AMERICAN INSTITUTES FOR RESEARCH®

Methodological Critique
Of Current NCES
Survey Measures of Instructional Processes

Lawrence Lanahan
Marion Scotchmer
Mary McLaughlin

American Educational Research Association

2004 Annual Meeting

April 12–16, San Diego, Calif.

Methodological Critique of Current NCES Survey Measures of Instructional Processes

Lawrence Lanahan, Marion Scotchmer, and Mary McLaughlin
Education Statistics Services Institute, American Institutes for Research

April 2004

Introduction

Over the last twenty years, education researchers have developed increasingly reliable and accurate measurements of educational inputs—such as fiscal expenditures, teacher characteristics, and material resources—and educational outputs as indicated by student academic assessments (Burstein et al., 1995). However, valid and reliable measurements of the processes that lay in between inputs and outputs—teaching and learning in the classroom—have proven elusive. Moreover, current survey measures of instruction demonstrate only few and very weak relationships to student achievement growth.

Efforts are currently underway to improve survey measures of classroom instructional processes (IP). The National Center for Education Statistics (NCES) has directed resources over the past decade toward developing more reliable and valid IP measures. NCES' development of national-level IP measures represents an effort to shed light on the "black box" of the classroom that has remained mostly in the dark.

As part of this effort, NCES and the Education Statistics Services Institute (ESSI) created the Instructional Processes Research and Design (IPRD) program with the intention of initiating a radical leap in the conceptualization and measurement of the instructional processes inside the "black box." The effort is aimed at development of items that reliably predict student achievement growth. The IPRD team's initial steps toward improving IP measurement have taken two tracks: 1) developing a deep comprehensive knowledge base concerning all relevant aspects of instruction, learning, and measurement related to IP, and 2) evaluating the most current, "state-of-the-art" NCES IP survey measures. These activities will provide a solid foundation from which the team can develop measures that incorporate useful aspects of previous measurements while using the knowledge base to make necessary advances in the conceptualization and measurement of IP.

This paper will present the findings of the IPRD team's evaluation of NCES IP survey measures. The evaluation was conducted in the context of the program's greater aim of gaining a full understanding of the issues surrounding IP measurement and using that understanding to advance the field to the point where valid, reliable IP measurement, with significant links to variation in student achievement, is possible.

Background

Most current IP survey items are the result of a movement to measure students' "opportunity to learn" (OTL) that began three decades ago. Early international assessments of student achievement included measures of curriculum to ensure that students had been provided an equal opportunity to learn the material that they were being tested on in the standardized exams (McDonnell, 1995). Eventually the OTL measures were expanded to include not only the "intended curriculum," but also the "implemented curriculum" (Floden, 2000). This distinction accounted for the fact that individual teachers can teach the same curriculum in different ways. Indicators of the implemented curriculum moved beyond content coverage to measure teachers' classroom strategies and practices.

These measures of classroom teaching received renewed attention in the mid–1980s as the education world searched for a remedy to a perceived decline in student achievement (National Commission on Excellence in Education, 1983). Researchers and policymakers began to call for instructional reforms and standards, and national indicators allowed the creation of a "comprehensive portrait of the schooling process as it occurred in schools and classrooms" (McDonnell, 1995, p. 309). Furthermore, some saw these measures as a way to monitor the equity of students' opportunities to learn and to hold schools and teachers accountable for the instruction they provided (Porter, 1995).

By the mid–1990s, many researchers were using OTL measures to collect information on variation in instructional processes (Brewer & Stasz, 1996; Porter 1995). NCES included IP measures on national-level surveys in order to provide a description of what happens in U.S. classrooms in terms of the content taught, instructional methods employed, and the resources used during instruction (Brewer & Stasz, 1996). These items covered a wide range of the classroom experience, including but not limited to content coverage, classroom resources, pedagogical styles, time spent by students and teachers on various activities, and the types of assessments used by teachers.

Although they have been in use for over a decade, current NCES IP survey measures are still considered to be in their "infancy," suffering from numerous methodological problems related to validity, reliability, and generalizability (Brewer & Stasz, 1996, p. 8). Furthermore, the items are only intended to provide a general portrait of teaching and are therefore too broad to be sufficiently predictive of student achievement (Brewer & Stasz, 1996).

Adding to the difficulty in linking teacher practice to achievement is the relatively small proportion of the variance in student achievement gains attributable to schools and teachers. Seventy or eighty percent of variation in student achievement on standardized tests is attributable to characteristics of the student taking the test; only 20 or 30 percent can be linked to schools or teachers (Rowan, 2000; Scheerens & Bosker, 1997).[1] Any

[1] Some researchers, Sanders (2000) in particular, would argue that the proportion of variance dependent on a student's classroom assignment is significantly larger than 30 percent. Analyses by Rowan (2002), using advanced statistical techniques currently being developed by Stephen Raudenbush, indicate that teacher effects have a larger effect on growth in student achievement than previously believed. However, different analysis techniques generate different effect sizes. Use of student achievement status results in smaller effects than use of achievement gains or cross-classified random effects (Rowan, 2002). For our purposes

attempt to find a significant link between student achievement and one of the many indicators of teaching practice comprising that small proportion of variance will be difficult.

Despite this difficulty, NCES still views as a worthwhile endeavor the attempt to develop methodologically sound measures of IP that explain as much of the variance in student achievement as possible. Rowan (2001) states, “there is a continuing role for survey research in the study of instructional effects on student achievement” (p. 25). Brewer and Stasz, in an NCES publication suggesting strategies for enhancing OTL measures, claim that despite the difficulties involved, “it is probably premature and overly pessimistic to abandon this line of research altogether” (1996, p. 5).

Problems in IP Measurement

The current track of item development is insufficient if NCES’s goal is to create measures that will pick up the small relationship between instruction and achievement. Before critiquing specific extant NCES IP items, it will be instructive to review the problems with IP measurement in general that have been documented in the literature in the areas of reliability, validity, and generalizability.

Reliability

To be reliable, measures of a particular construct must generate the same responses upon repeated administrations. This can be assessed in several ways. “Test-retest” reliability can be assessed by having a subject respond to a survey item and respond to the same item again at a later point in time. Entire instruments or batteries of items, such as a student assessment, can also be tested this way. Reliable items and instruments would generate the same responses regardless of the amount of time between administrations. By assessing inter-rater reliability, the consistency of observations can also be tested. This is accomplished by comparing scores given by different coders for an observation of the same behavior or phenomenon. Reliability of a measurement of a construct can also be assessed by comparing data collected by different methods. For example, a teacher may record data about her or his practice in a daily log and in a survey questionnaire. Responses from the two instruments should correlate if they are reliable. Discussions of reliability of IP measurement in this paper refer to the ability of measures of instructional processes to elicit consistent responses from teachers and students.

Many difficulties can arise in an attempt to generate reliable measures of an activity as expansive and complex as teaching. Teachers may not be able to recall their practice as far back as researchers would prefer, and hence might not give the same answer every time for a question about that practice (Ball, Camburn, Correnti, Phelps, & Wallace, 1999; Mayer, 1999a; Mullens & Gayler, 1999). Furthermore, teachers may not share the same vocabulary for describing their practice with survey researchers. If teachers are

---

it is more appropriate to use achievement status, as this allows one to compare students’ achievement status to their overall prior knowledge, not to the gains other students have made.

unsure of the type of activity or strategy an item is referencing, they may not answer the item consistently (Ball et al., 1999; Mayer, 1999a).

Studies investigating the reliability of IP survey measures have been mostly discouraging. A study commissioned by NCES to assess the reliability of IP measures from the 1994–95 Teacher Follow-Up Survey (TFS) found that every single one of the 22 IP items on the survey had reliability problems (Henke, Chen, & Goldman, 1999). Daniel Mayer (1999b) investigated measures of the frequency and duration of standards-related mathematics teacher practices such as “use manipulatives,” “teacher-led discussions,” and “work on individual projects” and obtained similarly dismal results for the reliability of individual items. The same questionnaire was administrated to teachers at two separate points in time. When Mayer (1999b) assessed the test-retest reliability of the items, only 3 of 17 items asking teachers to report the frequency and amount of time spent on specific practices were correlated at above .60, and 9 items were correlated below .30 (p. 38). Furthermore, although NCES (1997) reported that 73.3 percent of U.S. public school teachers used manipulatives or models once a week in 1994 and 1995, Mayer found that the correlation between survey reports of this practice at the beginning and end of the semester was only .44 (1999b, p. 43). Consequently, Mayer recommended against reporting frequencies of individual teacher practice items. When composites (groups of similar items) representing standards-related practice were compared, Mayer found a more encouraging test-retest correlation of .69, and the correlation between teacher self-reports and classroom observation was .85 (1999b, p. 38).

Two studies of IP measures, Smithson and Porter’s (1994) investigation of the accuracy of a survey designed to measure secondary school mathematics and science curriculum and a validation study of national curriculum indicators by Burstein et al. (1995), found more encouraging results concerning the reliability of IP survey measures. Burstein et al. found relatively high agreement on IP items between surveys conducted at the beginning and end of the semester: 60 percent of responses on IP items were exactly the same, and 90 percent were within one category. Both Smithson and Porter and Burstein et al. found what they claimed to be acceptable rates of agreement between surveys and logs on aspects of teacher practice.

Mayer, however, concluded that neither of these studies provided sufficient evidence to claim that their surveys were reliable or unreliable (1999a). Both studies compared teachers’ self-reports on surveys to their entries in a log. Mayer believes that independent observation is necessary to assess reliability, rather than comparing two different versions of self-reports from the same subjects. Furthermore, despite encouraging results using composite measures, Mayer concluded that more research would be necessary to improve reliability of both individual and composite measures of teacher practice (1999b).

The evidence suggests that reliability is still a major problem in large-scale surveys of instructional processes. Reliability is only part of the equation, however. One can develop items that respondents will likely fill out exactly the same every time, but if the items are not measuring what they were intended to measure, their reliability does not matter. The next section of this paper will address this issue by reporting on current knowledge about the validity of extant IP survey measures.

Internal Validity

To be useful measures of classroom processes, survey items must measure constructs as accurately and completely as possible. Even if an item's reliability is assured, it is still possible that the item is not measuring what it purports to measure. Internal validity is the extent to which a measure represents the qualities of the concepts or phenomena it is claimed to be measuring within a given sample. Internally valid measures of instructional processes should provide an accurate portrait of what happens in the classrooms under study as teachers are teaching and students are learning.

In a 1995 report detailing a validation study of national curriculum indicators, Burstein et al. noted that the validity of national-level IP measures had not yet been established:

> …failure to validate these indicators has remained a problem (p. xiv).
>
> ...at this point, none of the national survey data collected from teachers have been validated to determine whether they measure what is actually occurring in classrooms. Despite major advances in the design of background and school process measures, studies have generally developed only a few new items and 'borrowed' others from earlier studies (p. 8).

At the heart of the accuracy issue is the ability of teachers to report their own practice on survey questionnaires. To achieve internal validity, teachers must interpret a survey question as it was intended to be interpreted by the researchers, the response options provided must be organized in a way that makes sense to the teacher, and the teacher must answer the question honestly. When the intended meaning of a question does not match the meaning taken by the teacher, internal validity suffers. Items that are unclear, overly inclusive, or contain multiple meanings have poor validity; teachers may misinterpret the language and be, in actuality, answering a different question than that the researchers asked (Mullens, 1995; Mullens, 1998; Stecher & Foley, 2000). Furthermore, teachers in schools implementing instructional reforms may feel pressure to answer questions about their practice in a socially desirable and not necessarily honest fashion (Mayer, 1999a; Mullens, 1998). Also, teachers often honestly believe they have changed to a reform-oriented style, when in actuality they have merely incorporated reform elements into their original style (Cohen, 1990). This misunderstanding can undermine the internal validity of teacher self-reports as well. These examples illustrate that internal validity and reliability problems share similar roots. Both are compromised by teachers' problems with long-term recall and the language used to describe instruction.

Surveys that do not use a consistent set of response options throughout the questionnaire also tend to have internal validity problems (Mullens, 1998). For instance, a survey that alternated between "never/sometimes/always" and "everyday/once a week/two or three times a month/once a month/never" would not have as much internal validity as a survey that consistently used one or the other. Also, teachers are often unable to recall their practice over a long period of time, so items that ask teachers to recall practice over a year or semester have questionable internal validity (Mullens, 1998).

Internal validity can be tested and improved by employing alternative data collection methods to check the veracity of teacher reports. Other data collection methods can be used to validate survey items. However, problems exist here as well. Burstein et al. (1995) argue that logs cannot be used to validate surveys because both suffer from the same dangers of self-reporting. Observations can be used to validate teacher surveys but are more burdensome. Mayer (1999b) attempted to validate a teacher survey in this manner but was unable to determine how many observations were necessary to develop an accurate picture of a teaching style.

Obviously, validating national measures of instructional processes needs to be made a priority. The accuracy and scope of current measures is questionable. However, ensuring the accuracy and comprehensiveness of items is not enough for the purposes of this project. Measures must also be able to link teacher practice to changes in students' academic achievement.

Predictive Validity

Another way to validate survey items is to evaluate them based on their ability to predict some outcome that, based on the theory behind the development of the original item, should be related to the construct being measured. Classroom research assumes that teachers' instructional knowledge, strategies, and practices are deployed with the intent of their increasing their students' academic abilities. Therefore, the predictive validity of an item measuring teaching can be evaluated on its ability to predict changes in student learning.

