Educational Policy
2017, Vol. 31(1) 73–107

DOI: 10.1177/0895904815586848
epx.sagepub.com
SAGE

Assessing the Validity of an Annual Survey for Measuring the Enacted Literacy Curriculum

Eric M. Camburn[1], Seong Won Han[2], and James Sebastian[3]

Abstract

Surveys are frequently used to inform consequential decisions about teachers, policies, and programs. Consequently, it is important to understand the validity of these instruments. This study assesses the validity of measures of instruction captured by an annual survey by comparing survey data with those of a validated daily log. The two instruments produced similar rankings of the frequency with which teachers use particular practices but more than three fourths of the teachers in the study were found to overreport their instruction on the annual survey. Multilevel models revealed a number of teacher and school characteristics related to survey reporting error. The study's implications for users of survey evidence are discussed.

Keywords

teacher quality, survey research, classroom instruction

There is a widespread recognition that students' instructional experiences are one of the most important influences on their learning. In the United States,

[1]University of Wisconsin–Madison, WI, USA
[2]University at Buffalo, The State University of New York, NY, USA
[3]University of Missouri, Columbia, MO, USA

Corresponding Author:
Eric M. Camburn, University of Wisconsin–Madison, 253D Education Building, 1000 Bascom Mall, Madison, WI 53706, USA.
Email: ecamburn@education.wisc.edu

policymakers and educational leaders are increasingly interested in collecting valid evidence of the quality of instruction in hopes that it helps them understand how well teachers are teaching and teachers' influence on students. Self-report surveys are commonly used for these purposes (Camburn & Han, 2011; Desimone & Le Floch, 2004; Mayer, 1999). Self-report surveys have a number of distinct advantages over other strategies of measuring instruction—they are less expensive and less burdensome than most other options, and they are particularly well suited at demonstrating how instruction varies across large numbers of students, teachers, and settings. There is ample evidence of the validity of self-report surveys for measuring teaching and teachers' work from both validity studies (Burstein et al., 1995; Camburn, Huff, Goldring, & May, 2010), and empirical investigations of teaching and factors related to it. However, there is also evidence that casts doubt on the validity of measures of instruction from self-report surveys. For example, Desimone and Le Floch (2004) found that some teachers in their sample had difficulty understanding concepts asked about in survey items, and Mullens et al. (1999) found survey reports of learning objectives, classroom activities, and the use of instructional materials to be less accurate than daily log reports of these same aspects of instruction. The body of research that has examined the validity of teacher surveys is quite small and dated and has other significant limitations, including the use of small teacher samples, and relatively little attention given to factors related to survey inaccuracy. In addition, very little of this research is grounded in the well-established literature on survey research methods. The application of this literature has considerable potential for deepening our understanding of the validity of surveys used in education research as it provides useful explanatory frameworks for why and how survey response errors occur.

This study addresses these limitations by assessing the validity of measures of instruction from an annual survey from a large sample of elementary school teachers. We assess the instrument's validity by comparing data from the survey with those of a validated daily log administered to the same sample of teachers during the same time period. The two instruments measure the same dimensions of instruction and were designed to be compared in this fashion. This strategy of validating a self-report survey covering a long time period (e.g., an entire school year) against a more frequently completed instrument (e.g., a daily log, daily diary, or experience sampling instrument) is common in many social science disciplines. Based on prior research and cognitive perspectives on the survey response process, we conjectured that teachers' reports of their instructional practices on the instruction log and annual survey might differ markedly (Hilton, 1989; Lemmens, Knibbe, & Tan, 1988; Lemmens & Tan, 1992; Smithson & Porter, 1994). In addition to estimating the overall direction and degree of reporting error in measures of

instruction from the annual survey, this study also examined whether error in teachers' survey reports was more prevalent among certain types of teachers or in different types of schools.

Background

We begin with the premise that it is important for educational policymakers and researchers to understand how instruction varies across groups of people (teachers and students) and settings (e.g., classrooms, schools, geographic locations). Self-report surveys, a ubiquitous tool in education, are well suited for producing this sort of evidence. We distinguish between surveys that are administered *only once*, or administered *annually* as part of a longitudinal study, and closed-ended surveys that are completed *daily*. Strictly speaking, both kinds of instruments are self-report "surveys." However, the time frames of these two instruments differ dramatically, with the former requiring teachers to report on periods of a year or more and the latter covering a single day. To distinguish between these two kinds of self-report surveys, we adopt the language of prior studies and refer to one-time and annual surveys as "surveys" and daily instruments as "logs."

Use of Teacher Surveys in Educational Research

Among the various strategies commonly used to measure instruction, surveys are particularly well suited to revealing how instruction varies across groups and settings, in part because they are less expensive than classroom observations, logs, and videotaping. Because of their relatively low cost, surveys can be collected from larger samples, thus permitting analyses of broad patterns of instruction, statistical comparisons of group differences, and examination of relationships between instruction and other variables. Surveys administered to probability samples can be especially informative, because results from such surveys are generalizable to known populations (see, for example, Camburn & Han, 2011, who document a wealth of such evidence on instruction in the United States generated by studies that use surveys administered to nationally representative probability samples). There are numerous studies in which survey measures of various aspects of teaching and teachers' work have been found to be associated with other variables in theoretically predicted ways (see, for example, Bryk & Schneider, 2002; Camburn & Han, 2011; Cybulski, Hoy, & Sweetland, 2005; Milesi & Gamoran, 2006). This body of research provides evidence of the predictive validity of a variety of measurements from teacher surveys and underscores the potential of this tool for providing insight into education problems.

Survey data on how instruction varies across groups and settings inform policy conversations about teaching in a variety of ways. Data from nationally representative surveys are widely used to understand the kinds of instructional experiences that promote student learning and to understand how opportunity to learn (OTL) varies by student subgroups (Camburn & Han, 2011; Schmidt & McKnight, 1995). For example, evidence from national surveys is routinely used by the federal government to describe instructional patterns and the conditions of classroom instruction across the country (Mullens et al., 1999). Survey data on classroom instruction are also used to compare countries with one another and explain national differences in test performance (see, for example, Mullis, Martin, Foy, & Arora, 2012; Organisation for Economic Co-Operation and Development [OECD], 2013a, 2013b).

The validity of group comparisons based on survey data rests on the assumption that the survey measures all groups with relatively equal validity. When survey data are used for consequential purposes, it is important to gauge whether this assumption is met. Other researchers have argued that examining how the reliability and validity of survey data vary across settings and groups of people will help improve our understanding of the quality of data being used to guide policy and reform (Desimone, 2006; Desimone, Smith, & Frisvold, 2010). Desimone (2006) found that survey measurements of policy constructs varied considerably between teachers, principals, and district administrators. Desimone (2006) observed that these differences could be related to valid differences in the perceptions and experiences of the members of these three groups but might also reflect differential biases or errors in reporting. Clearly, the assumption of equal validity across groups does not hold if Desimone's latter conjecture is true. If systematic reporting errors are observed for a survey instrument, identifying individual or organizational characteristics that explain these errors could be useful for guiding subsequent statistical analyses. For example, it might be possible to adjust estimates for individual and organizational characteristics that are found to be associated with reporting error.

Because educational policy changes are often directed toward changing classroom instruction, the measurement of instruction at scale becomes critical to studying implementation and effectiveness of educational policies and reforms (Desimone & Le Floch, 2004; Desimone et al., 2010; Mayer, 1999). The research of Bryk, Sebring, Allensworth, Luppescu, and Easton (2010) is a good example of how survey data can be used for this purpose. These researchers used multiple measures of classroom instruction to understand differences in school performance over time and found instruction to be one of five essential elements for school success. As an outgrowth of this work,

principals of Chicago Public Schools regularly receive school-specific reports of strengths and weaknesses on multiple factors including classroom instruction, which they can utilize to guide assessment, policy decisions, and subsequent reform (Luppescu et al, 2007). In another recent example, researchers conducting the Measures of Effective Teaching (MET) study examined the use of student survey reports of classroom instruction to inform evaluation and accountability systems (Bill & Melinda Gates Foundation, 2010). In sum, surveys play a prominent role in research on classroom instruction and can be important tools in guiding policy discussions on issues related to learning and instruction. Because survey data are used for consequential decision-making by educational policymakers, practitioners, and researchers, we believe having a better understanding of how well surveys measure teaching is both timely and important (Desimone et al., 2010).

Prior Research on the Validity of Teacher Surveys

Inferences made from survey data are more valid to the extent that the data accurately represent intended instruction constructs. In this study, we focus on the validity of survey instruments for measuring the *enacted curriculum in reading comprehension and writing in third grade.* Porter (2002) defined the construct *enacted curriculum* as the amount of instructional time devoted to teaching various strands and/or topics in the school curriculum. This study operationalized the enacted third-grade literacy curriculum with 15 items on which third-grade teachers reported specific ways their students worked on the broad topics of reading comprehension and writing. The 15 items were asked on an annual survey and a daily log, thus permitting the validity study reported here.

