learning activities they orchestrate on a daily basis, we compared teachers' recordings of classroom activities on the daily log to the researcher's record on the observation form. If the items, teachers, and observers were each perfect, we could expect a 100 percent match on the occurrence and duration of all student learning objectives and instructional activities.

Items on student learning objectives. Independent observers validated 79 percent of teachers' recordings of the learning objectives used that day in their class, and agreement between teachers and observers was greater than 75 percent for four of seven objectives analyzed. The lowest agreements were for memorizing facts, definitions or formulae (66 percent), recognizing and solving story problems with unfamiliar or complex structures (71 percent), and building and revising theories (73 percent). Where nonagreement occurred, teachers were more likely than observers to report that a learning objective had been part of the observed lesson.

Teacher/observer agreement appeared to vary by the degree to which the objective was observable by a classroom visitor or was explicitly stated by the teacher to the class or to the observer. For example, it was usually clear to the observer when students were doing mathematical operations, but often difficult to observe

that students were memorizing or were expected to be memorizing. The learning objectives with the lowest rates of agreement, those generally less visible and more difficult to detect, may indeed have occurred but were simply not observed.

Teachers' estimates of the time spent on learning objectives were substantially verified by observers: teachers and observers strongly agreed on the minutes allocated toward the student learning objectives that occurred during that class period. In those instances where observers did not agree with teachers about the elapsed time, there was no clear pattern to the mismatches: teachers indicated either more or less time than the observers noted.

Items on teacher actions. Case study data show strong agreement between teachers and observers on the occurrence and duration of teachers' instructional activities. Teachers and observers agreed on 85 percent of all teacher activities occurring during all the lessons. In seven of the eight activities, agreement between teachers and observers about whether the activity occurred was 75 percent or greater. The highest rates of agreement between teachers and observers were for lecturing (98 percent) and providing individual or small group tutoring (95 percent); the lowest agreement was for stimulating student discussions of approaches to solving problems or explanations of their mathematical thinking (55 percent). Where there was nonagreement about an activity, teachers were more likely to report that it did happen than were observers. In 94 percent of the instances in which teachers and observers saw teacher activities differently, teachers indicated the activity had occurred and observers indicated they had not seen it.

We found a high level of agreement between teachers and observers on the minutes spent on each teacher instructional activity that occurred during the observed class period. Teachers and observers substantially agreed on the duration of all teacher activities except demonstrating a concept using two dimensional graphics.

Items on student activities. Teachers and observers agreed on 82 percent of all student activities recorded during the observed lessons. Agreement between teachers and observers on whether specific student learning activities occurred was 75 percent or greater for 13 of the 18 student activities included. High rates of agreement were recorded when students: listened to the teacher (100 percent), worked individually on exercises (93 percent), worked in small groups (93 percent), and worked on assignments due the next day (85 percent). Student activities with the lowest agreement between teacher and observer were the following: participate in whole-class discussion (56 percent), practice or drill on computational skills (63 percent), solve problems for which there are several appropriate answers or approaches (71 percent), and wait for completion of non-academic tasks (71 percent).

The low agreement rates for these activities reflect the gist of discussions following the observed classes in which teachers and observers reported differences of opinions on what constituted the first three activities. The majority of all nonagreements between teachers and observers on these items arose because the teacher saw the event as occurring but the observer did not: teachers indicated that student discussions involved the whole class, while observers were more likely to say that only a few students were actively involved; teachers thought students were drilling on basic skills, but observers saw no evidence; teachers more often said after class that they emphasized several approaches to a problem, while researchers observed only one.

There was strong agreement between teachers and observers on the length of time each student activity occurred during the observed class, ranging from 86 to 100 percent agreement per student activity. In the few instances where there was no agreement, there was also no pattern in the direction of nonagreement: observers used a clock or watch to record time as the activities occurred; teachers retrospectively over- and under-estimated elapsed time nearly equally.

Summary. Case study teachers' accounts of the student learning objectives, teachers' actions, and student activities occurring in the teachers' observed classes were substantially validated by the accounts of classroom observers on 24 of 33 items. Teachers' accounts of the length of time that student learning objectives were taught and that teachers and students engaged in activities were both substantially validated by independent observers on every item. Across the three types of items, teachers' time estimates were most accurate on those activities they used most frequently.

Determining the reliability of teachers' questionnaire responses. To assess the reliability with which teachers describe on a one-time questionnaire what they do throughout a semester, we compared teachers responses on the first questionnaire to their responses on the second questionnaire administered six weeks later. We assumed that the two sets of responses would be identical if their

first responses were accurate, if their implementation of instructional practices had not changed, and their opinions about their teaching had remained the same.

Items on student learning objectives. All of the nine subitems collecting information on student learning objectives had rates of agreement within one response option between the first and second questionnaire higher than 78 percent. There was high agreement on the frequency with which teachers' instructional objective was to have students understand concepts, relationships, theorems (100 percent); perform mathematical operations or execute algorithms (95 percent); and solve story problems with familiar structures (90 percent). The learning objective building or revising theories had the lowest agreement (79 percent) and, except for collecting data (by observing, measuring, or counting), was also the least used instructional objective during the case studies, according to log records.

Items on teacher actions. Ten of the twelve subitems assessing the frequency with which teachers use certain instructional techniques showed correspondence between teachers' responses on the two questionnaires at rates above 75 percent. Interestingly, the two items with low rates of agreement between questionnaires, leading students in recitation and drills and teacher time spent working on administrative tasks, are both forms of teaching considered to be more traditional. Additionally, teachers' responses on the typical length of time spent per class period on each instructional activity all showed more than 75 percent agreement between the two questionnaires.