Current survey measures of instructional processes could provide limited descriptive information on some aspects of teaching in U.S. classrooms if their validity and reliability were to be established. This data could provide the public with an opaque statistical glimpse into the classroom or inform policymakers of the extent to which teachers have adopted recommended instructional standards (Blank, Porter, & Smithson, 2001; Burstein et al., 1995). However, researchers have not yet found an adequate instrument to measure teaching in a way that convincingly links classroom instructional processes to student achievement outcomes (Brewer & Stasz, 1996).[2] Current IP measures are too broad to be sufficiently predictive of student achievement (Brewer & Stasz, 1996). This lack of predictive validity is a major obstacle to any attempt to assess the impact of classroom reforms on student achievement at a national level.

The absence of highly predictive IP measures is rooted in the failure of most attempts to empirically demonstrate "what works" in classroom instruction (Brewer & Stasz, 1996; Mayer, 1999a). Mayer (1999a) suggests that an overreliance on input-output studies has hindered development of more predictive measures. Several lines of input-output research, such as "process/product," "effective schools," and "educational productivity," have attempted to uncover an empirical relationship between aspects of the schooling process as inputs and variance in student achievement as an output. Research

[2] Rowan (2001) found significant overall teacher effects; however, when he examined the impact of particular instructional processes, links were extremely small and tainted by questions of reliability in the items used to collect the data.

has identified small relationships between student achievement and certain teacher characteristics, such as length of teaching experience and academic skill (Mayer, Mullens, & Moore, 2000). However, attempts to link actual classroom processes to student outcomes have been largely unsuccessful (Brewer & Stasz, 1996).

This lack of an empirical knowledge base about effective instructional processes has kept researchers from developing adequate IP measures. Most of what the research community knows about effective instructional practice comes from in-depth case studies of a small number of classrooms. Researchers have had difficulty generalizing the findings from these studies to surveys (Brewer & Stasz, 1996; Mayer, 1999a). As a result, most IP survey development has resulted in very broad measures of general pedagogical behavior based on instructional standards developed by subject-specific teacher associations (e.g., National Council of Teachers of English, 1996; National Council of Teachers of Mathematics, 1991) and process/product research (Brewer & Stasz, 1996; Burstein et al., 1995).

To establish validity in IP measures, it will be necessary to develop measures that reveal the quality of instruction (Mayer, 1999b). Most current instructional practice items measure only the frequency—and sometimes duration—of certain practices or behaviors. However, frequency and duration items do not measure whether the practices under consideration are used appropriately (Baker, LeTendre, & Zhang, 1999; Mayer, 1999b). Measures of the quality of instruction will necessitate moving beyond broad measures of teacher behavior, and they will necessitate moving beyond instructional standards and process-product research as foundations for item development.

Developing measures of the quality of instruction is not an easy task. The process of initiating gains in student achievement is too complex to measure completely with simple measures of the frequency of particular practices. Effective instruction involves fluid teacher responses to student needs. Teachers draw on a wide range of instructional techniques within the classroom context, none of which can be singled out as more effective than another (Baker et al., 1999). Furthermore, teaching strategies change with the content that is being taught. Current IP measures have been criticized for separating strategy from content (Stecher & Foley, 2000). Lastly, although there is usually just one teacher in a given classroom, there are many students, all of whom play a role in the link between instruction and learning. The current focus in IP measurement on teacher behavior in terms described in instructional standards ignores the important role that students play in classroom learning (Brewer & Stasz, 1996; Burstein et al., 1995).

Though difficult, this advancement is necessary if researchers are to achieve predictive validity in the measurement of instruction. New IP measures must move beyond teacher-centered frequency and duration items and begin capturing the quality of instruction. This entails sufficiently accounting for academic content, classroom context, and the role of students in the link between instruction and learning, for it is in the matching of technique to student, content, and occasion that effective teachers excel.

Despite the demonstrated difficulties in using survey items to measure instructional processes, Brewer and Stasz, who evaluated NCES's IP measures, recommended that efforts continue to develop IP items with better predictive validity:

> While these problems associated with answering the question "what promotes student learning?" have led to few strong conclusions, it is

> probably premature and overly pessimistic to abandon this line of research altogether (1996, p. 3–5).

More recent research reprises this recommendation (Rowan, 2001). Linking instruction to student learning will require addressing issues of instructional quality, as well as the involvement of students in the learning process (Brewer & Stasz, 1996; Burstein et al., 1995; Mayer, 1999b).

External Validity (Generalizability)

Once the reliability and internal validity of survey measures have been established, one must establish external validity, or generalizability. In other words, one must ensure that the findings for the chosen sample apply to the larger population and social context under study. Generalizability for national-level IP survey measures involves more than demonstrating that the findings for a sample of teachers applies to the entire U.S. teaching force. Findings related to a teacher’s instruction for a certain period of time, such as a lesson, unit, or semester, must also apply to her or his instruction throughout the course of an entire year.

The first dimension of external validity, generalizing from a sample to a population, does not pose any obstacles to IP items as they currently exist on NCES surveys. On these surveys, the teacher is usually the unit of analysis, and the items measure instances of an individual teacher’s behavior. The samples that NCES uses are representative of the U.S. teacher population, and national estimates of teacher practice can be computed from NCES IP survey items.

It is in the attempt to measure the quality of instruction that researchers encounter difficulty achieving external validity. In developing measures of instructional processes, researchers encounter a delicate tension when attempting to reconcile internal and external validity. Teachers have difficulty accurately recalling their practice over long periods of time, such as semesters (Mayer, 1999a; Mullens & Gayler, 1999). This would suggest that to achieve internal validity and reliability, one would ask teachers about their most recent practice. However, effective teachers’ instruction is dependent on the subject matter, classroom environment, and other contextual factors. It changes so much from day to day that it would be difficult to generalize from a survey item inquiring about the past week, unit, or lesson to a teacher’s instruction over a semester over a year (Ball et al., 1999). This appears to be a zero-sum game. Detailed measures of a teacher’s recent instruction gain predictive validity but cannot be generalized to her or his instruction throughout the year. Items that inquire about practice over a semester or year are more likely to capture teachers’ typical practice but are too broad to predict student achievement.

This tension between external and internal validity is why most of what the research world knows about effective instruction comes from in-depth case studies of small samples of classrooms, and not from large-scale surveys (Mayer, 1999a). By surveying literature on previous case studies, Leighton, Mullens, Turnball, Weiner, and Williams (1994) identified many aspects of instruction that were predictive of gains in student achievement, such as questioning strategies and reactions to student responses. However, these studies of instruction were done with small samples outside of the context of an actual classroom lesson. Consequently, converting those findings into generalizable, internally valid IP survey measures has been unsuccessful.

One potential remedy for internal validity in IP measurement, a shift in focus from the teacher to the student, raises its own generalizability issues. This shift has been suggested because current measures have been criticized for ignoring the student's role in the instructional process (Burstein et al., 1995). Indeed, most current measures only capture instruction as it is distributed by the teacher, and not as it is received by the student. Instruction for one student may be different than that of another. This is important to consider if measures are to be linked to student achievement outcomes. After all, a teacher's instruction may have different impact on different students' outcomes. A shift in focus toward the student would entail a reframing of the generalizability issue. Measures would not only have to consider whether data collected on the instruction as delivered by the individual teacher was generalizable to all teachers in the population, but also whether that data on instruction was applicable to all of the students in the class. If most students are receiving essentially the same instruction, however, this ceases to be an issue.

Even if all of the listed problems with external validity were resolved, researchers would have difficulty generalizing findings from one particular grade level or subject matter area to another. Over the past twenty years, researchers have increasingly emphasized the close relationship between content and pedagogy (Shulman, 1987). Because of this, it is expected that, because content changes across grades even within the same subject area, effective pedagogy would be different across both subjects and grades.

It is essential to address the problem of generalizability/external validity if advances are to be made in the measurement of instruction. Long-term measures of instruction are necessary to link instructional processes with achievement. However, the generalizability of these measures are challenged by requirements for internal validity and reliability.

Implications for Future Measurement Development

Obviously, there are many obstacles along the path to reliable, valid, generalizable measures of instructional processes. Even if the conceptualization of instructional processes underlying current measures remained the same, much work would still be required before items could be considered technically valid or reliable (Mayer, 1999b). Establishing predictive and external validity for IP measurement, however, would entail a major reconceptualization of IP. The development of successful new measures would require not only capturing the quality of instruction over the course of a year in a way that is predictive of student achievement gains, but capturing it in a way that is generalizable to all teachers and students in the population under study.

This is not to say, however, that current measures are useless or better measures should not be pursued. Self administered surveys are still the most efficient way to collect data on a national-level. Current survey IP measures can, at the very least, provide a broad descriptive portrait of teaching when analyzed at the composite level (Floden, 2000; Mayer, 1999b). This type of information is useful for determining whether teaching is consistent with expectations of reformers and policymakers (Burstein et al., 1995).

Furthermore, state-of-the-art measures of content coverage have shown small links to student achievement (Brewer & Stasz, 1996; Floden, 2000; Gamoran, Porter, Smithson, & White, 1997; Rowan, 2001). Further steps should be taken to improve the state-of-the-art measures in order to capture instructional processes that are predictive of student achievement (Brewer & Stasz, 1996).

Evaluation of NCES measures

Now that the general issues concerning IP measurement have been presented, this paper will focus upon the survey measures that NCES is currently using to collect data on IP in U.S. classrooms. It is important to review current NCES items because they are still used in reports that provide portraits of instruction in U.S. schools. Furthermore, it will aid the IPRD team in its own quest to learn about the reliability and validity of IP items and create a more compelling portrait of instruction.

Briefly described below are seven major NCES surveys—HS&B, NELS, SASS, NAEP, ECLS, Prospects, and TIMSS—their structure, components, and IP measures. Following this description, the review of the instructional processes items found within these questionnaires will be presented in detail.

Background on NCES Surveys

High School and Beyond (HS&B) was conducted by NCES as part of the National Education Longitudinal Studies program and as a follow-up to the National Longitudinal Study of the High School Class of 1972 (NLS–72). HS&B surveyed two cohorts, the 1980 senior and sophomore class, every two years through 1986, with an additional survey of the sophomore class in 1992. Principals provided basic school information, including information about instructional resources (e.g., per pupil expenditure, class size, teacher credentials, etc.) and general school policies. The 1984 survey included a supplementary Administrator and Teacher Survey, conducted separately for principals, guidance counselors, and teachers. Teachers were asked about their control of curriculum content, classroom resources, and teaching techniques; the importance of general goals in teaching; the frequency of class interruptions; allocation of classroom time; and detailed questions about their background and qualifications. However, teachers were not asked about either particular students or classes.

The National Educational Longitudinal Study of 1988 (NELS:88) was the first survey to link teachers with a national sample of students and was designed to provide trend data about young people as they experience critical transitions such as attending school and entering the workforce. NELS:88 surveyed a 1988 eighth-grade cohort every two years until 1994, with an additional follow-up in 2000. Further, standardized test scores and transcripts were collected for the sample students. For the base year of 1988 and the 1990 and 1992 follow-ups, questionnaires were completed by parents, teachers, and school administrators.

Teacher surveys for NELS:88 collected detailed information on at least one subject-specific class, including the teacher's educational background, goals, and instructional practices; class size; school resources; and curriculum content. Both topic coverage and degree of emphasis on objectives in mathematical learning were included in the curriculum content items; the instructional practices questions asked about the percentage of time spent teaching versus non-instructional activities. Finally, items about teaching methods were aimed at identifying and counting teacher-center and student-center activities.

The Schools and Staffing Survey (SASS) provides a comprehensive survey of public and private K–12 schools and was conducted in 1987–88, 1990–91, 1993–94, and 1999–2000. In the most recent administration, over 100,000 questionnaires were distributed to teachers, principals, school libraries, schools, and school districts, for a sample size of approximately 72,000 teachers, 10,000 public schools, 3,500 private schools, and 6,000 school districts. SASS has contained four core components since its inception: the Teacher Questionnaire, the Principal Questionnaire, the School Questionnaire, and the School District Questionnaire. Among the areas covered by SASS in 1999–2000 were school safety; teacher training and experience; professional development; parental involvement; charter schools; migrant student enrollment and services; computers; and uses of school performance reports.

The National Assessment of Educational Progress (NAEP), or the Nation's Report Card, has been conducted since 1969 and is Congressionally mandated to assess student performance at national, regional, and state levels. NAEP is divided into three assessments: National NAEP, State NAEP, and Long-Term Trend NAEP. National and State NAEP are conducted every two years at grades 4, 8, and 12. Long-Term Trend NAEP, which assesses students at ages 9, 13, and 17, is currently conducted every four years, with the next assessment in 2003. In NAEP, math and reading are assessed every two years in NAEP, science and writing every four years, and other subjects, including art, civics, history, and geography, less frequently. In each NAEP assessment background information is collected from students, teachers, and school administrators. Background information on students included information on student characteristics, eligibility of free or reduced-price lunches, participation in Title I programs, presence of a disability, and limited English proficiency status. Background information on teachers and schools included classroom practices, teacher training, availability of computers, and parent participation.

Begun in 1998, the Early Childhood Longitudinal Study (ECLS), is cosponsored by NCES and other federal agencies, including the National Center for Health Statistics, and other organizations with the goals of providing information about the health, care, and educational experiences of young children. The study follows two cohorts for six years: the kindergarten cohort, begun in 1998, and the birth cohort, begun in 2001. ECLS will provide data to test hypotheses about the influence of variables, including those of family, schooling, and community variables.