A common way of assessing the validity of survey data is to compare it with a benchmark data source that has significantly less measurement error. When the survey and benchmark data are collected during the same time period, this kind of analysis is called a test of *concurrent validity*. Potential benchmarks include evidence from other data collection instruments, existing records (medical records, administrative records, transcripts), and physical measurements (food intake, blood tests). The evidence from studies that have attempted to validate annual style surveys is mixed. A small number of concurrent validity studies, using daily logs and classroom observations as benchmarks, provide evidence supporting the validity of teacher self-report surveys. For example, Mullens et al. (1999) studied 41 middle and high school mathematics teachers using mail surveys, classroom observations, teacher logs, and interviews. Data from these sources were tightly integrated, as both teachers and observers completed instructional logs, follow-up

interviews referencing classroom observations were conducted, and logs were completed by teachers on the day of observation. These researchers found a substantial agreement between teachers' reports of their mathematics instruction on self-administered mail surveys and classroom observations and a strong agreement between surveys and logs for questions measuring the duration of instructional events. Similarly, Mayer (1999), who compared teachers' self-administered survey answers with observation data, concluded that the survey provided a fairly accurate picture of the duration of reform-oriented instructional practices in mathematics.

Other concurrent validity studies in education show that surveys can capture content coverage with considerable accuracy. Burstein et al. (1995) collected homework, quizzes, classroom exercises, projects, and exams from a sample of 70 high school mathematics teachers. Teachers in this study also completed daily logs for 5 weeks. Benchmarks of teachers' instructional activities were created by coding these artifacts, and this information was compared with teachers' survey responses. These researchers found survey measures of instructional content and instructional strategies provided a generally valid picture of classroom instruction, particularly at higher levels of generality (i.e., broad content areas as opposed to specific topics). Among 62 high school math and science teachers, Smithson and Porter (1994) similarly found that correlations between teachers' reports of broad content areas on logs and classroom observations were quite high, ranging between .60 and .93 for most constructs.

However, this same small set of benchmarking studies also produced evidence raising doubts about the accuracy of teacher self-report surveys. For example, based on their comparisons of survey results with coded artifacts, Burstein et al. (1995) concluded that surveys should not be used to measure teachers' instructional goals. Mayer (1999) found evidence that surveys did not provide good measures of the quality of interactions between teachers and students. Smithson and Porter (1994) found that correlations between measures of instructional activity from logs and surveys to generally be quite low—Most were .40 or lower.

There are many factors which may undermine the validity of teachers' self-reports on surveys. Our reading of the literature suggests three factors that may be particularly detrimental—Teachers may fail to understand survey questions, they may perceive social pressure to answer surveys in particular ways, and they may be unable to accurately recall and report what they do on a survey. A cognitive model of the survey response process and research on the role of memory in answering surveys, both drawn from the broader survey methodology literature, provide useful perspectives for understanding these sources of invalidity in teachers' survey responses.

Tourangeau and Rasinski (1988) articulated a general cognitive model for answering survey questions. The model stipulates that respondents go through four steps when answering a survey item: (a) they attempt to *understand* the meaning of the question, (b) they form a *judgment* based on this understanding, (c) they *format* their response in accordance with response alternatives presented in the question, and (d) they *edit* their response before communicating it, reflecting considerations of social desirability and self-presentation. Thus, before respondents can provide an answer on a survey they must first understand what the question means. At the most basic level, respondents must form an understanding of the literal meaning of the words used in the question. But beyond this literal understanding, respondents will also form an understanding of the researcher's intended meaning behind a question. Indeed, two studies have provided evidence that teachers' and researchers' definitions of language used in surveys sometimes differ, and when this occurs, the validity of teachers' responses is undermined (Desimone & Le Floch, 2004; Hill, 2005).

According to Tourangeau and Rasinski's (1988) model, after respondents have formulated an understanding of what a survey question is asking, they formulate an answer by first forming a judgment about what would be an appropriate response, and then editing the response before giving an answer. Respondents edit survey answers for social desirability and self-presentation in response to norms, incentives, and pressures in their social setting (Tourangeau, Rips, & Rasinski, 2000). Socially desirable responding that reflects how teachers think they *ought* to be teaching rather than how they *actually* teach is thus another potential source of error. In this study, three quarters of the schools in the sample were participating in a comprehensive school reform (CSR) program, which, in some cases, were implemented in strong accountability contexts where teachers may have perceived pressure to report instructional practices advocated by the reform programs. Our statistical analyses examine this issue.

Difficulty recalling past behaviors or events is a leading cause of measurement error in self-report surveys and is the primary source of error examined in this study. When reporting how often they engage in a behavior like teaching, respondents must search their memories of past events. Such memories are either stored with event-specific information (who, what, where), or as generic "categories of events and stereotypical sequences of events" (Tourangeau et al., 2000, p. 69). Event-specific information is believed to be stored in *episodic memory*, whereas generic information about events is stored in *semantic memory* (Tulving, 1983). Survey responses drawn from episodic memory are believed to be more accurate than responses based on semantic memory.

According to Menon (1994), the regularity of a behavior and its similarity from one occurrence to another affect the retrieval strategies used by respondents when answering survey questions about that behavior. Generally speaking, behaviors that occur more often are more likely to be stored in semantic memory, and less common behaviors are more likely to be stored in episodic memory. Menon's model further stipulates that behaviors that are similar from one occurrence to the next are more likely to be recalled from semantic memory, whereas behaviors that are less similar during each occurrence are more likely to be recalled from episodic memory. There is a considerable empirical support for Menon's model across a range of disciplines (Anderson, 1983; Linton, 1988; Thompson, Skowronski, Larsen, & Betz, 1996; Wagenaar, 1986; White, 1982).

The amount of time that elapses between an event and a reporting of the event on a survey also affects how memory is accessed and the accuracy of the survey response because longer reference periods require respondents to consider both greater amounts of time and more events (Tourangeau et al., 2000). In general, the shorter the elapsed time between an event and a survey response, the more likely it is that respondents will be able to directly access episodic memory. Conversely, the further away in time a survey response gets from the event, the more likely it is that respondents will rely upon semantic memory or estimation strategies. Consistent with these cognitive perspectives, daily logs and diaries, that are completed closer to when a behavior or an event occurs, have been found to be more accurate than surveys (Hilton, 1989; Lemmens et al., 1988; Lemmens & Tan, 1992).

Considering descriptive results from the daily log used in this study in light of these cognitive perspectives suggests how recall failure might cause errors in reports of instruction on annual surveys. Based on daily log data, we estimate that teachers report approximately 15 distinct instructional practices in literacy per day. This suggests that in completing a daily instrument, teachers may need to recall up to 15 different instructional events. Extrapolating this estimate to the typical 140-day school year that comprises the reporting period for the typical annual survey yields approximately 2,100 instructional events that teachers would have to consider in formulating a response to such a survey. Bear in mind that this is an estimate for a single student. Teachers' task of recalling instructional events is even more complicated to the extent that they differentiate instruction among students. Even if this is an overestimate (teachers may have used multiple log items to describe a single instruction event), it seems very likely to us that in reporting their instruction on an annual survey, teachers draw heavily upon less accurate semantic memory and use estimation strategies. Conversely, in reporting their instruction on a daily log, we believe that teachers will be more likely to directly access their

memories of specific instructional events throughout the day, thus yielding more accurate reports on the daily instrument.

While the existing studies provide important insight into the validity of surveys for measuring instruction, they also have limitations that are addressed by this study. None of the prior studies gauge the validity of survey reports of reading, English or language arts instruction; all were conducted with fairly small samples of teachers; the studies do not provide precise estimates of the magnitude of survey reporting error and finally, none of the studies examine teacher, classroom, and school characteristics associated with measurement error. Addressing these limitations, this study investigates two research questions:

> **Research Question 1:** How consistent are teacher reports of the enacted curriculum from daily logs and annual surveys?
> **Research Question 2:** What teacher, classroom, and school characteristics are associated with differences in annual survey and daily log reports of the enacted curriculum in reading comprehension and writing instruction?

In addressing these questions, we seek to strengthen the knowledge base on the validity of using surveys with long reference periods for measuring instruction.

Method

This study uses data from a sample of 245 third-grade teachers in 103 public elementary schools. Table 1 describes the teacher and school samples. Sample schools are located in large (central city population of 250,000 or more) and medium (central city population of less than 250,000) metropolitan areas comprised of a central city and the urban fringe surrounding the city. Compared with schools across the nation located in such metropolitan areas, sample schools serve higher percentages of African Americans and students receiving free/reduced-price lunch, and lower percentages of White students (Rowan & Miller, 2007). Sample schools are also more highly concentrated in the central cities of large metropolitan areas.

Data for this study come from an investigation of three widely adopted CSRs: Accelerated Schools Project (ASP), America's Choice (AC), or Success for All (SFA). Three fourths of sample schools participated in these programs, while the remaining schools did not. The schools that did not participate in the three CSR programs are located in the same districts as the CSR participants and are demographically similar.

Table 1. Characteristics of Teachers and Schools.

Variable	*M*	*SD*	Minimum	Maximum
Teachers (*n* = 245)				
Female	0.93	0.25	0	1
Hispanic	0.09	0.29	0	1
African American	0.22	0.41	0	1
White	0.60	0.49	0	1
Other race	0.09	0.29	0	1
Years of teaching experience	12.46	10.18	1	35
Content knowledge for literacy teaching	0.05	0.76	−1.72	2.04
Teacher individualizes instruction	0.68	0.47	0	1
Class size	19.9	5.69	3	36
Average classroom achievement	599.71	25.02	479.77	665.05
Schools (*n* = 103)				
School SES	−0.02	0.99	−2.43	1.75
Percentage of students eligible for free and reduced-price lunch	75.85	21.4	4	100
Percentage of non-White students	79.02	27.07	1	100
Average school achievement level	549.67	15.55	512.14	582.56
Accountability pressure	0.03	1.04	−2.21	2.89
CSR: Accelerated Schools project	0.26	0.44	0	1
CSR: America's Choice	0.24	0.43	0	1
CSR: Success for All	0.25	0.44	0	1

Note. SES = socioeconomic status; CSR = comprehensive school reforms.