Items on student activities. Of the 24 subitems assessing the frequency with which teachers have students engage in particular learning activities, 22 showed high correspondence between teachers' responses on questionnaires 1 and 2. The two items with low rates of agreement were practice or drill on computational skills (67 percent) and solving problems with more than one appropriate solution (74 percent). All 24 time-per-typical-use items had agreements greater than 75 percent.

Summary. Case study teachers' responses on the two questionnaires were substantially the same on 41 of the 45 items describing the student learning objectives and instructional activities used in the teacher's designated class.

Conclusions

These results suggest that teachers in disparate locations recognize and accurately interpret the named classroom activities, except for some glaring exceptions. Respondents' indications about whether or not certain activities occur (and for how long) coincide with those of independent observers, for the most part. Teachers' questionnaire responses about the math instructional practices they use (and for how long) are pretty reliable within one response option: teachers respond the same way to most questions on questionnaires administered six weeks apart.

The good news from this analysis is that we are confident that teachers recognize and identify most instructional activities in ways similar to the observers; they readily acknowledge their use of recitation and drill and admit to working on administrative record keeping tasks while their students wait, even though those activities may be out of favor with school reform advocates. When completing questionnaires administered six weeks apart, we know that teachers' responses to questions about instructional activities are consistent.

We suspect that three reasons may have contributed to occasional low levels of agreement between teachers and observers. Some items may have low agreement between teacher and observers because classroom observation itself is limited in its capacity to capture certain elements of classroom instruction. This is particularly true for unobservable instructional objectives such as memorization. Validation of items may also have been affected by the differences of opinion on the scale of classroom activities and was most noticeable on items that distinguish among number of participants, such as "whole-class discussion." Differences of opinion contributed to limited validation in other ways as well, such as when teachers and observers disagreed on whether problems had "more than one approach."

Recommendations

These conclusions suggest considerations for future questionnaire and fieldtest designs that may confirm and further our understanding of the accuracy and reliability with which teachers respond to self-report surveys. First, this fieldtest provided the strong basis for refining those few instructional practice items where wording that appeared unambiguous in pilot testing was subject to varying interpretation in wider use. Through fieldtesting, we identified specific issues for those select items that can now be reworded. Second, the next generation of designs might include multiple and concurrent techniques to validate the accuracy of teachers' descriptions of their daily instruction, especially information on student learning objectives whose occurrence can not be visibly or aurally confirmed by independent observers. This may

entail active observation by researchers, triangulation by multiple observers, or teacher/observer/student interactions to augment first-person observation. Finally, future efforts to validate questionnaire items would be well-served to equalize case study design lengths with the item referent periods: as long as a semester (equaling the referent period of the items tested here) or as short as two weeks. Shorter referent periods (with corresponding validation periods) are likely to result in more accurate responses by teachers, allow instructional variation across a large number of participants, and provide reliable data with which to estimate response accuracy.

We used these findings to modify items, to reduce ambiguity in problematic items, to identify particularly reliable and valid items, and to recommend a strong module of instructional process items for the 1999-2000 Schools and Staffing Survey.

References

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

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.

Leighton, M., Mullens, J., Turnbull, B., Weiner, L., & Williams, A. (1995). Measuring instruction, curriculum content, and instructional resources: The status of recent work (NCES 1995-11). U.S. Department of Education. Washington, DC: NCES Working Paper.

Mayer, D. (1999). Measuring Instructional Practice: Can Policymakers Trust Survey Data? Educational Evaluation and Policy Analysis, 21(1), 29-45.

Mullens, J. (1995). Classroom instructional processes: A review of existing measurement approaches and their applicability for the Teacher Followup Survey. (NCES 1995-15). U.S. Department of Education. Washington, DC: NCES Working Paper.

Mullens, J. & Gayler, K. (1999). Measuring classroom instructional processes: Using survey and case study fieldtest results to improve item construction (NCES 1999-08). U.S. Department of Education. Washington, DC: NCES Working Paper.

Mullens, J. & Kasprzyk, D. (1996a). Using qualitative methods to validate quantitative survey instruments. In 1996 Proceedings of the Section on Survey Research Methods. Alexandria, VA: American Statistical Association, 638-643.

Mullens, J. & Kasprzyk, D. (1996b). The Schools and Staffing Survey: Recommendations for the future (NCES 1997-596). U.S. Department of Education. Washington, DC: NCES Working Paper.

Mullis, I., Martin, M., Beaton, A., Gonzalez, E., Kelly, D., & Smith, T. (1998). Mathematics and science achievement in the final year of secondary school: IEA's third international mathematics and science study. Boston, MA: Center for the study of testing, evaluation, and educational policy, Boston College.

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

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.

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

Schmidt, W. (1996). Indicators of opportunity to learn in the Third International Mathematics and Science Study: What is the impact of this OTL information on U.S. public schools? Paper prepared for AERA Annual Meeting.

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

Stodolsky, S. (1996). Should SASS measure instructional processes and teacher effectiveness? In J. Mullens & D. Kasprzyk (Eds.) The Schools and Staffing Survey: Recommendations for the future (NCES 1997-596). U.S. Department of Education. Washington, DC: NCES Working Paper.

Acknowledgment: The authors would like to thank Daniel P. Mayer of Mathematica Policy Research, Inc. for his comments on an earlier draft.