Congressionally mandated as part of the reauthorization of Title I in 1988, the Prospects survey was designed to permit comparisons of students in Title I programs with those not. Prospects was a nationally representative longitudinal study of students in grades 1, 3, and 7 and was originally conducted in 1991, with follow-ups annually through 1993 for grade 7 and 1994 for grades 1 and 3. As part of the survey, background information was collected on students, parents, classrooms, and schools, with questionnaires going to parents, teachers, principals, and the school district. The teacher questionnaires included items on instructional processes, covering resources, content, and instructional pedagogy.

Conducted in 1995, with a follow-up in 1999, Third International Mathematics and Science Study (TIMSS) assessed nationally representative sampled students at fourth grade, eighth grade, and at the end of secondary school in mathematics and science. Sponsored by the International Association for the Evaluation of Education Achievement (IEA), forty-two countries participated in the initial administration. Additional TIMSS-R assessments are planned on a four-year cycle. TIMSS included five components: assessments, questionnaires, a videotape study, a curriculum analysis study, and case studies for policy topics.

TIMSS teacher questionnaires collected data on six areas relating to instructional processes, as identified by Baker et al. (1999). These areas included information on teacher background, the basic make-up of the classroom tested in TIMSS, taught topics (in two ways), average lesson taught, and pedagogical approach. Opportunity to learn (OTL) was measured in two ways, and the items on average lesson taught formed the crux of the TIMSS IP items. TIMSS utilized multiple approaches and measures of IP-related topics, including examining teacher performance in terms of use of teaching methods, range of such methods, and the relationship between teachers' understanding and practice.

Methodology

The review of the literature describing reliability and validity issues in IP measurement offers a set of criteria for assessing the strengths and weaknesses of current and recent IP survey items. In turn, reviewing current and recent efforts to measure IP can provide insight into further item development.

The centerpiece of the review is a Microsoft Access database created by IP staff in 1998 that contains nearly all of the survey items that NCES has used on a wide range of surveys to measure instructional processes.[3] The measures in the "item bank" were categorized by a scheme developed by Porter and Smithson (1997) that suited the purposes of this study. Their model organizes the instructional process, or in their terms, the "enacted curriculum," into six categories: instructional content, instructional pedagogy, instructional resources, assessment, homework, and instructional time. Each of these categories is divided into subtopics.

[3] Questionnaires not included in the item bank—NAEP, HS&B, TIMSS-Repeat, and the TIMSS "Opportunity to Learn" (OTL) questionnaire—were reviewed, and items were included for analysis when they differed significantly from the items already present in the item bank. None of the NAEP or HS&B IP items differed significantly enough to merit inclusion. Several TIMSS-Repeat and TIMSS OTL items were included; these items are noted in the text of the actual review. The Fast Response Survey System (FRSS) has also periodically conducted surveys of teachers. Three recent surveys asked teachers about their preparations and qualifications, use of computers and the Internet, and opinions on education reform. However, these surveys are designed to quickly collect information on a topic of interest. Instruction was not the main topic for any of these surveys, so they were excluded from the study.

We assigned each item to one or more subtopics, based on Porter and Smithson's taxonomy. When an item stem was broken into subitems, each subitem was assigned its own category. (For the purposes of this paper, each subitem unit will be referred to as an "item.") A count of the items within each of the six categories is presented in Table 1.

Table 1.—Frequency of appearance of items from Porter and Smithson's instructional categories on NCES surveys.

Porter and Smithson Category	Number of appearances
Instructional Content	982
Instructional Pedagogy	486
Instructional Resources	224
Assessment	164
Homework	147
Instructional Time	57

NOTE: These categories are not mutually exclusive. One item can belong to many categories and appear within one category multiple times as a member of different subcategories.

The item bank was too large—1876 items in all—to justify a review all of the items. We had to find some way to sample the items. Our analysis focused on the subcategories most often measured in NCES surveys. We identified the nine most frequently measured subcategories for analysis. Unfortunately, only five of the six categories were represented within these nine subcategories; none of the "instructional time" subcategories was represented. To remedy this, we replaced the items from the ninth most frequently measured subcategory with all of the items from the "instructional time" category so that all areas of the enacted curriculum would be represented. Ultimately, approximately three-quarters of all items were included in these categories and subsequently reviewed. Table 2 presents the categories selected for analysis and counts of the items within those categories.

Table 2.—Frequency of appearance of items from Porter and Smithson's instructional subcategories on NCES surveys.

Subcategory	Number of items in subcategory
Instructional Content What Topics are Covered	804
Instructional Content Student Cognition	162
Instructional Pedagogy Grouping Strategies	119
Instructional Pedagogy Student Activities	93
Instructional Resources Computers	70

Assessment Types Used	68
Assessment How Used	58
Instructional Time	57
Homework How Used	56

Items in each chosen subcategory were reviewed in the context of the IPRD project's goals and the IP measurement problems found in the literature and outlined in the previous section of this paper. In the presentation of the analysis, attention is given to typical items and items which stood out as "exceptions to the rule." Examples of both are provided throughout the critique.

Critique of Current NCES Items

Instructional Content: What Topics Are Covered

With 804 items, "What topics are covered" has the highest frequency of all categories in the IP databank. The items come from six surveys: ECLS Spring Administration, Prospects Teacher, NELS Teacher, TIMSS Teacher Population 1, TIMSS Teacher Population 2 Math, and TIMSS Teacher Population 2 Science.

Most items in this category ask the respondent whether and how often subjects were taught in their class, with a list of subjects or skills provided. For all questions, a broad individual category (e.g. Language Arts) is defined in the question stem, and the subjects or skills listed are components of that broad category. Response options for these items are mainly for the frequency with which the listed subjects or skills were taught. ECLS asks how often in terms of periods of time—daily, weekly, monthly. TIMSS asks about frequency in terms of how many lessons were taught, while NELS asks only if the subject was taught as new content or reviewed. The Prospects survey is an exception and calls for the respondent to assign percentages for each subject listed, with percentages totaling 100. Nearly all items had a referent period of the entire year and all respondents were teachers.

The following item (Table 3), from the ECLS Kindergarten Teacher Questionnaire (Spring) Part A, is typical of items in this category.

Table 3.—Typical "Topics Covered" item

29. For this school year as a whole, please indicate how each of the following **READING** and **LANGUAGE ARTS** skills is taught in your class(es)?
    NOT TAUGHT: Taught at a higher grade level, Children should already know;
    TAUGHT: One a month or less, 2–3 times a month, 1–2 times a week, 3–4 times a week, Daily

    a. Conventions of print (left to right orientation, book holding)
    b. Alphabet and letter recognition

c. Matching letters to sounds d. Writing own name (first and last) e. Rhyming words and word families f. Reading multi-syllable words, like adventure g. Common prepositions such as over and under, up and down h. Identifying the main idea and parts of a story i. Making predictions based on text j. Using context cues for comprehension k. Communicating complete ideas orally l. Remembering and following directions that include a series of actions m. Using capitalization and punctuation n. Composing and writing complete sentences o. Composing and writing stories with an understandable beginning, middle, and end p. Conventional spelling q. Vocabulary r. Alphabetizing s. Reading aloud frequently

SOURCE: ECLS Kindergarten Spring Teacher questionnaire

Some TIMSS items were the exception to this general trend, asking respondents to identify whether listed topics were part of their most recent lesson taught. The following example (Table 4) comes from the TIMSS Teacher Questionnaire (Mathematics) Population 2.

Table 4.—Item on topics taught in most previous lesson

13b. For each of the following mathematics topics, indicate whether or not it was the subject of the lesson. 1. Whole Numbers 2. Common and Decimal Fractions 3. Percentages 4. Number Sets and Concepts 5. Number Theory 6. Estimation and Number Sense 7. Measurement Units and Processes 8. Estimation and Error of Measurements 9. Perimeter, Area and Volume 10. Basics of One and Two Dimensional Geometry 11. Geometric Congruence and Similarity 12. Geometric Transformations and Symmetry 13. Constructions and Three Dimensional Geometry 14. Ratio and Proportion 15. Proportionality: Slope, Trigonometry and Interpolation 16. Functions, Relations, and Patterns 17. Equations, Inequalities, and Formulas 18. Statistics and Data 19. Probability and Uncertainty

20. Sets and Logic
21. Problem Solving Strategies
22. Other Mathematics Content

SOURCE: TIMSS Teacher Population 2 Mathematics questionnaire, 1995.

ECLS, conducted in 1998–9, and TIMSS, conducted in 1995, are the two most recent surveys in this sample and represent the largest number of items contained in this category. Of the 804 total items in this category, the three TIMSS questionnaires (Teacher Population 1, Teacher Population 2: Mathematics, and Teacher Population 2: Science) account for 648, or over eighty percent; eighty-eight of the items come from the ECLS Spring Administration. The concentration of content questions in TIMSS questionnaires is not unexpected, because TIMSS is the most recent in a string of student achievement studies to measure students' opportunity to learn. To ensure fairness on student assessments, researchers have included items on content coverage to ensure that students had the opportunity to learn the content being tested (McDonnell 1995). The questions themselves tend to cover a wide range of subject-appropriate content and are fairly lengthy and detailed. Below is an example (Table 5) from the Teacher Population 1 survey.

Table 5.—TIMSS content coverage item

How long did you spend teaching each of these topic areas to your class this year? Will you cover any of these topics in future lessons? Have taught this year: lessons completed: 1–5, 6–10, 11–15, > 15; will teach later this year; not taught the year; taught previous year

a) Whole Numbers

1. Place value and numeration
2. Whole number meanings, operations, and properties

b) Common and Decimal Fractions

1. Meaning, Representation and Uses of Decimal Fractions
2. Operations of Decimal Fractions
3. Properties of Decimal Fractions
4. Meaning, Representation and Uses of Common Fractions
5. Operations of Common Fractions
6. Properties of Common Fractions
7. Relationships Between Common and Decimal Fractions
8. Finding Equivalent Fractions & Forms
9. Ordering of Fractions (Common and Decimals)

c) Percentages

Concepts of percentage; computations with percentage; types of percentage problems

d) Number Sets and Concepts

Integers (negative as well as positive); rational, real, and other number sets; number bases other than ten; exponents

e) Number Theory

Prime numbers; factors of whole numbers; greatest common divisors; least common multiples; permutations; combinations; systematic counting

f) Estimation and Number Sense

Estimating quantity and size; rounding and significant figures, estimating the results of computations (including mental arithmetic and deciding if solutions are reasonable); scientific notation

g) Measurement Units and Processes

Ideas of measurement and units; standard (metric) units; length, area, volume, capacity, time, money and so on; use of measurement instruments

h) Estimation and Error of Measurements

Estimation of measurements other than perimeter and area; precision, accuracy, and errors of measurement

i) Perimeter, Area and Volume

Perimeter and area of triangles, quadrilaterals, circles and other two-dimensional shapes; calculating, estimating, and solving problems involving perimeters and areas; surface area and volume

j) Basics of One and Two Dimensional Geometry

Number lines and graphs in two dimensions; triangles, quadrilaterals, other polygons, and circles; equations of straight lines; Pythagorean Theorem

k) Congruence and Similarity

Concepts, properties and uses of congruent and similar figures, especially for triangles, squares, rectangles, and other plane shapes

l) Transformations and Symmetry

Patterns; tessellations; symmetry in geometric figures; symmetry of number patterns; transformations and their properties

m) Three Dimensional Figures and Constructions

Constructions with compass and straight edge; three-dimensional geometry; conic sections

n) Ratio and Proportion

1. Concepts and meaning

2. Applications and uses

Maps and models; solving practical problems based on proportionality; solving proportional equations

o) Functions, Relations, and Patterns

Number patterns; properties, uses, and graphs of functions; problems involving functions relations and their properties;

p) Equations and Formulas

1. Linear equations and formulas

Representing linear numerical situations; solving simple linear equations

2. Representing other numerical situations solving other simple equations; use of algebraic expressions and inequalities

q) Data Representation and Statistics

Collecting data from experiments and simple surveys; representing and interpreting data (tables, charts, plots, and graphs); means, medians and other simple statistics; samples; uses and misuses of simple statistics

r) Probability

Concepts of “more likely” and “less likely”; computing probabilities (including informal computation or estimation of probabilities)

s) Sets and Logic

Sets, set notation and set operations; classification; logic and truth tables

t) Problem Solving Strategies

Problem solving heuristics and strategies

u) Other Mathematics Content

Mark here for all content not in one of the earlier categories. This includes advanced topics such as: Computers (flow charts, programming languages, programs); history and nature of mathematics; proofs and advanced mathematics topics.

SOURCE: TIMSS Teacher Population 1 questionnaire, 1995.

Content questions from other surveys tend to be shorter and less comprehensive than those found in TIMSS; however, they take the same general approach of listing topics within a specified subject and asking teachers to identify the frequency or percentage of their time they have spent teaching the topics. All surveys ask about fine-grained and well-defined topics. The surveys, in sum, do a thorough job of covering and collecting information on content.

Critique. Items in this category have some of the problems related to assessing IP discussed earlier in this paper. All surveys include recall periods of the entire school year, with the exception of the TIMSS questions addressing the most recent lesson taught. These recall periods pose problems because, as mentioned before, teachers may not be able to recall their instruction over an entire semester, and results from items covering one lesson cannot be generalized to the body of a teacher's instruction over the course of a year.