While the CSR programs are not a major substantive focus of this study, we believe that variation in CSR participation among sample schools provides a useful window for examining reform implementation as a contextual variable that might influence teachers' reporting of their instruction. One way this might occur is through personal participation in the program. Participating in a CSR program, particularly one that advocates specific literacy instruction practices, might make a teacher more aware of those practices and might, in turn, affect the accuracy of their reports on an annual survey. A second way in which working in a school implementing a CSR program might affect teachers' survey responses is that reporting practices advocated by the CSR program might be perceived as socially desirable by teachers.

These unique characteristics of sample schools mean that the study's results cannot be generalized to the nation as a whole, and may also not generalize particularly well to the population of all urban schools. Instead, our

results are most directly applicable to elementary school teachers in public schools located in large urban and disadvantaged settings.

Instruments

The annual teacher survey that is the focus of this study is a self-administered questionnaire completed by teachers at the end of the school year. The survey is a 28-page booklet that took approximately 60 min to complete. The survey design and data collection procedures were modeled after Dillman's (1991) *Total Design Method*, which outlines empirically tested procedures that have been shown to reliably yield high response rates and reduce measurement error. Procedures used in this study included the following: advance notification letters, multiple questionnaire mailings, the inclusion of questions that were salient to teachers, and the use of questionnaire formatting that eased the task of understanding and answering questions. Response rates for the teacher survey varied slightly from year to year but averaged about 72%. This study provides the first concurrent validity test of the annual survey. A number of studies that have utilized data from the annual teacher have found theoretically predicted, statistically significant relationships between variables measured by the survey, thus providing evidence of the predictive validity of the instrument (Camburn & Han, 2009; Camburn, 2010).

Like previous studies in education (Burstein et al., 1995; Mullens et al., 1999; Smithson & Porter, 1994) and other fields (Hilton, 1989; Lemmens et al., 1988; Lemmens & Tan, 1992), we used data from a daily instrument as a benchmark for assessing the validity of the teacher survey. The daily language arts log used for this purpose is a four-page self-administered questionnaire on which teachers reported the instructional experiences of a target student for a single school day. A random sample of eight students was selected and recruited from each classroom. Teachers were trained by field staff to use the log and were provided with a 26-page glossary containing item-by-item definitions of terminology used in the log. Response rates for the instructional log ranged from 82% to 92% over the course of the study, with an average of 85%. The annual teacher survey, daily logs, and log glossaries can be downloaded from this website http://sii.soe.umich.edu/instruments/.

Teachers selected for analysis had to have completed a teacher questionnaire *and* at least two daily instructional logs. On average, teachers in the sample completed 41.7 logs and only 9% of our analytic sample completed less than 10 logs.

Rationale for Validity Analyses

We assessed the validity of an annual survey for measuring the enacted third-grade literacy curriculum through a test of concurrent validity in which data from the annual survey were compared with benchmark data from a daily log. We offer four reasons for using data from the daily log as a validation benchmark. First, there is considerable evidence that the daily log accurately measures the enacted curriculum in third-grade classrooms, thus indicating the construct validity of evidence from the daily log. A rigorous validation of the daily log was conducted as part of this study's parent project, the Study of Instructional Improvement (SII; Camburn & Barnes, 2004). In the validation study, teachers were observed by two researchers on a day on which they completed a log. In addition to recording a narrative of classroom activity, observers also completed daily logs themselves. At the end of the day, observers and teachers participated in a follow-up interview in which differences between teachers' and observers' logs were discussed. Evidence from classroom observations and follow-up discussions with teachers provided direct evidence of the enacted curriculum on observation days. Two published reports of the log validity study found a generally strong correspondence between this direct evidence and evidence from teachers' daily logs (Camburn & Barnes, 2004; Hill, 2005). Based in part on the log validity study results, SII researchers chose daily logs over third-party observations to measure the enacted curriculum in a large-scale investigation of instructional improvement in urban schools (Rowan, Camburn, & Correnti, 2004).

Second, in developing the daily log, considerable steps were taken to use language that would be commonly understood by teachers. During pretesting of the daily log, teachers' understanding of the meaning of terminology used in log items was assessed in a series of focus groups, and numerous items were revised as a result.

Third, the annual survey and daily log were designed to measure the enacted literacy curriculum in third grade, and were designed to conduct the comparative analyses undertaken for this study (Rowan et al., 2004). A set of 15 common items measuring the enacted literacy curriculum were included in both instruments. With only a few exceptions, the wording of items was exactly the same on both instruments. Teachers also reported on comparable groups of students on the two instruments. On the teacher survey, teachers reported on a target class to whom they taught reading, while on the log, they reported on a random sample of students to whom they taught reading. The 15 questions common to both instruments are listed in Table 2.

Finally, Menon's (1994) cognitive perspectives on survey response and a considerable body of evidence supporting those perspectives provide further

Table 2. Correlations Between Teachers' Annual Survey and Daily Log Reports for 15 Items.

Item	Correlation
Wrote letters, strings, or words	.19
Worked on a literature extension project	.34
Wrote extensive answers to questions	.29
Did a thinkaloud or explained skill or strategy use	.34
Revised writing through elaboration	.31
Revised writing by refining or reorganizing	.36
Generated their own questions	.23
Edited word use, grammar, or syntax	.33
Worked on concept maps, story maps, text structure, frames	.36
Edited capitals, punctuation, or spelling	.26
Wrote brief answers to questions	.27
Summarized important details	.20
Answered questions requiring inferences	.21
Activated prior knowledge	.17
Answered questions with answers directly stated in text	.32

Note. All correlations are significantly different than 0 at the .05 level of less.

support for our stance that daily logs provide a reasonable benchmark for assessing the validity of an annual survey. Consistent with Menon's ideas about the influence of a survey's recall period on memory retrieval, a number of studies have found that daily logs and diaries that are completed closer to when a behavior or an event occurs are more accurate than surveys with longer reporting periods such as the annual teacher survey under study (Hilton, 1989; Lemmens et al., 1988; Lemmens & Tan, 1992).

Like any measurement strategy, however, daily logs are not immune to reporting errors. For example, Unge et al. (2005) found a good agreement for the occurrence of work tasks between diaries and observations but found that brief tasks were underreported on diaries. In assessing the validity of the daily log, Camburn and Barnes (2004) found that teachers sometimes did not report frequently occurring classroom activities on the log. Despite these limitations, we believe that the daily log provides a useful benchmark against which to evaluate the teacher survey.

Data from the two instruments were recoded to place them on comparable scales. Daily log data were recoded in two steps. First, log reports were aggregated into a single data record per teacher resulting in a set of items, which measured the proportion of days on which a teacher checked a particular item. These proportions were then multiplied by 20 (the typical number of

instructional days per month) yielding items measuring the number of instructional days per month that teachers used a particular practice. Annual survey items were placed on a comparable "days per month" response scale by recoding the original ordinal scale as follows: "never" = 0, "less than once a month" = 1, "1-3 times per month" = 2, "1-2 times per week" = 6, "3-4 times per week" = 14, and "everyday" = 20.

There is some evidence that recoding procedures such as those used for the annual survey data might affect the variance of the imputed variable and subsequently reduce its correlation with other variables (Heeringa, Little, & Raghunathan, 2002). For this study, reduced correlations between teachers' annual survey and daily log reports would underestimate the amount of agreement between the two instruments. To test whether our recoding of annual survey data had this effect, we fit two multilevel models, which nested teachers' mean responses to multiple daily log items within teachers within schools. The outcome variable for both models was thus the proportion of days teachers used various instructional practices. In one model, we used teachers' raw responses to annual survey items as a predictor of their mean log response, and in another model we used the imputed version of the annual survey items as a predictor. The results of these two models were nearly identical leading us to conclude that our choice of imputation procedures did not substantially affect the estimation of the difference in teachers' responses to the two instruments and inferences drawn about those differences.

Statistical Analysis

To investigate factors associated with measurement error in teachers' survey reports, we fit two-level hierarchical linear models (HLMs). Variables included in statistical models are described in Appendix A. The outcomes in these models are the differences in teacher reports between the teacher survey and daily log for the 15 items measuring the enacted curriculum. Visual inspection of the outcome variables revealed that they were positively skewed because of the relatively high proportion of teachers whose annual survey and log responses differed by less than 5 days per month. Despite this skewness, the outcome variables still had a generally normal shape. The HLMs estimated the differences in individual teachers' responses on the daily log and annual surveys as a function of teacher and classroom characteristics (Level 1) and school contextual characteristics (Level 2). The specific teacher-level characteristics we used were teachers' gender, race, teaching experience, a measure of their content knowledge, and whether they individualized instruction. The classroom characteristics we examined included class size and the average achievement level of the class. The school

characteristics we examined were school average socioeconomic status (SES), average achievement, a measure of the accountability pressure faced by the school, and the school reform program a school participated in. The statistical equations for these models are included in Appendix B.

Predictors for all models were centered around their grand means, which means that model intercepts represent the expected outcome for a teacher whose value on all predictors is equal to the grand mean of those predictors. In addition, all continuous predictor variables were standardized ($M = 0$, $SD = 1$), so that coefficients can be interpreted as standardized regression coefficients. We chose this metric because it permits direct comparisons of the sizes of the effects of different independent variables that are measured on different scales.

In addition to the HLM analyses, we also present descriptive statistics which illustrate patterns of discrepancies between teachers' reports on the survey and log.

Results

We first attempted to understand the general patterns in the differences between teachers' reports on the annual survey and daily log by correlating teachers' responses on the two instruments. Table 2 presents correlations for each of the 15 items measuring the enacted third-grade literacy curriculum. From these correlations, we can see if teachers, who reported using a particular instruction strategy more often, also tend to report doing that same strategy more often on the daily log. Correlations of teachers' reports of the enacted curriculum in third grade on the two instruments were all positive, and low to moderate in magnitude, ranging from .17 for the item "Activating prior knowledge" to .36 for the items "Worked on concept maps, story maps, text structure, frames" and "Revised writing by refining or reorganizing." The size of the correlations is similar to that of correlations between teacher logs and surveys reported by Smithson and Porter (1994). The fairly wide range of the correlations suggests that discrepancies between teachers' survey and log reports vary considerably across instructional practices. From this first analysis, we conclude that teachers' reports of their instruction on the annual survey are only modestly related to their reports on the daily log.

While the correlations show the general degree of correspondence between teachers' reports of their instruction on the two instruments, they do not indicate the direction or the magnitude of the differences in teachers' reports. We conducted additional descriptive analyses to get a better understanding of the practical implications of the modest correlations between the two instruments.

For each of the 15 instructional practices measured, Table 3 displays the estimated means from the annual survey and daily log, the mean difference between teachers' reports on the two instruments, the percentage of teachers whose log and survey reports matched exactly, and the percentages of teachers whose survey reports under- and overreported their instruction when compared with their log report. To highlight differences between the two instruments, items are sorted in ascending order of the means of our benchmark instrument, the instructional log. Table 3 makes three points readily apparent: (a) the two instruments rank order instructional practices similarly with respect to their estimated frequency of occurrence, (b) the survey systematically overstates the frequency of *all* instructional practices measured, and (c) the degree of overstatement varies considerably from practice to practice.

One of the most common uses of survey measures of instruction is to compare the incidence of different instructional practices. In their review of two decades of research on instruction from large-scale surveys, Camburn and Han (2011) found that more than half of the studies they reviewed relied exclusively on descriptive statistics to describe the incidence of different practices. Table 3 shows that estimates from the annual survey and daily log rank order instructional practices with respect to frequency very similarly. For example, we found that both instruments indicated that the third-grade students in the classrooms of sampled teachers spent the most time activating prior knowledge and answering questions that have answers directly stated in the texts they read. Similarly, both the daily log and the survey estimated that three kinds of writing tasks—having students write letter strings and words, write extensive answers to questions, and produce written literature extension projects—were the least frequently implemented practices among the third-grade teachers in the sample.

Despite these similarities, the survey consistently overstated the frequency of all instructional practices measured. For all 15 items, a majority of teachers overreported their instruction on the annual survey, and for every item, the difference between the mean log response and the mean survey response is statistically significant.[1] Overall, more than three quarters of all teachers (77%) overreported their instruction on the survey with percentages for specific instructional practices ranging from 63.35% to as high as 90.39%. Overall, we estimate that teachers' annual survey reports overstate the frequency with which they use a particular practice by 4.41 days per month with estimates for specific items ranging from approximately 3 to 5 days per month. For example, on the daily log, teachers said students answered questions that have answers directly stated in the text and activated prior knowledge approximately 8 days per month. These same teachers reported on the annual survey that their students engaged in these

Table 3. Descriptive Statistics by Item.

Item	Days per month–instruction log (*M*)	Days per month–annual survey (*M*)	Days per month–instruction log (*SD*)	Days per month–annual survey (*SD*)	*M* difference in survey and log responses	% exact match	% underreporting	% overreporting
Wrote letters, strings, or words	0.95	5.81	1.97	6.96	4.82	18.15	18.51	63.35
Worked on a literature extension project	1.46	4.46	2.23	4.75	2.99	6.76	19.93	73.31
Wrote extensive answers to questions	1.58	6.65	2.43	4.90	5.05	2.85	6.76	90.39
Did a thinkaloud or explained skill or strategy use	3.14	8.91	3.62	6.02	5.74	2.49	13.52	83.99
Revised writing through elaboration	3.59	7.66	3.55	5.55	4.05	1.42	17.79	80.78
Revised writing by refining or reorganizing	3.77	7.18	3.48	5.69	3.38	1.42	23.84	74.73
Generated their own questions	3.84	8.40	4.57	6.47	4.55	2.49	25.27	72.24
Edited word use, grammar, or syntax	3.98	8.50	3.70	6.05	4.51	1.78	18.86	79.36
Worked on concept maps, story maps, text structure frames	3.99	8.18	4.34	5.86	4.17	1.42	19.93	78.65
Edited capitals, punctuation, or spelling	4.77	9.66	3.97	6.30	4.87	1.42	19.57	79.00
Wrote brief answers to questions	5.74	10.77	4.53	5.24	5.00	0.36	17.08	82.56
Summarized important details	6.32	11.47	4.47	5.50	5.13	0.00	22.06	77.94
Answered questions requiring inferences	7.31	11.28	4.68	5.69	3.93	0.71	27.76	71.53
Activated prior knowledge	7.90	12.19	5.11	6.27	4.28	0.71	25.98	73.31
Answered questions with answers directly stated in text	8.06	11.71	4.72	5.88	3.63	0.36	28.11	71.53

activities approximately 12 times per month. These results are inconsistent with those of Mullens et al. (1999) who found that when teachers' survey and log reports disagreed, survey reports tended to *understate* the frequency of instructional practices. There are considerable differences between our study and that of John E. Mullens et al. (1999), which might account for the differences in findings. The two studies examined instruction in different academic subjects (we focus on literacy, Mullens et al., 1999, focused on mathematics) and education levels (we focus on third grade; Mullens et al., 1999, focused on Grades 8-12). Reporting reference periods for the two studies also differed. For this study, the reporting period for the survey was a year but was a single semester for Mullens et al. (1999). And for this study, the reporting period for the logs was three, 6- to 8-week periods spread throughout the school year, whereas the log reporting period for Mullens et al. (1999) was a single 4-week period.

The descriptive analyses in Table 3 also show that the degree to which annual reports of instruction differed from daily reports varied considerably from practice to practice. The percentages of teachers who under- and overstated different teaching practices on the survey were also quite variable. While much of this practice-to-practice variation appeared unsystematic, the percentage of teachers whose log and survey answers matched exactly *did* systematically vary along with the frequency with which practices occurred (as measured by the log). Specifically, we found that the percentage of teachers whose answers on both instruments matched exactly decreased as the frequency with which an instructional practice occurred increased. This result suggests that teachers' survey reports of *less* frequently occurring practices might have been more accurate than their reports of *more* common practices which is generally consistent with cognitive perspectives on the survey response process, which stipulate that less common behaviors are more likely to be stored in episodic memory and thus more accurately recalled (Menon, 1994; Tourangeau et al., 2000). The significance with which we regard this result is tempered given the relative rarity of teachers providing the same answers on both instruments.

Tables 2 and 3 shed light on the direction and magnitude of differences between teachers' answers on the annual survey and daily log but leave many questions unanswered. A significant limitation of these results and those of the other benchmarking studies previously discussed is that they do not address whether survey reporting error is more likely for particular kinds of teachers or particular kinds of schools. The next stage in the analysis was to examine whether teachers with different background characteristics and from different kinds of schools were more or less likely to over- or underreport their teaching practices on the annual survey. To address this question,

separate HLMs were fit for each of the 15 items measuring the enacted curriculum. The results of these models are displayed in Tables 4, 5, and 6.

We first used unconditional models (no predictors at any level) to examine what proportion of the variation in discrepancies in teachers' survey and log reports were associated with teachers themselves and what proportion was connected with the schools in which they worked (Table 4). For many items, we found that knowing what school a teacher worked in accounted for very little of the discrepancies between their answers on the survey and daily log. For example, for 5 of the survey items less than 5% of the variation in discrepancies between the two instruments was associated with teachers' schools. In contrast, for 3 of the 15 items, survey/log discrepancies varied substantially from school to school. For teachers' reports of how much their students revised their writing by refining and reorganizing, refined their writing through elaboration, and answered questions that required inferences, 23% or more of the variation in reporting discrepancies was attributable to teachers' work environments rather than their personal characteristics. These results suggest to us that mis-reporting instruction on surveys likely has multiple causes, some of which may be associated with certain school contextual conditions, while others may be related to the background, beliefs, and motivations of the teachers completing the survey. These results also suggest to us that the factors that influence reporting error on teacher surveys do not affect all aspects of instruction equally.