Two approaches were taken to describe the nature of the presentation of content. ECLS and TIMSS ask about frequency of content, and NELS and Prospects about the emphasis placed on the content. Neither of these approaches can establish how the content was taught, how effective the teaching was, or, in general, the quality of the instruction. In addition, all surveys in this category were teacher surveys, and thus focus on teacher activities, not on student engagement. They are therefore unable to capture how students are receiving the instruction and how they are interacting with their teachers.

The exceptions in this category are those TIMSS items examining the most recent lesson taught. The questions ask many details about the lesson, including the topic, in what part of the larger unit it was in, homework assignments, lesson activities and order, uses of grouping, and teacher and student interaction on incorrect answers. These questions, unlike the majority of those in this category, integrate content with activities and address teacher and student interaction. However, there is still no measure of the effectiveness or quality of the teacher. Furthermore, because the questions deal with only one lesson, the generalizability of the results to a teacher's larger body of instruction is compromised. The TIMSS "most recent lesson" items, as well as another innovative and unclassifiable TIMSS item on teachers' pedagogical knowledge, will be covered in more depth in a later section.

Instructional Content: Student cognition

One hundred sixty-two items from the IP databank were classified as "student cognition," making it the second largest in the IP survey. Five questionnaires were represented in this category: ECLS Spring Administration, Prospects Teacher, NELS Student, NELS Teacher, and SASS Teacher Follow-Up.

Since classification of items is not limited to one category, many questions classified as "What topics are covered" are also listed in "Student cognition." Many ECLS questions, for example, are assigned to both categories in the item bank. The difference between the two categories is that while coverage is concerned only with the taught curriculum, student cognition attempts to measure the skills and objectives teachers try to convey to their students. Student cognition is by definition more intangible and, in general, is not the main focus of the surveys included in the item bank. Although student cognition represents the second largest number of items, its 162 total is well below that of content's 804 and is more evenly dispersed across the surveys. ECLS accounts for the largest number of cognition items; TIMSS, which had far and away the largest number of questions classified as content, has no items classified as student cognition.

Most of the items refer to a full year, although the SASS Teacher Follow-Up asks about the most recent semester. All but the NELS Student survey had teacher respondents. Most, with the exception of ECLS and TIMSS, referred to only one class rather than all the respondent's classes. (NELS Teacher asks separately about all of the teacher's classes.) ECLS and SASS ask how often teachers teach or emphasize listed subjects or skills in their classes. Both NELS surveys ask how much emphasis the teacher places on listed objectives. Response options for ECLS and SASS ask how often in terms of periods of time—daily, weekly, monthly. The NELS surveys give four response options for degree of emphasis—"major," "moderate," "minor," and "none." The following example (Table 6), from the NELS Teacher Questionnaire, illustrates this type of question.

Table 6.—NELS item illustrating typical response options

2_14. In this math class, how much emphasis do you give to each of the following objectives? Major, moderate, minor, none.

a. Understanding the nature of proofs
b. Memorizing facts, rules, and steps
c. Learning to represent problem structures in multiple ways (e.g., graphically, algebraically, numerically, etc.)
d. Integrating different branches of mathematics (e.g., geometry, algebra) into a unified framework
e. Conceiving and analyzing effectiveness of multiple approaches to problem solving
f. Performing calculations with speed and accuracy
g. Showing importance of math in daily life
h. Solving equations
i. Raising questions and formulating conjectures
j. Increasing students' interest in math

SOURCE: NELS Teacher questionnaire, 1992.

In contrast, the Prospects Teacher survey asks what percentage of instruction is conducted in the following modes, including "learning mathematics facts and concepts, learning skills and procedures needed to solve word problems, and develop reasoning and analytic ability." Another item from this survey asks how much emphasis is placed on teaching students certain skills: "major," "moderate," "minor," and "none." Table 7 illustrates a typical item from the Prospects survey.

Table 7.—Item illustrating subitems that total to 100 percent

I29A. Please estimate what percent of your math instruction with this class is conducted in the following
modes. Total should equal 100 percent.

a. Whole numbers/whole number operations
b. Problem solving
c. Common fractions, decimal fractions, and/or percent
d. Ratio and proportion
e. Measurement and/or tables and graphs
f. Geometry
g. Algebra (formulas and equations)
h. Trigonometry
i. Probability and statistics
j. Calculus
k. Learning mathematics facts and concepts
l. Learning skills and procedures needed to solve word problems
m. Develop reasoning and analytic ability
n. Learning to communicate ideas in math
o. Applications of math skills to life
p. Appreciation for importance of math
q. Student confidence in ability to do math
r. Develop perception of math as enjoyable
s. Awareness of application-math to life

SOURCE: Prospects Teacher questionnaire, 1991.

ECLS, the most recent of the surveys in this category and the questionnaire with the greatest number of cognition questions, examines student cognition in depth. Its questions are content specific and somewhat fine-grained. An example (Table 8) follows.

Table 8.—In-depth student cognition item

34. For this school year as a whole, please indicate how each of the following **SCIENCE** or **SOCIAL STUDIES** topics or skills is taught in your class(es)? CIRCLE ONE NUMBER ON EACH LINE.

NOT TAUGHT: Taught at a higher grade level, Children should already know;
TAUGHT: One a month or less, 2–3 times a month, 1–2 times a week, 3–4 times a week, Daily

a. Human body
b. Plants and animals
c. Dinosaurs and fossils
d. Solar system and space
e. Weather (e.g., rainy, sunny)
f. Understand and measure temperature
g. Water
h. Sound
i. Light
j. Magnetism and electricity
k. Machines and motors
l. Tools and their uses

m. Health, safety, nutrition, and personal hygiene
n. Important figures and events in American history
o. Community resources (e.g., grocery store, police)
p. Map-reading skills
q. Different cultures
r. Reasons for rules, laws, and government
s. Ecology
t. Geography
u. Scientific method
v. Social-problem solving

SOURCE: ECLS Kindergarten Spring Teacher questionnaire, 1999.

Other surveys, including Prospects, SASS, and NELS, are also fine-grained, but, in contrast to ECLS, ask not about specific topics covered but about teacher objectives.

From the NELS Teacher questionnaire (Table 9):

Table 9.—Teacher objective item

2_14. In this math class, how much emphasis do you give to each of the following objectives?

A Understanding the nature of proofs
B Memorizing facts, rules, and steps
C Learning to represent problem structures in multiple ways (e.g., graphically, algebraically, numerically, etc.)
D Integrating different branches of mathematics (e.g., geometry, algebra) into a unified framework
E Conceiving and analyzing effectiveness of multiple approaches to problem solving
F Performing calculations with speed and accuracy
G Showing importance of math in daily life
H Solving equations
I Raising questions and formulating conjectures
J Increasing students' interest in math

SOURCE: NELS Teacher questionnaire, 1992.

In addition, some questions from the NELS Student survey are classified in this category and ask students to identify how much emphasis their teachers has placed on listed objectives.

Critique. Items in this category suffer from some of the problems of IP measurements discussed earlier. The recall periods for the surveys are either a full school year (ECLS, NELS, and Prospects) or a semester (SASS); long recall periods make it difficult for teachers to accurately remember and report their activities over the time period. They also do not capture typical practices across the year or semester, as instructional content and practices may change throughout the period.

The validity of the items is also weakened because of the detailed and lengthy items included on the surveys; such time-consuming detail can prove a burden to respondents.

In addition, there is no measuring of teacher quality or student engagement. Although the questionnaires do well in their coverage of the quantity of instructional processes, they do not capture the quality of the instruction or how well the students receive it.

Instructional Pedagogy Items: Grouping Strategies

Under instructional pedagogy the item bank had 119 items in the subcategory of "grouping strategies." The items cover many aspects of grouping, including the frequency with which teachers divide students into groups, the size of average groups, activities used during grouping, and teachers' reasons for creating groups. Response options include multiple choice responses on frequency (e.g., "never/once a month/once a week" etc.), duration (e.g., 1–15 minutes/day, more than 60 minutes/day, etc.), and reasons for using groups (e.g., ability, age, disciplinary), as well as numerical options for number of groups and number of students in groups. Some items ask for the percentage of total class time spent with students divided into groups. Most items ask teachers to recall over an entire year or semester, or to answer in terms of a "typical day." TIMSS collects detailed information on grouping and other aspects of teachers' "last lesson."

All grouping strategy items were found on teacher questionnaires except for one TIMSS item asked of both science and mathematics students, which can be seen in Table 10. Most items were asked in the context of a general subject area such as mathematics or science.

Table 10.—TIMSS Student Questionnaire item on grouping strategies

How often does this happen in your math lessons? We work together in pairs or small groups.
Most lessons, some lessons, never

SOURCE: TIMSS Student Population 1 questionnaire, 1995.

Critique. Grouping strategy items provide useful contextual information for understanding instruction. Also, the items are not particularly difficult for teachers to answer. Teachers do not need to do much interpretation when reporting the social organization of the classroom. Furthermore, most surveys gather fairly comprehensive data on grouping beyond just the frequency with which teachers split their students into groups. NCES surveys provide data on the number of groups used, the number of students in each group, and the type of assistance provided to students in groups. Some items move beyond simple indicators of the frequency of grouping occurrences by asking the length of those occurrences (Table 11). Lastly, teachers are asked about their primary basis for splitting students into groups, such as differences based on age, ability, or English language proficiency (Table 12). Knowing why a teacher uses grouping is more informative than knowing simply how often he or she uses grouping.

Table 11.—ECLS items asking about duration of grouping instances within the context of frequency of instances

16. How often do you divide your class(es) into achievement groups for reading and math activities or lessons? CIRCLE ONE NUMBER ON EACH LINE.
    Never, Less than once a week, Once or twice a week, Three or four times a week, daily
    - a. Reading
    - b. Math
17. On days when you use achievement grouping, how many groups do you have and how many minutes per day are your class(es) usually divided into achievement groups for reading and math activities or lessons? WRITE NUMBERS ON LINES BELOW, AND CIRCLE ONE NUMBER ON EACH LINE. IF YOU DO NOT USE ACHIEVEMENT GROUPING IN THE SUBJECT LISTED, PLEASE WRITE "0" ON THE LINE AND SKIP TO THE NEXT QUESTION.

Number of achievement groups: _______

1–15 minutes/day, 16–30 minutes/day, 31–60 minutes/day, More than 60 minutes/day
- a. Reading
- b. Math

SOURCE: ECLS Kindergarten Spring Teacher Questionnaire, 1999.

Table 12.—Prospects item on primary basis for grouping

Please indicate the primary basis on which instruction groups are formed for math in this class. Circle one
- 01 = Similar math ability
- 02 = Diversity of abilities
- 03 = Same language other than english
- 04 = Handicapping condition
- 05 = All compensatory education students grouped together
- 06 = Unit topics or subject matter
- 07 = No basis/random
- 08 = Other basis

NOTE: The response options appear five times, so that the teacher can fill the item out for each class.
SOURCE: Prospects questionnaire, 1991.

Unfortunately, the items are too broad to reveal the quality of instruction taking place in groups. Items on the frequency of grouping are unlikely to be reported reliably, and are unable to gauge the effectiveness of the actual instances of grouping. Even the newer ECLS items that move beyond frequency and ask the duration of various grouping scenarios in a typical day (as seen in Table 11) are unlikely to result in any stronger link to achievement. An effective teacher using primarily whole group instruction would probably be able to create a more lasting impact on student achievement than a less knowledgeable teacher using long grouping sessions. The simple act of splitting students into groups is not sufficient to improve the achievement of individual students. The "variable of interest," as Leighton et al. (1994) refer to it, is the "extent to which the lesson structure engages students," not group work *per se*, which is not a "predictor of learning" (p. 27).

Other essential contextual information for understanding the quality of instruction is absent from the NCES grouping strategy items. No information on the academic content being covered in the groups is given beyond the subject area, such as language arts or science. Furthermore, teachers can use grouping for varying reasons dependent not only on the content under study, but also the point in time within a given unit or semester.

Referent periods used in some grouping strategy items are problematic. Some items only inquire about teachers' "typical," "usual," or "average" instruction. This may not be how teachers think about instruction. Their instruction can be thought of as having a high "standard deviation"; in other words, the instruction varies widely dependent on context. This is best illustrated by an item from the Prospects questionnaire, in which teachers are asked about the average size of their groups (Table 13). Their groups may vary from 2 to 20 depending on the characteristics of the students, how far the class has progressed in the lesson, class size, the content being presented, the characteristics and cognitive abilities of the students receiving the instruction, or many other contextual factors. In this case, an average group size may be uninformative or even misleading. In another Prospects item, teachers are asked to choose one of a list of reasons for using grouping as their "primary" basis (Table 14). Teachers may use all of those equally, or they may have many "primary" reasons, depending on the same contextual factors just listed. In either case, the language used in the item does not convey the way teachers might think about their own teaching and could elicit answers that are not valid.

Table 13.—Prospects item on average group size

I–24. FOR YOUR MATH GROUPS, WHAT IS THE AVERAGE GROUP SIZE?
_________ (number of students)

NOTE: The response options appear five times, so that the teacher can fill this out for each class
SOURCE: Prospects questionnaire, 1991.