Teacher and Classroom Characteristics Associated With Survey Reporting Discrepancies

We next investigated whether discrepancies between teachers' annual survey and daily log reports were greater for different kinds of teachers and classrooms. Statistically significant effects for a subgroup might indicate that teachers in the subgroup had greater difficulty accurately remembering and reporting their practices on the annual survey. Significant effects might also indicate that reports of subgroup members were more affected by socially desirable response bias. When interpreting the findings of the statistical models presented here, it is worth bearing in mind that *null* results are desirable because they indicate that the survey measures particular subgroups of teachers, classrooms, or schools equally well.

Level 1 of the models contained predictor variables for nine teacher and classroom characteristics (independent variables are described in Appendix A). Recall that the outcome measures in the HLMs are the differences between a teachers' answer on a survey item and their answer on the equivalent daily log item. The metric of both sets of items was days per month, so the

Table 4. Proportion of Variation in Differences Between Teachers' Annual Survey and Log Reports Lying Between Teachers and Schools.

	Proportion of variation	
	Between teachers	Between schools
Wrote letters, strings, or words	0.980	0.020
Worked on a literature extension project	0.960	0.040
Wrote extensive answers to questions	0.990	0.010
Did a thinkaloud or explained skill use	0.985	0.015
Revised writing through elaboration	0.767	0.233
Revised writing by refining or reorganizing	0.733	0.267
Generated their own questions	0.874	0.126
Edited word use, grammar, or syntax	0.805	0.195
Worked on concept maps, story maps, and structure frames	0.952	0.048
Edited capitals, punctuations, or spelling	0.839	0.161
Wrote brief answers to questions	0.845	0.155
Summarized important details	0.824	0.176
Answered questions requiring inferences	0.758	0.242
Activated prior knowledge	0.887	0.113
Answered questions with answers directly stated in text	0.896	0.104

outcomes indicate by how many days per month teachers' reports on the two instruments differ. Positive values indicate that teachers reported engaging in the practice more frequently on the annual survey than on the log. Negative values indicate that the opposite was true. The reference teacher for the results presented in Tables 5 and 6 is a White male in a comparison school who was at the mean on all continuous independent variables.

We found that discrepancies between teachers' survey and log reports of their instruction were generally not strongly associated with their demographic characteristics nor the characteristics of their classrooms. Indeed, the vast majority of relationships tested for teacher-level predictors in the HLMs did not reach statistical significance. Female teachers were no more likely than males to over- or underreport their instruction on the survey, nor did we observe any significant differences among racial/ethnic groups. Two predictors that characterize teachers' classrooms—class size and average achievement level—were not statistically significant for most items, though teachers whose average classroom achievement was 1 standard deviation above the mean were predicted to overreport how often their students edited capitals,

Table 5. Models Predicting Factors Associated With Differences in Reports of Instruction Between an Annual Teacher Survey and Daily Log-Items 1-8.

	Item 1			Item 2			Item 3			Item 4			Item 5			Item 6			Item 7			Item 8		
	Coefficient	*SE*	Significance	Coefficient	*SE*	Significance	Coefficient	*SE*	Significance	Coefficient	*SE*	Significance	Coefficient	*SE*	Significance	Coefficient	*SE*	Significance	Coefficient	*SE*	Significance	Coefficient	*SE*	Significance
Intercept	4.89	0.40	***	2.97	0.27	***	5.05	0.26	***	5.79	0.30	***	4.21	0.36	***	3.70	0.37	***	4.38	0.35	***	4.54	0.42	***
Level 1 teachers																								
Female	1.68	−1.36		0.98	0.80		0.75	−1.08		−1.26	1.34		−1.57	1.75		−1.47	1.89		1.21	1.64		−1.78	1.50	
Hispanic	−4.05	−2.26		0.21	1.39		0.61	−1.35		−0.17	1.37		−0.30	1.94		−0.03	1.76		0.26	1.94		−0.22	2.35	
African American	−1.08	−1.46		−1.05	1.10		1.42	−0.97		−0.09	1.28		1.14	1.62		2.36	1.38		−0.67	1.20		−1.36	1.64	
White	−3.42	1.41	*	−1.37	0.99		−1.61	−0.91		−1.77	0.99		−2.11	1.47		−1.52	1.26		−1.66	1.12		−2.96	1.55	
Teaching experience	0.31	−0.40		−0.08	0.30		−0.41	−0.28		−0.45	0.31		−0.20	0.29		0.17	0.33		0.04	0.35		−0.36	0.40	
Content knowledge	−0.73	−0.44		−0.96	0.28	**	−0.78	0.27	**	−0.86	0.34	*	−0.36	0.36		−0.48	0.34		−0.64	0.37		−0.64	0.37	
Individualized instruction	2.16	0.92	*	0.92	0.62		0.69	−0.76		−0.12	0.85		0.16	0.82		0.63	0.75		−0.52	0.79		−0.06	0.92	
Class size	−0.16	−0.46		0.63	0.28	*	−0.14	−0.35		0.46	0.34		0.38	0.33		0.01	0.35		0.01	0.38		−0.05	0.51	
Average achievement	−0.06	−0.49		0.23	0.34		0.39	−0.32		0.23	0.37		0.41	0.30		0.57	0.35		0.05	0.46		0.27	0.56	
Level 2 schools																								
School SES	0.50	−0.43		0.15	0.32		−0.17	−0.32		0.51	0.40		−0.47	0.50		0.07	0.51		−0.38	0.52		−0.08	0.63	
Average achievement	0.60	−0.57		−0.11	0.33		−0.10	−0.39		0.34	0.49		−0.21	0.60		0.26	0.59		−0.10	0.57		0.10	0.67	
Accountability pressure	−0.44	−0.37		0.27	0.24		−0.05	−0.24		0.25	0.31		0.30	0.33		0.29	0.34		−0.37	0.32		0.15	0.37	
Accelerated Project	2.54	1.28	*	0.90	0.79		−0.12	−0.78		−0.02	0.85		1.62	0.97		1.63	0.98		−0.80	1.00		−0.55	1.14	
America's Choice	−0.44	−1.33		1.53	0.93		−0.47	−0.86		−1.56	1.00		2.05	1.29		2.76	1.27	*	−0.78	1.28		−0.39	1.55	
Success for All	−0.37	−1.31		1.38	0.87		−2.07	0.97	*	−2.14	1.09		−1.35	1.02		−1.14	1.11		−3.28	1.04	**	−3.67	1.30	**
Proportion of variance																								
Between teachers	0.998			0.968			0.998			0.996			0.809			0.762			0.945			0.893		
Between schools	0.002			0.032			0.002			0.004			0.191***			0.238***			0.055			0.107*		

Note. Item 1: Wrote letters, strings, or words; Item 2: Worked on a literature extension project; Item 3: Wrote extensive answers to questions; Item 4: Did a thinkaloud or explained skill use; Item 5: Revised writing through elaboration; Item 6: Revised writing by refining or reorganizing; Item 7: Worked on concept maps, story maps, and structure frames; Item 8: Generated their own questions. SES = socioeconomic status.
*$p \leq .05$. **$p \leq .01$. ***$p \leq .001$.

Table 6 Items 9-15. Models Predicting Factors Associated With Differences in Reports of Instruction Between an Annual Teacher Survey and Daily Log.

	Item 9			Item 10			Item 11			Item 12			Item 13			Item 14			Item 15		
	Coefficient	*SE*	Significance	Coefficient	*SE*	Significance	Coefficient	*SE*	Significance	Coefficient	*SE*	Significance	Coefficient	*SE*	Significance	Coefficient	*SE*	Significance	Coefficient	*SE*	Significance
Intercept	4.65	0.40	***	5.19	0.42	***	5.16	0.31	***	5.18	0.37	***	4.05	0.39	***	4.41	0.44	***	3.72	0.32	***
Level 1 teachers																					
Female	0.07	1.75		0.26	1.71		1.83	1.15		−1.56	1.31		−1.70	1.84		−1.48	2.37		−1.32	1.50	
Hispanic	0.18	1.99		1.76	2.10		0.82	1.84		−0.77	1.48		−1.18	2.01		0.07	2.07		−1.05	1.85	
African American	−0.38	1.71		0.10	1.52		−0.40	1.78		−1.79	1.65		0.41	1.55		−2.80	1.76		1.26	1.42	
White	−2.47	1.57		−2.18	1.38		−1.77	1.52		−3.57	1.31	**	−2.47	1.41		−2.15	1.52		−1.49	1.32	
Teaching experience	0.12	0.34		0.62	0.34		−0.61	0.32		−0.54	0.39		−0.28	0.42		−0.46	0.48		−0.66	0.40	
Content knowledge	−0.34	0.35		−0.04	0.42		−0.66	0.32	*	−0.56	0.38		−0.44	0.39		−0.67	0.39		−0.28	0.35	
Individualized instruction	0.29	0.87		−0.76	0.88		−0.56	0.86		−0.18	0.90		0.75	0.79		0.90	1.03		0.09	0.80	
Class size	−0.56	0.38		−0.03	0.40		−0.09	0.34		0.40	0.41		0.07	0.36		0.33	0.51		0.22	0.34	
Average achievement	0.65	0.34		0.95	0.35**		0.31	0.38		0.17	0.54		0.09	0.47		0.14	0.61		0.26	0.47	
Level 2 schools																					
School SES	−0.47	0.62		−0.90	0.62		0.08	0.42		0.05	0.51		0.17	0.46		−0.19	0.63		−0.01	0.41	
Average achievement	−0.37	0.59		−1.00	0.60		0.32	0.50		0.17	0.59		0.03	0.62		−0.26	0.69		0.50	0.53	
Accountability pressure	0.18	0.37		−0.11	0.39		−0.29	0.34		−0.24	0.34		−0.79	0.38	*	0.18	0.40		−0.33	0.35	
Accelerated Project	1.32	0.98		−0.24	1.08		−1.26	0.94		1.12	1.05		1.05	1.22		1.12	1.15		−1.31	1.06	
America's Choice	2.39	1.51		0.32	1.40		−1.96	1.16		2.00	1.33		0.39	1.30		0.71	1.55		−0.36	1.16	
Success for All	−0.71	1.18		−1.19	1.23		−5.85	1.17	***	−3.60	1.41	*	−4.61	1.21	***	−3.33	1.48	*	−5.48	0.98	***
Proportion of variance																					
Between teachers	.814			.849			.993			.915			.872			.927			.998		
Between schools	.186***			.151**			.007			.085			.128*			.073			.002		

Note. Item 9: Edited word use, grammar, or syntax; Item 10: Edited capitals, punctuations, or spelling; Item 11: Wrote brief answers to questions; Item 12: Summarized important details; Item 13: Answered questions requiring inferences; Item 14: Activated prior knowledge; Item 15: Answered questions with answers directly stated in text. SES = socioeconomic status.