Table 14.—Prospects item on average group size

I–25. PLEASE INDICATE THE PRIMARY BASIS ON WHICH INSTRUCTION GROUPS ARE FORMED FOR MATH IN THIS CLASS. CIRCLE ONE

01 = SIMILAR MATH ABILITY
02 = DIVERSITY OF ABILITIES
03 = SAME LANGUAGE OTHER THAN ENGLISH
04 = HANDICAPPING CONDITION
05 = ALL COMPENSATORY EDUCATION STUDENTS GROUPED TOGETHER
06 = UNIT TOPICS OR SUBJECT MATTER
07= NO BASIS/RANDOM
08 = OTHER BASIS

NOTE: The response options appear five times, so that the teacher can fill this out for each class
SOURCE: Prospects questionnaire, 1991.

Referent periods can cause other problems. Items inquiring about instruction over a year or semester may not be reliable or valid because teachers have trouble recalling their practice accurately that far back. TIMSS items inquiring about a teacher's "last lesson" are as problematic as items on "typical instruction" or instruction over a semester or year. They are simply too specific to be generalized to a teacher's instruction over the course of a year.

NCES grouping strategy items are likely to provide descriptively interesting variation in the social organization of classrooms. However, because the items lack the ability to describe the quality of the instruction taking place in the groups, that variation is unlikely to be related to variation in student achievement.

Instructional Pedagogy: Student Activities

The item bank contained 93 items on "student activities." These items inquire about a wide range of activities in which students would feasibly participate during the course of instruction in a classroom. The main sources of the items were the ECLS-K Spring Teacher Questionnaire, Prospects, NELS Student Questionnaire, NELS Teacher Questionnaire, TIMSS Student Questionnaire, TIMSS Teacher Questionnaire, and the SASS Teacher Followup Survey.

"Student activity" items mostly require teachers to choose one of four or five options detailing the frequency with which they have their students engage in various instructional activities. The items represent dozens of activities, ranging from the general (e.g., "use a textbook"—Teacher Followup) to the specific (e.g., "listen to you read stories where they see the print" and "listen to you read stories but they don't see the print"—(ECLS K Spring). Response options on these frequency items range in specificity from "always, sometimes, never" to "daily, three or four times a week, once or twice a week, two or three times a month, once a month or less, never."

NCES questionnaires follow a typical pattern in the length of referent periods used and the generality of the subject matter context established for individual IP items. Typical items asked teachers to recall their practice over the course of a semester or year. Most of the typical items were asked in the context of a general content area such as mathematics or reading. Teachers of multiple subjects were sometimes asked questions outside of any subject matter context.

Most of the typical student activity items were found on teacher questionnaires. However, a few were included on the student questionnaires from the NELS:88 and TIMSS surveys. Student items were similar to the teacher items; the students were asked how often they participated in a particular activity.

Table 15 shows a representative item on student activities from a NELS:88 teacher questionnaire. Table 16 includes a representative item from a TIMSS student questionnaire.

Table 15.—Typical student activity item from NELS:88 Teacher Questionnaire

2_19. How often do you do each of the following activities in this science class?
Never/rarely, 1–2 times/month, 1–2 times/week, almost every day, every day
a. Have students do an experiment or observation individually or in small groups
c. Require students to turn in written reports on experiments or observations
e. Have students use computers for data collection and analysis
g. Have students give oral reports
h. Have students independently design and conduct their own science projects

NOTE: The response options appear five times once for each class a teacher might teach.
SOURCE: NELS:88 Teacher Questionnaire, 1992.

Table 16.—Typical student activity item from TIMSS Student Questionnaire 1

22. How often does this happen in your math lessons?
most lessons, some lessons, never

b) We copy notes from the board.
c) We have a quiz or test.
d) We work from worksheets or textbooks on our own.
e) We work on math projects.
f) We use calculators.
g) We use computers.
h) We work together in pairs or small groups
i) We use things from everyday life in solving math problems
k) We can begin our homework in class.
m) We check each other's homework.
n) We discuss our completed homework.

SOURCE: TIMSS Student Questionnaire 1, 1995

Critique. The typical student activity items suffer from a number of the problems outlined earlier in this paper. We know that instructional activities are intimately intertwined with the subject matter content taught. The typical student activity items do not address subject matter content at the depth that would provide sufficient context for understanding the activity. Asking teachers what happens in "math lessons" or "this science class" does not provide sufficient context to understand the nature of the student activity taking place. Furthermore, asking only the frequency of an activity provides only information on *how often* that activity took place. It does not tell *how* the activity took place, or *how effective* it was. In other words, the quality of the instruction is not revealed to any extent.

The use of long referent periods reduces the usefulness of the items, as well. These items usually ask teachers to recall typical practice over the course of an entire school year. The activities in question, however, may occur only at certain points in time throughout the year or differ in their frequency across subject matter topics taught. Even frequency response options such as "every day" do not allow the teacher to express a situation in which the activity was employed "every day" for several weeks and then "once a month" as the year progressed. Furthermore, the effectiveness of student activities is dependent on many other contextual factors such as the resources available and the characteristics of the students in the class.

Some student activity items did not fit the typical "frequency of activity" mold. An item on the Prospects questionnaire asks the teacher about the exact percent of total classroom time in which students are "actively engaged in academic activities." This appears to be a useful item, because time on task is an appropriate indicator of students' opportunity to learn, and time spent engaged in academic activities has demonstrated a small link to student achievement (Rowan, 2001). However, this item is flawed because, as mentioned before, more effective teachers are more likely to be able to diagnose student engagement. Consequently, some teachers will report on this item less accurately than others.

Lastly, student activity questions that address students as respondents suffer from regular differences among students, especially in the elementary grades, in how well they can respond accurately to the questions. The accuracy with which students respond to these items will also vary with the students' success in the class. In other words, students who have not been engaged with the content presented by the teacher during her or his instruction will not be able to report as accurately on that instruction. Regardless of students' ability to accurately respond to the item, the response options provided to the students ("most lessons, some lessons, never") are too broad to provide enough variation to link to variation in the students' achievement.

There are some positive aspects to the student activity items. Whereas most instructional practice items focus on the distribution of instruction by the teacher, these items provide some perspective on how instruction is actually received by the student, and they give the researcher some information on variation in the nature of interactions between teachers and students. Furthermore, items on student questionnaires, despite having questionable validity, might be useful in validating teacher reports.

One item, shown in Table 17, asks students about an important conceptual issue: engagement in the classroom. Information garnered through student reports must be questioned, but it is encouraging to see items focusing substantively on student engagement.

Table 17.—NELS item on student engagement

17. In your current or most recent SCIENCE class, how often do/did you do the following:
    Never, rarely, sometimes, often, always
    - Pay attention in class?
    - Complete your work on time?
    - Do more work than was required of you?
    - Participate actively in class?

SOURCE: NELS Student Questionnaire, 1992.

Another substantively interesting—if methodologically flawed—item is found on the Prospects questionnaire (Table 18). This item makes an attempt to capture a construct, teacher knowledge of the student, which the IPRD team has determined to be essential to effective teaching. The item measures the frequency of instances of teachers attempting to clarify the degree to which a student is engaging productively with content. Although this item is well oriented conceptually, it is a poignant reminder of the limitations of survey questionnaires. Despite covering an area believed to be substantively essential, the item does not provide any information on the quality of interactions between students and teachers. The item asks only the frequency of these instances in an overly broad subject area over the course of an entire year, and consequently falls victim to the same methodological barriers as other frequency items in this category. The results from this item might show variation, but they will not provide information on the effectiveness of teachers' efforts to understand better the degree of their students' engagement with the content.

Table 18.—Prospects item on student content engagement

I–31. On average, how often do the following occur when you teach *math*? Almost never, sometimes, frequently, almost always

a. Students seek clarification about directions

c. Students are asked questions to check for understanding

NOTE: The response options appear five times, once for each class a teacher might teach.
SOURCE: Prospects questionnaire, 1991.

Newer "frequency of student activity" items have made significant advances. An item on the 1999 ECLS-K teacher questionnaire, shown in Table 19, is very specific about the content associated with the activity, and it provides enough detailed subitems—23 in all—to allow researchers to create robust composite measures of teaching styles. This attention to content and long list of activities increases the likely amount of variation and likelihood of impact on student achievement. However, the ability of the item to uncover the quality of instruction is still hampered by the "frequency of activity" format, which is susceptible to reliability problems and does not capture the quality of a teacher's execution of an activity. For example, teachers may have dozens of different ways to "play math-related games," all with varying degrees of effectiveness. Also, the length of the item, taken in the context of the longer questionnaire of which it is a part, greatly increases the burden on the teacher-respondents, which can decrease overall validity and reliability. Reliability of the item is also damaged by the length of time over which teachers are asked to recall their practice.

Table 19.—ECLS item on student activities in mathematics

How often do children in this class do each of the following **MATH** activities?
Never, once a month or less, two or three times a month, once or twice a week, three or four times a week, daily

a. Count out loud

b. Work with geometric manipulatives.

c. Work with counting manipulatives to learn basic operations .

d. Play math-related games

e. Use a calculator for math

f. Use music to understand math concepts

g. Use creative movement or creative drama to understand math concepts

h. Work with rulers, measuring cups, spoons, or other measuring instrument

i. Explain how a math problem is solved

j. Engage in calendar-related activities

k. Do math worksheets

l. Do math problems from their textbooks

m. Complete math problems on the chalkboard

n. Solve math problems in small groups or with a partner

o. Work on math problems that reflect real-life situations

p. Work in mixed achievement groups on math activities

q. Peer tutoring

SOURCE: ECLS-K Field Test Spring Teacher Questionnaire, 1999.

Instructional Resource Items: Computers

Seventy items in the item bank covered the presence of or use of computers as instructional resources in schools and classrooms. The ECLS Kindergarten Spring Teacher Questionnaire, NELS student and teacher questionnaires, TIMSS student and teacher questionnaires, and the Teacher Followup Survey each included several items. The Prospects questionnaire, however, included a detailed section of computer items.

Most of the surveys inquire about the number of computers present, adequacy of equipment, subject areas and types of instructional activities that include computers, frequency of instructional use, and barriers that teachers encounter in using computers for instruction. The Prospects questionnaire addresses two additional topics. The first topic is the use of specific instructional software programs, such as Pogrow's Higher Order Thinking Skills and IBM's Writing to Read. The second inquires about the goals toward which teachers use computers in the classroom, such as rewarding students for completing their work and understanding concepts in mathematics and science. Some items are not specifically focused on computers, but included computers as subitems in a list of classroom resources such as textbooks and basal readers.

Response options are similar to those on most other IP items. Most items provide four or five closed-ended response options asking teachers to choose levels of frequency of use, levels of adequacy of the equipment, and degrees to which various barriers inhibited the use of computers. Examples of this type of response options are found in Tables 20 and 21.

Table 20.—NELS Teacher Questionnaire item on classroom computer use

How often do you use the following teaching methods or media? Never/rarely, 1–2 times/month, 1–2 times/week, almost every day, every day

b. Use computers

Source: NELS Teacher Questionnaire, 1992.

Table 21.—ECLS-K Spring Teacher Questionnaire item on adequacy of computer equipment

In general, how adequate is each of the following for your class(es)?
I don't use these at this grade level, Never adequate, Often not adequate, Sometimes not adequate, Always adequate

- Computer equipment
- Computer software

Source: ECLS-K Spring Teacher Questionnaire, 1999.

For the most part, computer items did not contain much contextual information on the period of time for which teachers are asked to recall their use of computers or the content matter involved. Most items did not specify a referent period, although some asked about teachers' "regular" practice or "so far this year." The TFS items are more specific, asking teachers to think back over the past semester. The computer items do not set much context in terms of content. Usually, as with most IP items reviewed so far, just a subject, such as mathematics or reading, is given. However, some items, such as the ECLS item in Table 22, ask about frequency of computer use for content-specific activities.

Table 22.—ECLS-K Spring Teacher Questionnaire item on adequacy of computer equipment

How often do children in your classes use computers for the following purposes?
Never, Once a month or less, Two or three times a month, Once or twice a week, Three or four times a week, Daily

- To learn reading, writing, or spelling
- To learn math
- To learn social studies concepts
- To learn science concepts
- To learn keyboarding skills
- To create art
- To compose and/or to perform music
- For enjoyment (e.g., games)
- To access information (e.g., to connect to Internet or local network)

Source: ECLS-K Spring Teacher Questionnaire, 1999.

Critique. Items collecting information in the instructional use of computers are not as inherently problematic as other areas of instruction. The language of the items is straightforward and reports on material resources require less interpretation than reports on strategies and activities. These types of items are also more amenable to validation with student questionnaire responses.

However, the items are still too broad to collect information that will link to student achievement. There is not enough contextual information within the items. For instance, a TIMSS item asks teachers how often they use computers in mathematics class (Table 23). This information is not very useful to someone interested in the quality of a teacher's instruction. It has no information on content other than the subject matter, and it does not say what types of activities the computers are being used for.

Table 23.—TIMSS item on frequency of computer use

In your mathematics lessons, how often do you ask students to use computers? Never or almost never, some lessons, most lessons, every lesson

SOURCE: TIMSS Teacher Questionnaire Population 2 Mathematics, 1995

Another problem with the computer items is that they are simply too numerous, considering other instructional resources that may be important to measures. Other resources, such as textbooks and worksheets, are probably used in classrooms more often than computers, yet NCES surveys do not go into nearly as much detail measuring them.

Assessments: Types Used

Sixty-eight items in the item databank were categorized as types of assessments used. These items represented four questionnaires: Prospects Teacher (1991), TIMSS Teacher Population 2 Math and Science (1995), and the SASS Teacher Follow-Up (1998). The majority (thirty-eight total items) appeared in the SASS questionnaire.

All surveys had teacher respondents, and all had a referent period of a school year. The Prospects survey questions applied to all classes for the respondent, while TIMSS and SASS asked about designated classes. TIMSS asked exclusively about the weight given to listed types of assessments in assessing the work of students, including standardized tests, teacher-made tests, homework, projects or exercises, observation, and student responses in class. SASS similarly asked about the importance given to types of assessments in determining grades or progress reports. The following question (Table 24) comes from the TIMSS Teacher Population 2: Science survey and illustrates the typical importance questions.

Table 24.—Assessment item on importance of various assessment modes

22. In assessing the work of the students in your science class, how much weight do you give each of the following types of assessment?
    - a) standardized tests produced outside the school
    - b) teacher-made short answer or essay tests that require students to describe or explain their reasoning
    - c) teacher made multiple choice, true-false and matching tests
    - d) how well students do on homework assignments
    - e) how well students do on projects or practical/laboratory exercises
    - f) observations of students
    - g) responses of students in class

SOURCE: TIMSS Teacher Population 2 Science questionnaire, 1995.

Prospects and SASS contain questions on the use and frequency of assessments or other means of determining student grades or progress. SASS includes five response options ranging from almost every day to never, while Prospects asks if each assessment was used at time of entry, other times, or both. Prospects also asks about the frequency (daily, weekly, monthly, once or twice a year, not at all, and not applicable) of consultation with others when evaluating student progress.

SASS includes several questions about the use of portfolios in the classroom, including broad subject areas in which portfolios are used, what is included, how are decisions made about portfolios, how often portfolios are used, and about the teacher's agreement (strongly agree, somewhat agree, somewhat disagree, and strongly disagree) with statements about portfolios.

Table 25.—Portfolio assessment item

49. How often did you use student portfolios in your DESIGNATED CLASS last semester or grading period for the following purposes?
    - a. Training students to reflect upon and/or assess each piece of work
    - b. Training students to reflect upon and/or assess their overall progress
    - c. Communicating student progress to parents
    - d. Determining student grades or other formal progress reports
    - e. Planning for future lessons
    - f. Diagnosing student learning problems
    - g. Making informed decisions about student placement

h. Making informed decisions about student graduation
i. Providing information for program or school accountability

SOURCE: SASS Public School Teacher Survey, 1993.

Critique. Three of the four questionnaires, Prospects and both TIMSS surveys, collect only general information about teachers' use of assessment. Prospects asks what measurements are used to judge student progression and often teachers consult other members of the faculty; TIMSS about the weight given to different types of assessment. In contrast, SASS collects detailed information not only about frequency of student assessments and weight given to assessments, but also about the use of portfolios in the designated class. Besides the previously cited example, the questionnaire includes items on what types of work are included in portfolios, how decisions were made about portfolios, and teacher agreement with statement about the portfolio policy and process. No other surveys mention portfolios.

The items in this category show some of the problems associated with measuring IP. All the surveys have recall periods of the entire school year, making their accuracy questionable, and do not make accommodation for different practices during different units, subjects, or parts of the year. In addition, the surveys provide no link between teachers' activities in regards to assessment and the achievement of their students or between instructional content and assessment. There is also no information about interaction between students and teachers.

Assessment: How Used

The IP survey of selected NCES questionnaires found fifty-eight items that were classified under how assessments were used. These items appeared in five of the surveys contained in the IP databank: Prospects Teacher, SASS Teacher Field Test, TIMSS Teacher Questionnaire 2 Mathematics, TIMSS Teacher Questionnaire 2 Science, and SASS Teacher Follow-Up.

Both TIMSS surveys ask how often assessment information gathered from students is used for listed items, including providing feedback to students, reporting to parents, planning future lessons, etc. SASS includes a similar question. Prospects' one item in this category asks how frequently information about student progress is shared with each group—students, parents, and other teachers. Items from the SASS Teacher Follow-Up Survey make up the majority of items in this category. Besides the item similar to TIMSS, SASS asks how much importance is given to listed components (effort, class participation, test results, relative achievement, etc.) in determining student grades. SASS also includes a significant number of questions about portfolios and their use, including how often they were used in assessing students.

Most of the survey items classified in Assessment: How Used covered an entire year (some SASS items asked only about the last semester), were teacher respondent, and covered one subject area. Most asked how often certain practices were conducted, although a few asked about the importance or content of types of assessments. All questions were measuring practice. For items asking about the frequency of certain practices, 4–5 response options were offered and generally included "none," "little," "a lot,” and “a great deal” or similar options. For the SASS item about the importance given to the listed items, response options were “extremely important,” “very important,” “somewhat important,” and “not important.” The following question (Table 26), from the TIMSS Teacher Population 2: Science questionnaire, is typical of items in this category.

Table 26.—Typical use of assessment item

23. How often do you use the assessment information you gather from students to...
a) provide students’ grades or marks?
b) provide feedback to students?
c) diagnose students’ learning problems?
d) report to parents?
e) assign students to different programs or tracks?
f) plan for future lessons?

SOURCE: TIMSS Teacher Population 2 Science questionnaire, 1995.

In addition, some questions in this category also appear in that of types of assessments used, including items from SASS on the use of portfolios in the classroom. An example of this type of question is below (Table 27).

Table 27.—Example of item that appears in “Assessment: Types Used” and “Assessment: How Used”

46. In what content areas were PORTFOLIOS used with your DESIGNATED CLASS?
English/language arts
Math
Reading
Social studies
Science
Art
Music
Home economics
Foreign language
Other
None

SOURCE: SASS Public School Teacher Survey, 1993.

Critique. The surveys in this category do a good job of covering the use of assessments. The items themselves are not necessarily very fine-grained, but they do gather general information on how teachers utilize assessments. Prospects gathers information on how frequently teachers share information about student progress. TIMSS asks about the use of assessment information in providing grades, feedback, building future lesson plans, etc. SASS, in addition to the portfolio questions, includes items on the use of information, similar to TIMSS, and on the importance given to various factors, including effort, participation, and test results, in determining grades. SASS also collects information about the work included in portfolios, the use of portfolios, and teachers’ agreement with statements about portfolio policies and practices.

However, questions in this category illustrate the problems of measuring IP in questionnaires. The surveys do not link teacher activities and use of assessment with student achievement. They rely upon teacher self-reports and have recall periods of either a full semester (in SASS) or a full year, calling into question their accuracy, and do not allow for fluctuations in teacher practices as subject matter or units change. In addition, student and teacher interaction are not linked.

Instructional Time Items

NCES surveys included 57 items measuring instructional time. Items were included in this category if their focus was on the specific amount of time spent by students and/or teachers on activities related to instruction. Most items were found on the ECLS-K Fall and Spring teacher questionnaires and Prospects, although a few were included in the NELS and TIMSS teacher questionnaires.

Items cover a wide range of topics and use a variety of formats for their response options. A few items ask teachers to report the amount of time spent teaching per week. Most, however, ask how that time is spent. Items inquire about time spent on various subjects, instructional time (as opposed to administrative and preparation time, time devoted to students' social development, behavioral management, and other kinds of time), grouping formations, various instructional activities, and physical settings (e.g. laboratories). Response options vary throughout the instructional time items. Some ask teachers to choose from a range of options, such as "1–15 minutes" or "three hours or more" (Table 28). Others ask for exact minutes or hours per day or week. Some items ask teachers to report time on various activities on percentages that total to 100 percent. A few items first ask how frequently certain activities occur, and then ask how much time is spent in one typical instance of that activity (Table 29).

Table 28.—Typical instructional time item with multiple choice response options

In a typical day, how much time do the children spend in the following activities? CIRCLE ONE NUMBER ON EACH LINE. DO NOT INCLUDE LUNCH OR RECESS BREAKS. IF YOU TEACH MORE THAN ONE CLASS, CONSIDER ALL CLASSES WHEN MARKING YOUR RESPONSES. No Time, Half hour or less, About one hour, About two hours, Three hours or more
a. Teacher-directed whole class activities?
b. Teacher-directed small group activities?
c. Teacher-directed individual activities?
d. Child-selected activities?

Source: ECLS K Fall Teacher Questionnaire, 1999

Table 29.—ECLS item asking duration of instruction in a content area within context of frequency of that instruction

How often AND how much time do children in your class(es) usually work on lessons or projects in the following general topic areas, whether as a whole class, in small groups, or in individualized arrangements?
How Often: Never, Less than once a week, 1–2 times a week, 3–4 times a week, Daily
How Much Time: 1–30 minutes a day, 31–60 minutes a day, 61–90 minutes a day, More than 90 minutes a day

a. Reading and language arts
b. Mathematics
c. Social studies
d. Science
e. Music
f. Art
g. Dance/creative movement
h. Theater/creative dramatics
i. Foreign language
j. English-as-a-second-language (ESL)

SOURCE: ECLS K Fall Teacher Questionnaire, 1999

Instructional time items are similar to other NCES IP items in the level of detail regarding the subject matter content, but deviate from other items in the referent period over which they ask teachers to recall their instruction. Although some items do not ask about content at all, most are similar to items in other categories in that they inquire as to the general subject, such as reading or mathematics, in which the time was spent on the listed activities. Referent periods in instructional time items are necessarily different. Most of the other categories ask teachers to recall the frequency of various aspects of their practice over the past year or semester. Time items, however, ask about instruction in a "typical" day, week, or month, except for the TIMSS item that asks about the "last lesson."

Time items can be slightly more useful than frequency items, simply because they are more specific. They give the duration of a given activity, whereas frequency items leave the researcher in confusion over just how much time was spent on the listed activity. Even more telling than items that collect the amount of time spent on an individual activity are items that ask for the amount of time on a particular activity in relation to time spent on other activities. Items like the one in Table 30 ask teachers for time spent not in terms of minutes or hours per day, but as a percentage of total class time.

Table 30.—Item asking for instructional time as a percentage of total class time

In a typical month of lessons for your mathematics class, what percentage of time is spent on each of the following activities? (*write in a percentage for each activity*)

The total should add to 100%

a. administrative tasks
___________%

b. homework review
___________%

c. lecture-style presentation by teacher
___________%

d. teacher-guided student practice
___________%

e. re-teaching and clarification of content and procedures
___________%

f. student independent practice
___________%

g. tests and quizzes
___________%

h. other
___________%

SOURCE: TIMSS Repeat Mathematics Teacher Questionnaire, 1999.

Critique. These items do pose methodological problems. Teachers are asked to give the amount of time spent during a "typical" day, week or month, which they may not be able to do. An effective teacher modifies her or his instruction to match the students being taught, the content being presented, and how far into the lesson the class has gone, among other factors. This dependence of instruction on the dynamics of the classroom context might result in a teacher feeling that he or she cannot not picture a "typical" day, week or month. Some teachers will not be able to give an approximation of the amount of time spent on an activity within a typical span of time without providing information on the content being taught or the students who are learning the content. The instructional time items on the NCES surveys do not collect this contextual information.

Even if teachers could respond to instructional time items accurately, the items would not collect sufficient information. One might find variation in the amount of time spent on different activities, but this information will be merely descriptive because the amount of time spent on a given activity will not indicate the quality of instruction that took place during that time. Some teachers can use particular activities more effectively in less time than other teachers. Therefore, an instructional time item is unlikely to predict any change or variation in student achievement.

Data on instructional time is worth collecting on IP surveys. It is the most simple, straightforward way to break down the activities in a classroom. However, teacher self-reports on time are subject to reliability and validity problems, and items will not predict student achievement if they do not collect enough contextual information to gauge a teacher's effectiveness during the time measured.

Homework: How Used

Fifty-six items in the IP survey covered how homework was used. These items came from six questionnaires: NELS Teacher, TIMSS Student Population, TIMSS Teacher Population I, TIMSS Teacher Population 2 Math and Science, and SASS Teacher Follow-Up.

The majority of items in this category refer to activities in one class for the entire year. The NELS survey asks for information on each of the teacher's classes. Most questions, including those from NELS, TIMSS Teacher Population 2, and SASS, ask how often homework is checked or used for listed purposes. Response options for these items generally were "never," "rarely," "sometimes," and "always." An example (Table 31) from NELS is below.

Table 31.—Example of item from "Homework: How Used"

2_9. How often do you do each of the following with homework assignments?	
A	Keep records of who turned in the assignment
B	Return assignments with grades or corrections
C	Discuss the completed assignment in class

SOURCE: NELS Teacher Questionnaire, 1992

Two items were exceptions in this category. TIMSS Teacher Population 1 asked teachers to recall the content for their most recent lesson, to place the listed choices in the order in which they were taught, and to identify how many minutes were spent on each one. The SASS Teacher Follow-Up questionnaire included an item on whether listed types of student work were included in portfolios. The following example (Table 32) comes from TIMSS.

Table 32.—Example of order of lesson item

27a. How did the lesson proceed? (order in lesson and minutes spent)
review of previous lesson(s)
review or correction of previous lesson's homework
assignment of student homework
students work on homework in class

SOURCE: TIMSS Teacher Population 1, 1995.

In addition, this category includes items from the TIMSS student survey, which had as its respondents students rather than teachers, as is the case with the other surveys. The one question on the use of homework in this survey is below (Table 33).

Table 33.—Example of student questionnaire use of homework item

23. How often does this happen in your science lessons?
l) The teacher checks homework.

SOURCE: TIMSS Student questionnaire, 1995.