*$p \leq .05$. **$p \leq .01$. ***$p \leq .001$.

punctuation, or spelling by an additional 1 day per month. The general lack of significant effects of teacher-level predictors also suggests to us that, at least in terms of the factors examined, reporting errors on the annual survey generally do not reflect bias among specific teacher subgroups.

A measure indicating whether teachers individualized instruction for their students was included in the models. We conjectured that the instructional practices of teachers who provide individualized instruction to their students would tend to be more variable and less similar from one occurrence to the next. According to Menon's (1994) cognitive model, highly variable activities like these are more likely to be stored in episodic memory, which is less readily accessed when surveys require recall over long periods of time. Consequently, we hypothesized that individualizing instruction would make it more difficult for teachers to accurately report what they did on the survey. We found very little support for this hypothesis as the predictor for individualizing instruction was statistically significant for only one item. Teachers' who reported individualizing instruction *more* tended to report the frequency with which their students wrote letters, strings, or words with *less* accuracy.

We also tested whether survey reporting error is associated with teachers' experience and expertise. We found evidence supporting the idea that teachers with greater content knowledge for literacy teaching reported their literacy instruction on annual surveys with slightly more accuracy than average. For 4 of the 15 items, we found statistically significant negative effects of teachers' literacy teaching content knowledge, indicating that the higher the teachers' score on the knowledge measure, the smaller the discrepancies between their survey and log answers. More specifically, teachers with higher content knowledge for literacy teaching appeared to more accurately report their students' work on literature extension projects, writing extensive answers to questions, participating in thinkalouds, and writing brief answers to questions. For example, whereas the reference teacher was predicted to overstate how often their students worked on literature extension projects by 2.97 days per month, a teacher, whose content knowledge for literacy teaching was a standard deviation *above* the mean, was predicted to overstate this practice on the survey by only 2 days per month (2.97 + −0.96). We conjecture that teachers with stronger content knowledge for teaching might be more aware of what they teach and when they teach it than less knowledgeable teachers, and that this awareness might have allowed them to report their instruction on the annual survey with greater accuracy. There is some basis for this conjecture in research on expertise, which indicates that expertise grows out of experience and is exercised in part by recalling past experiences from memory. However, while our evidence supports fairly robust inferences about the impact of elapsed time on survey reporting error, we are on less

solid ground when it comes to drawing inferences about the influence of teachers' knowledge on reporting error.

School-Level Factors Associated With Measurement Error

In the final stage of the analysis, we tested whether characteristics of the contexts in which teachers worked were associated with greater discrepancies in their annual survey and daily log reports. In particular, we were interested in seeing whether teachers' contexts affected the accuracy of their survey responses, and whether teachers in particular contexts were more prone to socially desirable reporting due to perceived pressures in their environment. The majority of the school-level predictor variables had insignificant effects. Neither schools' average SES level nor schools' average achievement level was a significant predictor of the accuracy of teachers' reports of the enacted third-grade literacy curriculum on the survey.

We conjectured that schools' participation in a CSR program might be a significant contextual factor affecting teachers' reports of their instruction. As mentioned previously, three quarters of all sample schools participated in a CSR program. In fact, two of these programs prescribed specific literacy instruction routines for teachers to implement. A number of schools in the sample were also encouraged to adopt CSR models as a response to accountability pressures to improve student achievement. We conjectured that teaching in a school that adopted a CSR program might affect teachers' survey responses by exerting social pressure to report practices advocated by the program. We also conjectured that participating in professional development in literacy instruction and reflecting on one's literacy teaching as part of a CSR program might make teachers more attentive to what they were teaching, thus making their survey reports more accurate. To test these conjectures, we added dummy variables to the Level 2 models indicating schools' participation in the three CSR programs. We also included an explicit measure of accountability pressure placed on schools in the Level 2 models.

We found that for all 15 items, discrepancies in the survey and log reports of teachers in schools implementing SFA were *smaller* than those of teachers in comparison schools (the reference group) and for more than half of the items (8 of 15), the difference between SFA and comparison school teachers was statistically significant. The difference between SFA and comparison school teachers ranged from 2.07 days per month (writing extensive answers to questions) to 5.85 days per month (writing brief answers to questions). Adding the estimates for SFA teachers to model

intercepts produces an estimate of the average discrepancy between the survey and log reports of SFA teachers. For two items (having students work on concept maps, having students activate prior knowledge), the average discrepancy between SFA teachers' survey and log reports was approximately 1 day per month, whereas the predicted discrepancy for comparison school teachers on these items was 3 days per month. For three items (having students generate their own questions, having students write brief answers to questions, having students answer questions requiring inferences), SFA teachers' survey and log reports differed by less than 1 day per month, while differences for comparison school teachers on these same items ranged from 3.67 to 5.85 days.

The SFA program stipulates in great specificity teaching methods teachers should use, with an emphasis on cooperative learning arrangements for students and direct instruction. Because of this great level of specificity, the observed SFA effects may indicate that the program narrows teachers' focus on a particular set of instructional practices in such a way that they become more aware of how frequently they use the practices. SFA also makes use of a number of routines such as a daily 90-min reading block and periodic assessments. It could also be that these routines bring a regularity to teachers' schedules that makes it easier for them to remember what they did and then report it on a survey. As previously noted, cognitive perspectives on the survey response process indicate that behaviors that are more similar from episode to episode can be more easily recalled from episodic memory.

While the SFA program placed a strong emphasis on reading comprehension, the AC program placed a strong emphasis on developing teachers' capacity to provide effective writing instruction. For AC teachers, their annual survey reports of how often their students revised their writing by refining or reorganizing exceeded the average by nearly 3 days per month (2.76). Given the strong emphasis on this practice by the program, AC teachers' apparent overstatement by more than 6 days per month (3.70 + 2.76) of how often their students engaged in this practice may indicate socially desirable reporting on the part of teachers implementing this program.

Discussion

This study adds to an existing body of research, which suggests that policymakers and researchers should interpret evidence about important educational processes like instruction with considerable care (Desimone & Le Floch, 2004; Mayer, 1999; Mullens et al., 1999). Like other studies,

we found *mixed* evidence about the validity of an annual survey for measuring instruction. A basic criterion we examined was whether the two instruments produced similar portraits of the relative emphasis teachers placed on different literacy instruction practices. On this score, the annual survey fared quite well as the pattern of instructional emphasis produced by the survey closely mirrored that of the log. A second criterion we examined was the degree to which teachers' answers on the annual survey were correlated with their answers to comparable items on the daily log. On this criterion, the survey fared less well as teachers' responses to the two instruments were only modestly correlated with each other. A third question we examined was whether reporting discrepancies between the two instruments was essentially random, or whether it was greater among different kinds of teachers or different school settings. Here, again the evidence was mixed. Many teacher and school characteristics examined were unrelated to discrepant reporting, while a handful of such characteristics were found to be statistically significant predictors of reporting discrepancies.

It is important to note a number of limitations of this study. First, it is worth noting that the results of this study might not be broadly generalizable because the study was conducted primarily in schools in large, disadvantaged, urban districts. Another distinctive feature of the sample of teachers is that study participants used two different instruments to document their teaching practice during the study period. Could it be that reporting their practice on logs throughout the school year might have influenced how teachers reported their practice on the annual survey? To test this concern, we reestimated all the HLMs with the number of logs each teacher completed as a predictor variable. By and large, the number of logs completed did not significantly predict differences between the questionnaire and daily log instrument. A significant association between the number of logs a teacher completed and their reporting accuracy was observed for only one item (wrote letters, strings, or words, $\beta = -.05$, $p = .036$). We conclude from these results that the teachers' completion of daily logs did not have a major influence on how they responded to the annual surveys. The most widespread indication of invalidity was evidence that the annual survey substantially and systematically overstates the amount of time teachers spend on the enacted literacy curriculum. Overall, more than three fourths of the teachers in this study overreported their instruction on the annual survey, and we predict that, on average, teachers' survey reports overstate the frequency with which they implement a given instructional practice by nearly 4.4 days per month. The consistency of results across the 15 instructional practices examined suggests

to us that the annual survey may possess a general "response bias," which, Groves et al. (2009) said, occurs when the direction of error in a survey measurement (i.e., its departure from a true value) is consistent across multiple "trials." The multiple items used to measure literacy instruction and the multiple teachers who completed the two instruments can be viewed as multiple trials. Examining mean differences in teachers' responses to the two instruments, we observe complete consistency in the direction of survey response errors. In every case, the annual survey means were higher than daily log means. However, this study did not find any consistent patterns across the 15 instructional practices (e.g., writing-related instructional practices show larger magnitude of overreports on survey than other instructional practices). While the preponderance of teachers had annual survey reports that exceeded their log reports, there were some teachers for whom the opposite was true. Given that measurement error was not entirely limited to overreporting, we conclude that the evidence approaches but ultimately falls slightly short of indicating a general response bias associated with the annual survey.

Although relatively few in number, statistically significant results in the HLM analyses indicate that reporting errors on the annual survey were greater among some groups of teachers than others. In particular, we found that survey/log discrepancies were considerably smaller for teachers with greater knowledge for teaching literacy content areas. We also found evidence suggesting that the context in which teachers work may influence how teachers answer surveys, and subsequently, response accuracy. And while our results do not permit us to make definitive causal inferences about contextual influences on teachers' survey responses, they do suggest to us that some contextual influences may increase reporting accuracy while others may diminish it. The finding that teachers in schools implementing the SFA program provided considerably more accurate reports of their literacy instruction than teachers in a set of comparison schools may indicate a positive contextual influence. In contrast, the finding that teachers in schools implementing the AC program overstated their use of an instructional practice that was a focus of that program even more than comparison school teachers might indicate a social desirability bias among AC teachers. These are conjectures, however, as our data do not support strong causal inferences about the effects of social desirability. Moreover, the findings about the association between social desirability and measurement errors in the annual survey should be interpreted with caution because social desirability may affect both instruments equally. In other words, teachers' reports on both instruments might be affected by social desirability. Taken together, the statistically significant HLM results are clearly a cause for concern because when they interpret results

from surveys, policymakers and others assume that all groups of interest in all settings are measured with equal error. Our findings indicate that this assumption does not hold for this sample. We urge further research that examines the specific causes of invalidity in teacher survey responses.

The deep and long-standing literature on cognitive perspectives on the survey response process proved insightful to us, and we believe that future application of these perspectives by policymakers and researchers could prove to be equally useful. The perspectives provided us with a theoretically and empirically grounded justification for regarding the log as a benchmark, and they helped us make sense of the study's results. In particular, our interpretation of the finding that teachers with greater knowledge for teaching literacy appeared to report their literacy instruction with greater accuracy struck us as consistent with Tourangeau and Rasinski's (1988) cognitive model of the survey response process. Similarly, Menon's (1994) model which posits that more highly variable activities will be more difficult to recall over long periods of elapsed time informed our inference about teachers who individualized their instruction. Our interpretation of the effects associated with schools' CSR implementation was also informed by this literature. For example, our conjecture that survey reports of writing instruction by teachers in schools implementing AC may have reflected socially desirable response biases has a basis in (a) our knowledge that some AC schools in the sample were under considerable pressure to implement new writing instructional practices and (b) Tourangeau and Rasinski's (1988) model which stipulates that respondents *edit* their responses in response to considerations of social desirability and self-presentation. Similarly, our interpretation of results for SFA schools reflected (a) an understanding of that program as providing highly specified instructional routines for teachers and (b) the role of understanding, judgment, and episodic similarity as stipulated in the models of Tourangeau and Rasinski (1988) and Menon (1994). It is our observation that the rich bodies of conceptual knowledge and empirical evidence in the survey research methodological literature have gone largely untapped in the field of educational research. In future inquiries, we urge researchers to make a greater use of this body of research.

Based on these results, we would also urge policymakers and researchers to use estimates of instruction from annual surveys with caution. In particular, these results suggest that point estimates of specific instructional practices based on annual survey data likely overstate teachers' use of the practices to a significant degree. Because of this, we would caution against making precise statements about the incidence of a specific instructional practice based on annual survey results. Users of annual survey data might

also consider conducting analyses like those presented here as a preanalysis strategy to identify individual and organizational correlates of reporting error. If significant correlates are found, statistical estimates based on survey data might be improved by controlling for the correlates.

There have been recent recommendations to use student self-report surveys to measure teachers and their teaching effectiveness (Bill & Melinda Gates Foundation, 2010) despite the findings of prior research, which found fairly modest correlations between the survey reports of teachers and students (Desimone et al., 2010). We conjecture that estimates of instruction based on annual surveys given to students are likely prone to the same sorts of reporting errors observed here. Certainly, the incentives for reporting particular instructional activities are different for students than for teachers, and therefore, socially desirable response biases will likely be different for these two groups. However, to a substantial degree, the systematic overreporting on the annual survey reported here strikes us as being a function of the inherent difficulty of recalling details about a complex activity like instruction after a long period of time has elapsed. We conjecture that students experience instruction in a similarly complex way as teachers, and therefore, are likely to have similar kinds of difficulty reporting their instructional experiences on an annual survey. While our results are limited to measures of instruction, we surmise that survey measurement of other important education variables, which require respondents to consider many events over long expanses of time, will also suffer from the same kind of reporting error that we observed.

Finally, our results led us to consider the potential gains in accuracy that might be achieved if the amount of time about which teachers report was reduced. In this study, the difference in reference periods for the two instruments was dramatic—a single day compared with a whole school year. Our results suggest that self-reports of instruction are likely to be substantially more accurate if teachers are asked to report their instruction in a way that allows them to tap their memories of their complex classroom experiences with greater accuracy. Daily logs can be costly and burdensome, and therefore may not be feasible in some circumstances. But there may be middle ground methodologies that require fewer responses from teachers than a daily log but still provide more accurate measurement of instruction than annual surveys. In our view, surveys are often chosen when cost is given primary consideration. As policymakers and researchers continue to look for valid yet cost-effective measures of teacher quality and student learning experiences, we urge further investigations on alternative methodologies that provide a more optimal balance of accuracy, cost, and teacher burden.

Appendix A

Descriptions of Independent Variables.

Variable	Description
Level 1 teachers	
Female	Dummy variable coded "1" if a teacher is female, "0" if a teacher is male.
Race	Set of four dummy variables (Hispanic, African American, White, and Other) indicating a teacher's race.
Years of experience	The number of years the teacher has taught.
Content knowledge for literacy teaching	An IRT scale score measuring teachers' content knowledge for teaching literacy.
Teacher individualizes instruction	Dummy variable indicating if a teacher reported he or she individualizes instruction.
Class size	The number of students in a teacher's classroom.
Average classroom achievement level	The average achievement level of students in a teacher's classroom as measured by the TerraNova achievement test.
Level 2 schools	
School SES	This measure characterizes the socioeconomic conditions of the school. It was formed by combining block-level census data for the school, percent of free lunch, and the mean of a student-level measure of SES.
Average school achievement level	This variable was created by taking the average of schools' mean reading and math scores as measured by the TerraNova assessment.
Percentage of students eligible for free and reduced-price lunch	The percentage of students who were eligible for the Federal free and reduced-price lunch program. Percentages were obtained from the National Center for Education Statistic's Common Core of Data.
Percentage of non-White students	The percentage of students in a school whose race is not White. Percentages were obtained from the National Center for Education Statistic's Common Core of Data.
Accountability pressure	The variable was a combination of six items which measured the degree of parent and community member dissatisfaction with student achievement in the school, demands for improvement placed on the school, and whether the school had been formally identified as "in need of improvement."
CSR program	Set of three dummy variables indicating whether a school participated in the Accelerated Schools Project, America's Choice, or Successful for All comprehensive school reform programs.

Note. IRT = item response theory; SES = socioeconomic status; CSR = comprehensive school reforms.

Appendix B

Statistical Models

The equations used to estimate the HLMs are as follows.

Level 1.In the Level 1 model, the average level of disagreement between annual survey and daily log reports for each teacher, π_{0ij}, is modeled as a function of teacher and classroom characteristics:

$$\left(\text{Difference Between Survey and Log}\right)_{ij} = \beta_{0j} + \beta_{1j} \times$$
$$\left(\text{Teacher Predictor } 1_{ij}\right) + \cdots + \beta_{xj} \times \left(\text{Teacher Predictor } X_{ij}\right) + r_{ij},$$

where β_{0j} is the average level of disagreement between survey and log reports for teachers in school *j*. Variation in the outcome that is unique to each teacher is captured in the term r_{ij}.

Level 2.At Level 2, β_{0j} is the average level of disagreement in school *j* between teachers' answers on the annual survey and daily log is predicted as a function of γ_{00}, the overall level of disagreement across all schools, school-level predictors, and a random effect associated with each school, u_{00j} .

$$\beta_{0j} = \gamma_{00} + \gamma_{01} \times \left(\text{School Predictor } 1_j\right) + \cdots$$
$$+ \gamma_{0y} \times \left(\text{School Predictor } Y_j\right) + u_{0j}.$$

The descriptions of teacher-level and school-level predictors are provided in Appendix A.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received the follwoing financial support for the research, authorship, and/or publication of this article: The research reported here was supported by grants from the U.S. Department of Education to the Consortium for Policy Research in Education at the University of Pennsylvania and to the Center for the Study of Teaching and Policy at the University of Washington; by the National Science Foundation's Interagency Educational Research Initiative; by the Atlantic Philanthropies, USA; and by the William and Flora Hewlett Foundation.