Critique. The surveys in this category do a fairly good job of collecting information on the uses of homework. Although the questions are not as finely grained as those from other categories, they are nonetheless fairly extensive and provide a good amount of coverage on the possible uses of homework, with the NELS and TIMSS student questionnaires asking a limited amount about homework and the TIMSS teacher and SASS questionnaires collecting more information. SASS also includes a question about the type of work included in student portfolios and is the only survey to gather information on portfolios.

Despite this, items in this category suffer from some of the problems in IP measurement discussed earlier. The recall period for most surveys is a full year, which is too far back for teachers to recall their practice accurately and reliably. The long recall period also makes it difficult to capture typical practices, those that are habitually used and not clustered in limited places, across the time frame. Further, the TIMSS questions ask only about one class, making generalizability of the information problematic.

The use of teacher self-reports does not take into account the effectiveness of the teacher and teacher activities, and there can be no connection made between teacher activities and student achievement. The TIMSS student questionnaire, the only questionnaire with a non-teacher respondent, provides some information about student engagement; however, student reports may be invalid since students at different achievement levels are likely to report different levels of engagement.

Innovative TIMSS Items

The Third International Mathematics and Science Study, conducted in 1995, featured innovative instructional processes items that differed significantly from items included on surveys such as NELS:88 and SASS. One set of items from the teacher questionnaires asked teachers to provide a great deal of information about their most recent lesson. To date, items had asked teachers to describe “typical” days or recall their practice over a semester or year. Another set of items, in the Opportunity to Learn questionnaire, assessed teachers’ knowledge of pedagogy in the context of very specific academic content. Both of these groups of items represented methodological improvements over past items. Unfortunately, both encountered new methodological problems that resulted in the items being excluded from the 1999 administration of the TIMSS Repeat Study (E. Gonzalez, personal communication, September 24, 2001; P. Gonzales, personal communication, March 13, 2002).

Table 34 shows the set of items addressing a teacher's "last lesson." These items sidestep the recall problems of other items that ask teachers to describe their practice as far back as a year. Furthermore, the items are extremely comprehensive and holistic. A teacher's instruction is described within a very rich context, including how many minutes were spent on the lesson, a very specific description of the academic content under study, the social organization of the classroom, how far into the lesson the class had advanced, whether homework was assigned, and how long the teacher expected students to take to complete the homework. Within this context, a teacher is presented with an extensive list of classroom activities and is asked to 1) present the activities in the order in which they occurred and 2) write down how many minutes each activity lasted. This descriptive richness represents a major advance over previous IP items, which had provided very little context for understanding teachers' classroom practices.

Table 34.—TIMSS item set covering "last lesson"

How many minutes was the last mathematics lesson you taught to your class? ______ minutes

For each of the following mathematics topics, indicate whether or not it was the subject of the lesson.

1. Whole Numbers
2. Common and Decimal Fractions
3. Percentages
4. Number Sets and Concepts
5. Number Theory
6. Estimation and Number Sense
7. Measurement Units and Processes
8. Estimation and Error of Measurements
9. Perimeter, Area and Volume
10. Basics of One and Two Dimensional Geometry
11. Geometric Congruence and Similarity
12. Geometric Transformations and Symmetry
13. Constructions and Three Dimensional Geometry
14. Ratio and Proportion
15. Proportionality: Slope, Trigonometry and Interpolation
16. Functions, Relations, and Patterns
17. Equations, Inequalities, and Formulas
18. Statistics and Data
19. Probability and Uncertainty
20. Sets and Logic
21. Problem Solving Strategies
22. Other Mathematics Content

Was this lesson…

- the introduction of the topic?
- a continuation of a previous lesson on the same topic?
- the end of the coverage of this topic?

Did you assign homework after the class lesson?

If yes, how long would it take a typical student to complete this homework?

NOTE: Think of the last lesson in which you taught mathematics to your class. (If this lesson was atypical, e.g., an examination or a field trip, pick the previous one.)

How did the [math] lesson proceed?

The following presents a list of activities that may occur during a lesson. Although the list is not exhaustive of what happens in a classroom, most classroom activities may be considered as variations if those listed below. Using this list, indicate how your lesson developed. In the blanks on the right, write in the order in which the activities used in the lesson took place (1 = first, 2 = second, and so on) and estimate the amount of time you spent on each one. Ignore activities you used that do not fit into the descriptions listed. Write in the order and the approximate number of minutes for each activity. NOTE: If you did not do a certain activity write zero in the blank next to it.

Order:

Minutes:

- review of previous lesson(s)

  ____

  ____

- a short quiz or test to review previous lesson

  ____

  ____

- oral recitation or drill (students responding aloud)

  ____

  ____

- review or correction of previous lesson's homework

  ____

  ____

- introduction of a topic (class discussion, teacher explanation/demonstration, film, video, use of concrete materials etc.)

  ____

  ____

- development of a topic (class discussion, teacher explanation/demonstration, group problem solving, film, video, etc.)

  ____

  ____

- small group activities (with or without teacher)

  ____

  ____

- students do paper-and-pencil exercises related to topic (not the same as homework)

____
____

- assignment of student homework

____
____

- students work on homework in class

____
____

- student laboratory or data collection activity (not a separate laboratory hour) or hands-on session

____
____

In this class lesson did the students work in small groups? none of the time, some of the time, all the time

SOURCE: TIMSS Teacher Population 1 questionnaire, 1995.

Advantages gained in the richness of the description, however, were offset by the confusing language of the item and researchers' inability to generalize the results to teachers' larger bodies of instruction throughout the year. These problems resulted in data that "didn't make any sense" (P. Gonzales, personal communication, March 13, 2002). Teachers do not think about their instruction in terms of the order of activities and the number of minutes spent on each activity. This casts the reliability of the item into doubt. Also, because teaching is so dependent on the contextual information garnered through this new instrument, it is likely that a teacher's instruction would change from lesson to lesson. This means that all one learns about a teacher from these items is how a teacher instructs students in one particular finely grained academic content area at one particular point in a unit. It might be possible to statistically link a teacher's instruction in this topic to the change in a student's ability to master that small area of content on an assessment. However, this would provide little information on a teacher's ability to teach a general subject area or the change in a student's academic ability in that general subject area over the course of an academic year.

Table 35.—TIMSS pedagogical knowledge items

To better understand what teachers believe about how science is best taught, we are asking you to respond to two out of the three following hypothetical teaching situations. Several possible approaches are presented for each situation. Some of the situations may involve topics that are tangential to your current teaching field, and therefore, may or may not be likely to occur in your classes. Nevertheless, we are interested in what you believe would be the best approach or sequence of approaches to help students learn in these situations regardless of whether they may occur in your classes.

- Respond to the two situations with the science content most similar to your background and experience.
- Imagine yourself in each situation.
- Assume that there are no time or equipment constraints.

For item 1:

- Indicate how strongly you agree or disagree with each of the four statements made about the teaching approach presented.

For items 2 and 3:

- Choose what you believe, based on your own principles and beliefs, to be the best approach or sequence of approaches to help students learn.
- Number the boxes next to each approach in the order in which you would consider using them. If you would use only one approach, place a '1' in that box only. Write zero in the box for any approach you would not consider using.

Remember, respond to only two of the next three items: 1, 2, and 3.

1. A teacher began instruction on a new topic in energy (e.g., chemical energy, mechanical energy, energy in life or earth processes). The first thing the teacher did was ask the students "What do you think energy is?" Student responses ranged from very accurate to quite incorrect. What is your opinion about this approach?

Check one box in each row.

strongly disagree
disagree
agree
strongly
agree

a) This approach should be avoided because some students might get confused by other students' inaccurate ideas about energy. ....

b) The teacher should have begun instruction by first explaining what energy is ....

c) This approach was useful because the teacher became aware of the students' ideas about energy...

d) The teacher should have begun instruction with a demonstration of the effects of energy followed by a discussion of the concept of energy.....

2. A student puts his hand in the water in the class aquarium and says, “Look! My hand swells up in the water. See how much bigger than normal it is?” Although the student’s hand does appear to be larger than normal when in the water, the student’s reasoning is not an accurate explanation of this phenomenon.
   If you were working with a class in which you suspected many students shared the belief stated by the student above and if there were no time constraints on what you might do in responding to this, what approach or sequence of approaches do you believe would best help students learn?
   Place a ‘1’ in the box next to the approach you believe to be the best. If you believe other approaches would also be acceptable, place a number in the box next to each one indicating the order in which you would consider using it. You need not choose more than one approach. Write a zero in the box for any approach you do not consider acceptable.

a) I would explain to the students how water affects the angles of reflection making an object appear larger than it really is ....

b) I would ask the class questions about how the different objects in the aquarium appear in and out of the water to lead them to understand that the above explanation is inaccurate. ....

c) I would give the students an experiment to do measuring the size of different objects both outside and in several different places inside the aquarium to generate data contrary to the above statement...

d) I would ask the students to design and conduct an experiment on the size of objects outside and inside an aquarium that would help them decide whether the explanation above is correct.

e) I would have the students read relevant information from their textbooks.......................

f) I would demonstrate an experiment or show a filmstrip on how water affects the appearance of objects which would provide a basis for arriving at a more accurate explanation................

g) I would have the students compare their ideas about why objects would appear to be different sizes inside and outside of the aquarium in a discussion ....

h) Which of the approaches listed above do you believe to be the least acceptable approach? *Place the letter (a–g) of that approach in the box* ....

SOURCE: TIMSS Teacher Opportunity to Learn (Science) questionnaire, 1995.

The TIMSS OTL item on pedagogical knowledge was a complete change from previous IP items. Instead of asking teachers to describe actual practice, these items present them with a hypothetical classroom situation and ask them to choose what they think is the best instructional practice in that situation. Not only are they asked to choose the best one, but they also are asked to rank the remaining practices. Teachers had been asked about their pedagogical philosophies and beliefs before on NCES surveys, but they had never been asked to use those beliefs and philosophies in a hypothetical situation. These items are more like an assessment of teacher knowledge than a survey of teacher practice.

This move toward assessment of teacher knowledge is encouraging. The IPRD team's studies of learning and instruction suggest that effective instruction is based upon knowledge of the content being presented, the level of students' understanding, and the pedagogical tools available. These items address all three aspects. A very specific area of content is being taught, a particular misunderstanding on the part of a student based on knowledge about how that particular content is learned is presented, and a range of pedagogical approaches are offered.

Unfortunately, the data resulting from these items also had problems. There was considerable controversy over the questions, and researchers had a "hard time making sense of" the resulting data (E. Gonzalez, personal communication, September 24, 2001). Consequently, the items were among the first to be deleted during the development of the TIMSS Repeat questionnaire. However, the items hold promise. This particular TIMSS questionnaire, which had only four teacher knowledge items, may not be predictive of student achievement because the four items cover too thin a topical area to be generalized to instruction in a given subject over the course of a year. However, a larger battery of such items would likely be more generalizable. Preliminary work by Rowan and his colleagues toward developing this kind of battery of items indicates that measures of teachers' knowledge of subject-specific pedagogy show promise for linking to growth in student achievement (Rowan, 2001).

Conclusion

NCES survey measures of instructional processes are almost necessarily plagued by many problems because of the complex nature of teaching. Individual problems with the measures add up to two disappointing conclusions. First, current measures are not predictive of the main outcome of teaching—changes in student academic achievement. Second, most items suffer from a variety of problems that damage their reliability and validity. This limits NCES' ability to report even descriptive results with an adequate degree of confidence.

The lack of predictive validity among NCES IP items stems from their inability to capture the quality of instruction, and a lack of focus on conceptual areas that are essential for a full understanding of classroom learning and teaching. Most NCES items lack sufficient context to capture the quality of a particular teacher's instruction. The instructional strategy a teacher chooses is dependent upon the prior knowledge and other characteristics of the individual students in the class, the content being taught, and many other factors that current measures do not capture. Furthermore, NCES has few items that address areas of instruction with strong impacts on student learning. Specifically, current measures focus too narrowly on teacher activities, ignoring the essential role of the student in learning. Also, current items ignore the knowledge—of content, pedagogical options, and their particular students' cognitive structure and capabilities—that effective teachers use as the basis for their instructional decisions.

Not only are current measures unable to predict student achievement, but they also suffer from problems that damage their reliability, validity, and generalizability. Relying solely on student or teacher reports casts questions of reliability and validity over all of the items. Many IP items are burdensome to the respondent, either because they contain too many subitems, or because they are buried in longer surveys that cover many areas of schooling with little depth. High burden reduces validity because teachers become less and less likely to devote the effort necessary to answer questions accurately. Most items inquire about the frequency of specific practices or situations, even though research has shown frequency items to have questionable reliability. The referent periods given to respondents in many items are questionable. Many items inquire about "typical" practice, even though instruction is so dependent on content and classroom dynamics as to render typical practice in a general subject area a moot point. Items that cover instruction over a semester or year may be asking teachers to recall so far back in time that most will not be able to report accurately. Other items that ask about recent instruction, such as in a teacher's "last lesson," are not generalizable to that teacher's larger body of instruction over the academic year.

Despite these problems, current NCES IP items do provide a steppingstone toward better reliability and validity. Some newer items on ECLS and TIMSS couch instruction in a richer context. Instead of presenting instruction separate from content, these items ask about the social organization of the classroom, the specific content being presented, whether instruction was reviewing or introducing a topic, whether homework was assigned, and other important contextual variables. Some items ask about a sufficient number of practices as to be analyzed at the composite level, which is more reliable than simply reporting findings on the frequency of individual practices (Mayer, 1999b). Some items address appropriate conceptual areas, such as student engagement and teacher knowledge of pedagogy within a specific topical area. The teacher pedagogy items from the TIMSS Opportunity to Learn questionnaire, in particular, signify a move in the right conceptual direction.

Recommendations

Improvement of NCES measurement of instructional processes can proceed along two tracks. The first is a low to medium-level effort aimed at developing valid and reliable descriptive survey measures of instruction. This effort would entail improvement of current items and development of new items. The second approach, which is more ambitious and useful, is the implementation of a major measurement development program aimed at creating data collection tools that will pick up all of the variation in student achievement—likely between one-tenth and one-third—thought to be impacted by variation in classroom instructional processes. This would entail a major conceptual and methodological reorientation and intensive in-field development activity.

Descriptive measures of instructional processes

Our review of the literature and critique of current items found that it is possible to obtain valid and reliable survey measures of instruction on a descriptive level. Previous research outside of NCES has shown that although little variation currently exists in instruction, survey measures are capable of picking up what variation is there (Blank et al., 2001; Burstein et al., 1995; Gamoran et al., 1997; Rowan, 2001). In order to develop adequate descriptive measures of instruction, NCES must: 1) pay close attention to lessons learned in the research concerning development of methodologically sound indicators of instruction; 2) focus on areas of instruction that are most amenable to measurement or have been measured reliably in the past, and; 3) develop these measure for use in a survey that focuses solely on instruction. Below, we discuss each recommendation in turn.

If NCES is to develop adequate descriptive measures of instruction, it must heed lessons learned in the past. Recommendations from researchers who have spent considerable time and resources on developing measures of instruction can form a framework for future development. For instance, response options must be developed carefully; teachers are unable to distinguish between “daily” and “once or twice a week” when reporting the frequency of particular practices (Burstein et al., 1995, p. 45). Also, measures of the frequency of practice are not usually reliable at the individual level. Reliability improves, however, when individual measures are analyzed and presented as composites (Mayer, 1999b). Most importantly, once measures have been developed, NCES must perform rigorous validity and reliability testing (Burstein et al., 1995).

The domains of instruction chosen for measurement must consider the success of previous research attempting to measure those areas, as well as the level of inference and interpretation necessary by teachers to report on their practice in those areas. Instructional content is the area most ripe for inclusion on teacher self-reports. As the knowledge to be learned, it is the most elemental aspect of instruction. Furthermore, instructional content has a demonstrated link to student achievement (Gamoran et al., 1997). Moreover, it has proven to be an area that teachers can report with relative ease and little interpretation (Burstein et al., 1995; Leighton et al., 1994). Porter and his colleagues have developed sophisticated measures of content in mathematics and science for elementary, middle, and high school grades that could be utilized on NCES surveys (see Blank et al., 2001; Gamoran et al., 1997; Smithson & Porter, 1994). The method Porter and his colleagues use to report findings collected with these measures, however, is also sophisticated and may prove awkward for a general audience. Measures of content in the TIMSS Opportunity to Learn questionnaires can also serve as a model of a sound, detailed inventory of content.

Other areas of instruction that require little abstraction and interpretation are instructional resources and the social organization of the classroom. Measuring these areas in addition to content would give audiences a description of what the classroom actually looks like, and what resources teachers in instruction. Furthermore, the unequal distribution of instructional resources may inhibit some teachers’ “opportunity to teach” (Brewer & Stasz, 1996, p. 13). Current NCES measures can be used as building blocks for these areas, but improvements must be made. The validity and reliability of measures must be tested, and measures of resources must be expanded beyond computers to include textbooks, workbooks, and other essential materials.

Development of measures of instructional strategies and classroom activities have proven troublesome to researchers. However, to present a comprehensive descriptive portrait of instruction, these must be measured. Current NCES measures of these areas must be expanded to account for the classroom context in which they take place. This entails couching strategy and activity items within fairly specific areas of content, with as much rich detail as possible on other aspects of classroom context.

Broad, generalizable measures should not be the only type of indicators employed. The "structure of last lesson" item from TIMSS, although not generalizable, contains so much contextual information that it is worth improving for inclusion in a survey of instruction. Although the content area covered is too specific to generalize to a teacher's larger body of instruction, this item would provide a useful and rich complement to a general descriptive portrait of the classroom. The lack of generalizability would be offset by the "meaningful chunks" of the heretofore unaddressed area of classroom interaction that it would provide (Leighton et al., 1994, p. 28).

The last recommendation for the development of adequate descriptive measures of instruction is to place all items in one instrument devoted solely to instruction. The IP items reviewed in this critique were found in long survey questionnaires devoted to multiple topics. Respondent burden has always been one of the most significant obstacles to valid and reliable measures of instruction (Burstein et al., 1995; Leighton et al. 1994). For this reason, we recommend an instrument that collects as much detail on instruction as possible without burdening the respondent to the point of providing inaccurate or unreliable responses. This is not possible as long as items on instruction are squeezed into surveys that focus on other areas, such as school staffing or general teacher background information.

Predictive measures of instructional processes

We suggest that NCES opt for the second recommended route: an intensive effort toward development of measures that are predictive of variation in student achievement. To achieve this goal, NCES must 1) reorient its conceptualization of instructional processes, building on knowledge that the IPRD program has accumulated and generated over the course of the past year, and 2) develop new data collection methods and tools that will allow researchers to capture the link between instruction and changes in student academic achievement.

Knowledge accumulated and generated over the last year by the IPRD program indicates that NCES must reconceptualize the instructional process in order to capture the actual quality of teachers' instruction, instead of merely describing variation in classroom practices and settings. An understanding of the quality of instruction must account not only for instruction as distributed by teachers, but also as received by students. Measurement of students' role in the learning process and the nature of how they experience the curriculum will provide a much richer portrait of instruction (Brewer & Stasz, 1996). Furthermore, students must be considered if the goal of measurement is to link to changes in their academic achievement. This effort will also require the development of new data collection tools, because surveys alone are not adequate for capturing instruction in a way that will be predictive of student achievement. A survey would entail too much burden if it asked about everything of interest (Brewer & Stasz, 1996). Furthermore, classroom activities are too dependent on context to capture with a survey instrument. Instruction varies day-to-day depending on so many factors that a survey cannot possibly capture all the necessary context (Brewer & Stasz, 1996).

A conceptual focus on student learning further increases the necessity of non-survey collection tools (Brewer & Stasz, 1996). Much of the literature on the development of instructional processes measures suggests using alternative tools, such as logs, teacher interviews, classroom artifacts, and observations, not just for validating surveys, but for actual data collection (Brewer & Stasz, 1996; Burstein et al., 1995; Leighton et al., 1994; Smithson & Porter, 1994). Furthermore, through the TIMSS Video study, NCES has established a precedent for incorporating observational tools toward the documentation of instructional processes (Baker et al., 1999).

We also suggest another methodological advancement for measuring instruction: classroom experiments. If a main purpose of IP measurement is to predict student achievement, NCES should attempt to uncover as large an effect as possible. Experiments allow the researcher to manipulate the independent variable—in this case, instruction—to better observe its effect on the dependent variable—in this case, achievement. There is a growing sentiment in the literature toward using experiments to understand what is happening—and what is successful—in schooling (Jencks & Phillips, 1998).

Accurate, reliable, generalizable, and predictive measurement of instructional processes is an intimidating task, but with the proper resources and determination, it can be realized. We reassert the recommendation made by Brewer and Stasz (1996) in a previous NCES report: “While these problems associated with answering the question ‘what promotes student learning?’ have led to few strong conclusions, it is probably premature and overly pessimistic to abandon this line of research altogether” (p. 5).

References

Baker, D.P., LeTendre, G., & Zhang, Y. (1999). *The measurement of instructional practices in TIMSS: Description and critique*. State College, PA: Pennsylvania State University.

Ball, D.L., Camburn, E., Correnti, R., Phelps, G, & Wallace, R. (1999). *New tools for research on instruction and instructional policy: A web-based teacher log*. Seattle, WA: Center for Study of Teaching and Policy.

Blank, R.K., Porter, A., & Smithson, J. (2001). *New tools for analyzing teaching, curriculum and standards in mathematics & science*. Washington, DC: Council of Chief State School Officers.

Brewer, D.J., & Stasz, C. (1996). Enhancing opportunity-to-learn measures in NCES data. In G. Hoachlander, J.E. Griffith, and J.H. Ralph, *From data to information: New directions for the National Center for Education Statistics* (NCES 96–901). U.S. Department of Education, National Center for Education Statistics. Washington, DC: U.S. Government Printing Office.

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

Cohen, D.K. (1990). A revolution in one classroom: The case of Mrs. Oublier. *Educational Evaluation and Policy Analysis 12*(3), 327–345.

Floden, R.E. (2000). *The measurement of opportunity to learn*. Washington, DC: National Academy of Sciences/National Research Council.

Gamoran, A., Porter, A.C., Smithson, J., & White, P.A. (1997). Upgrading high school mathematics instruction: Improving learning opportunities for low-achieving, low-income youth. *Educational Evaluation and Policy Analysis 19*(4), 325–338.

Henke, R.R., Chen, X., & Goldman, G. (1999). *What Happens in Classrooms? Instructional Practices in Elementary and Secondary Schools, 1994–95* (NCES 1999–348). U.S. Department of Education, National Center for Education Statistics. Washington, DC: U.S. Government Printing Office.

Jencks, C., & Phillips, M. (1998). *The black-white test score gap*. Washington, DC: Brookings Institution Press.

Leighton, M.S., Mullens, J.E., Turnball, K., Weiner, L.K., & Williams, A.S. (1994). *Measuring Instruction, Curriculum Content, and Instructional Resources: The Status* *of Recent Work* (NCES 95–11). U.S. Department of Education. Washington, DC: National Center for Education Statistics.

Mayer, D.P. (1999a). Invited commentary: Moving toward better instructional practice data. *Education Statistics Quarterly 1*(2), 17–20. Available at *http://nces.ed.gov/pubs99/quarterlyjul/2-feature/2-esq12-c.html*

Mayer, D.P. (1999b). Measuring instructional practice: Can policymakers trust survey data? *Educational Evaluation and Policy Analysis 21*(1), 29–45.

Mayer, D.P., Mullens, J.E., & Moore, M.T. (2000). *Monitoring school quality: An indicators report* (2001–030). U.S. Department of Education, National Center for Education Statistics. Washington, DC: U.S. Government Printing Office.

McDonnell, L.M. (1995). Opportunity to learn as a research concept and a policy instrument. *Educational Evaluation and Policy Analysis 17*(3), 305–322.

Mullens, J.E. (1995). *Classroom instructional processes: A review of existing measurement approaches and their applicability for Teacher Followup Survey* (NCES 95–15). U.S. Department of Education. Washington, DC: National Center for Education Statistics.

Mullens, J.E. (1998, October 31). *Validating teacher responses on instructional practice self-report surveys: Findings from a recent field test*. Paper presented at the Association for Public Policy Analysis and Management Research Conference, New York.

Mullens, J.E., & Gayler, K. (1999). *Measuring classroom instructional processes: Using survey and case study fieldtest results to improve item construction* (Working Paper No. 1999-08). U.S. Department of Education. Washington, DC: National Center for Education Statistics.

National Commission on Excellence in Education. (1983). *A nation at risk: The imperative for educational reform*. Washington, DC: U.S. Government Printing Office.

National Council of Teachers of English. (1996). *Standards for the English language arts*. Urbana, IL: Author.

National Council of Teachers of Mathematics. (1991). *Professional standards for teaching mathematics*. Reston, VA: Author.

Porter, A.C. (1995). *Developing opportunity-to-learn indicators of the content of instruction*. Madison, Wisconsin: Wisconsin Center for Education Research.

Porter, A.C., & Smithson, J.L. (1997). *Enacted curriculum survey items catalogue: Middle school and High School Mathematics and Science.* (2nd edition). Madison, Wisconsin: Wisconsin Center for Education Research.

Rowan, B. (2000). *Assessing teacher quality: Insights from school effectiveness research*. Ann Arbor, MI: University of Michigan.

Rowan, B. (2001). *What large-scale, survey research tells us about the effects of teachers and teaching on student achievement.* Consortium for Policy Research in Education, Study of Instructional Improvement, Research Note S-5. Ann Arbor: University of Michigan.

Rowan, B., Correnti, R., & Miller, R. J. (2002). What large-scale, survey research tells us about the effects of teachers and teaching on student achievement: Insights From the *Prospects* Study of Elementary Schools. *Teachers College Record* 104(8), 1525–1567.

Sanders, W. (2000). Schools and kids. Online discussion available at: *http://www.washingtonpost.com/wp-srv/liveonline/00/schools/schools0315.htm.*

Scheerens, J., & Bosker, R.J. (1997). *The foundations of educational effectiveness*. Oxford: Elsevier Science Publishers.

Shulman, L.S. (1987). Knowledge and teaching: foundations of the new reform. *Harvard Educational Review 57*(1), 1–22.

Smithson, J.L., & Porter, A.C. (1994). *Measuring classroom practice: Lessons learned from efforts to describe the enacted curriculum—the reform up close study*. Madison, WI: Consortium for Policy Research in Education.

Stecher, B.M., & Foley, B. (2000). *Developing survey measures of standards-based instructional practices in fifth grade mathematics: Progress report (June–December 1999)*. Santa Monica, CA: RAND.