Note

1. To test the difference in the means for the two instruments, simple *t* tests are not appropriate as the data have a nested structure (teachers within schools). Using two-level models, we estimated the outcome variables measuring the differences between teachers' log and survey responses using models with no predictors at either level. The intercept of these models is the estimated mean difference in teachers' responses on the two instruments. For each model, this intercept was found to be significantly different than 0 at the .001 level. Estimates of the intercepts in Tables 5 and 6 show that these statistically significant differences in teachers' responses to the two instruments persisted even after controlling for characteristics of teachers and schools.

References

Anderson, J. R. (1983). *The architecture of cognition*. Cambridge, MA: Harvard University Press.

Bill & Melinda Gates Foundation. (2010). *Learning about teaching: Initial findings from the Measures of Effective Teaching project* (MET Project Research Paper). Seattle, WA: Author.

Bryk, A. S., & Schneider, B. (2002). *Trust in schools: A core resource for improvement*. New York, NY: Russell Sage Foundation.

Bryk, A. S., Sebring, P. B., Allensworth, E., Luppescu, S., & Easton, J. Q. (2010). *Organizing schools for improvement lessons from Chicago*. Chicago, IL: University of Chicago Press.

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

Camburn, E. M. (2010). Embedded teacher learning opportunities as a site for reflective practice: An exploratory study. *American Journal of Education*, *116*, 463-489.

Camburn, E. M., & Barnes, C. A. (2004). Assessing the validity of a language arts instruction log through triangulation. *Elementary School Journal*, *105*, 49-73.

Camburn, E. M., & Han, S. W. (2009). Investigating connections between distributed leadership and instructional change. *Distributed Leadership: Different Perspectives*, *7*, 25-45. doi:10.1007/978-1-4020-9737-9_3

Camburn, E. M., & Han, S. W. (2011). Two decades of generalizable evidence on U.S. instruction from national surveys. *Teachers College Record*, *113*, 561-610.

Camburn, E. M., Huff, J. T., Goldring, E. B., & May, H. (2010). Assessing the validity of an annual survey for measuring principal leadership practice. *Elementary School Journal*, *111*, 314-335.

Cybulski, T. G., Hoy, W. K., & Sweetland, S. R. (2005). The roles of collective efficacy of teachers and fiscal efficiency in student achievement. *Journal of Educational Administration*, *43*, 439-461.

Desimone, L. M. (2006). Consider the source: Response differences among teachers, principals, and districts on survey questions about their education policy environment. *Educational Policy*, *20*, 640-676.

Desimone, L. M., & Le Floch, K. C. (2004). Are we asking the right questions? Using cognitive interviews to improve surveys in education research. *Educational Evaluation and Policy Analysis*, *26*, 1-22.

Desimone, L. M., Smith, T. M., & Frisvold, D. E. (2010). Survey measures of classroom instruction: Comparing student and teacher reports. *Educational Policy*, *24*, 267-329.

Dillman, D. A. (1991). The Design and Administration of Mail Surveys. *Annual Review of Sociology*, *17*, 225-249.

Groves, R. M., Fowler, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2009). *Survey methodology* (2nd ed.). Hoboken, NJ: Wiley.

Heeringa, S., Little, R. J. A., & Raghunathan, T. (2002). Multivariate imputation of coarsened survey data on household wealth. In R.M. Groves, D.A. Dillman, J.L. Eltinge & R.J.A. Little (Eds) *Survey nonresponse* (pp. 1-19). New York, NY: Wiley.

Hill, H. C. (2005). Content across communities: Validating measures of elementary mathematics instruction. *Educational Policy*, *19*(3), 447-475.

Hilton, M. E. (1989). A comparison of a prospective diary and two summary recall techniques for recording alcohol consumption. *British Journal of Addiction*, *84*, 1085-1092.

Lemmens, P., Knibbe, R. A., & Tan, F. (1988). Weekly recall and diary estimates of alcohol consumption in a general population survey. *Journal of Studies on Alcohol*, *49*, 131-135.

Lemmens, P., & Tan, E. S. (1992). Measuring quantity and frequency of drinking in a general population survey: A comparison of five. *Journal of Studies on Alcohol*, *53*, 476-486.

Linton, M. (1988). Real-world memory after six years: An in-vivo study of long-term memory. In M. M. Gruneberg, P. E. Morris, & R. N. Sykes (Eds.), *Practical aspects of memory: Current research and issues* (pp. 69-76). Chichester, UK: Wiley.

Luppescu, S., Hart, H., Rosenkranz, T., Montgomery, N., Sporte, S., Bender Sebring, P., & Mazzeo, C. (2007). *CCSR's 2007 Survey Reports for Chicago Public Schools*. *Chicago, IL: Consortium on Chicago School Research*. Retrieved on 5/12/15 from https://ccsr.uchicago.edu/sites/default/files/publications/rpt7777.pdf.

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

Menon, G. (1994). Judgments of behavioral frequencies: Memory search and retrieval strategies. In N. Schwarz & S. Seymour (Eds.), *Autobiographical memory and the validity of retrospective reports* (pp. 161-172). New York, NY: Springer-Verlag.

Milesi, C., & Gamoran, A. (2006). Effects of class size and instruction on kindergarten achievement. *Educational Evaluation and Policy Analysis*, *28*(4), 287-313.

Mullens, J. E., Gayler, K., Goldstein, D., Hildreth, J., Rubenstein, M., Spiggle, T., . . . Welsh, M. (1999). *Measuring classroom instructional processes: Using survey and case study fieldtest results to improve item construction* (Working paper no. 1999-08). Washington, DC: U.S. Department of Education, Office of Educational Research and Improvement, National Center for Education Statistics.

Mullis, I. V. S., Martin, M. O., Foy, P., & Arora, A. (2012). *TIMSS 2011 international results in mathematics*. Chestnut Hill, MA: Boston College, TIMSS & PIRLS International Study Center.

Organisation for Economic Co-Operation and Development. (2013a). *PISA 2012 results: Ready to learn—Students' engagement, drive and self-beliefs* (Vol. III). Paris, France: Program for International Student Assessment, Organisation for Economic Co-Operation and Development.

Organisation for Economic Co-Operation and Development. (2013b). *PISA 2012 results: What makes schools successful? Resources, policies and practices*. Paris, France: Program for International Student Assessment, Organisation for Economic Co-Operation and Development.

Porter, A. C. (2002). Measuring the content of instruction: Uses in research and practice. *Educational Researcher*, *31*(7), 3-14.

Rowan, B., Camburn, E., & Correnti, R. (2004). Using teacher logs to measure the enacted curriculum: A study of literacy teaching in third-grade classrooms. *The Elementary School Journal*, *105*(1), 75-101.

Rowan, B., & Miller, R. J. (2007). Organizational strategies for promoting instructional change: Implementation dynamics in schools working with comprehensive school reform providers. *American Educational Research Journal*, *44*(2), 252-297.

Schmidt, W. H., & McKnight, C. C. (1995). Surveying educational opportunity in mathematics and science: An international perspective. *Educational Evaluation and Policy Analysis*, *17*(3), 337-353.

Smithson, J. L., & Porter, A. C. (1994). *Measuring classroom practice: Lessons learned from efforts to describe the enacted curriculum—The reform up close study* (CPRE Research Report Series Report No. 31). New Brunswick, NJ: Consortium for Policy Research in Education.

Thompson, C. P., Skowronski, J. J., Larsen, S. F., & Betz, A. L. (1996). *Autobiographical memory: Remembering what and remembering when*. Mahwah, NJ: Lawrence Erlbaum.

Tourangeau, R., & Rasinski, K. A. (1988). Cognitive processes underlying context effects in attitude measurement. *Psychological bulletin*, *103*(3), 299.

Tourangeau, R., Rips, L. J., & Rasinski, K. A. (2000). *The psychology of survey response*. Cambridge, UK: Cambridge University Press.

Tulving, E. (1983). *Elements of episodic memory*. Oxford, UK: Oxford University Press.

Unge, J., Hansson, G.-A., Ohlsson, K., Nordander, C., Axmon, A., Winkel, J., & Skerfving, S. (2005). Validity of self-assessed reports of occurrence and duration of occupational tasks. *Ergonomics*, *48*, 12-24.

Wagenaar, W. A. (1986). My memory: A study of autobiographical memory over six years. *Cognitive Psychology*, *18*, 225-252. doi:10.1016/0010-0285(86)90013-7
White, R. T. (1982). Memory for personal events. *Human Learning*, *1*, 171-183.

Author Biographies

Eric M. Camburn is an Associate Professor at the University of Wisconsin-Madison. His interests include school improvement, particularly in urban schools, and survey research methods.

Seong Won Han is an Assistant Professor at the University at Buffalo, State University of New York. Her research interests include international and comparative education, gender inequality in STEM, educational policy and teacher quality.

James Sebastian is an Assistant Professor at the University of Missouri-Columbia. His research interests include the study of school leadership, organizational theory and behavior, organizational learning, and urban school reform.