ASSESSING THE VALIDITY OF AN ANNUAL SURVEY FOR MEASURING PRINCIPAL LEADERSHIP PRACTICE

ABSTRACT

Because research has shown that principals' practices can significantly impact teaching and learning in their schools through multiple avenues, it is critical to understand the validity of tools that measure principal practice. While the field has relied heavily on self-report surveys, little is known about their validity. This study compares survey measures of principal leadership practice with comparable daily log measures. We construct a multitrait-multimethod (MTMM) matrix showing intercorrelations among 4 dimensions of principal leadership captured with the 2 instruments. Correlations were evaluated against criteria used for MTMM analyses. Considerable evidence attested to the principal survey's validity, with correlations between log and survey measures of all 4 dimensions exceeding .50. Results suggest that surveys may be better at measuring principals' management of the school building and instructional leadership than measuring more irregular activities like planning and personal professional development. We discuss steps for improving the validity of principal surveys.

Eric M. Camburn
UNIVERSITY OF WISCONSIN—MADISON

Jason T. Huff
UNIVERSITY OF TENNESSEE

Ellen B. Goldring
VANDERBILT UNIVERSITY

Henry May
UNIVERSITY OF PENNSYLVANIA

THERE is clear evidence that principals can substantially impact the quality of teaching and learning in their schools by shaping school structures and influencing people (Leithwood, Seashore Louis, Anderson, & Wahlstrom, 2004). The evidence suggests that principals exercise this influence through multiple practices including setting clear and strategic directions; recruiting, developing, and retaining effective teachers; and creating organizational conditions that

 0013-5984/2010/11102-0005 $10.00

support effective teaching and learning (Leithwood & Duke, 1999; Leithwood et al., 2004; Pounder, Ogawa, & Adams, 1995). Given principals' central role, it is vitally important that we have valid and reliable measurement tools to provide trustworthy evidence about what they do.

Self-report surveys are widely used to measure principal leadership (Hallinger & Heck, 1996); their relatively low cost, low respondent burden, and ease of administration make their ubiquitous use among leadership scholars understandable. Beyond economic advantages, the scientific value of self-report surveys in educational research finds support in a considerable body of research that establishes the predictive validity of survey measures. For example, Bryk and Schneider (2002) found that levels of trust among teachers, principals, and parents measured by teacher surveys were positively associated with school-level improvements in reading and mathematics achievement. Bryk and Schneider (2002) grounded measures of trust and statistical analyses in a carefully articulated theoretical framework, and consequently their results attest to the validity of the survey measures. Similarly, Cybulski, Hoy, and Sweetland (2005) used survey measures of teacher collective efficacy to test a theoretical set of hypotheses about the relationship between collective efficacy and student achievement, finding evidence of a positive relationship between these two variables and thus attesting to the predictive validity of the survey measures. Finally, Camburn and Han (2011) found 49 studies that examined the relationship between classroom instruction and student achievement using generalizable evidence on instruction from large-scale surveys. Many of these studies reported positive associations with achievement that were predicted by investigators, again attesting to the validity of the survey measures used.

Despite the ubiquity and utility of surveys in research on principal leadership, there is relatively little evidence about their accuracy or validity. The most comprehensive study to date was conducted as part of the Schools and Staffing Survey (SASS) and examined a principal survey that collected data on school personnel (Levine, Chambers, Ixtlac, & Hikido, 1998). Evidence about the validity of the survey was mixed. While the survey captured some information with greater accuracy than district records, cognitive interviews revealed that principals sometimes failed to follow directions in the survey protocol, and principals' and researchers' understandings of a number of concepts differed, including basic terms such as which individuals in a school qualified as staff (Levine et al., 1998).

The widespread use of self-report surveys has been identified as a significant shortcoming of research in the field of organizational behavior (Donaldson, Ensher, & Grant-Vallone, 2000; Mersman & Donaldson, 2000). This concern appears to have a basis in the survey methods literature because self-report surveys have been shown to be particularly prone to two kinds of response errors: Respondents leave more questions unanswered in self-report surveys than in interviewer-administered surveys (Tourangeau, Rips, & Rasinski, 2000), and self-report surveys commonly require respondents to recall specific behaviors or events that occurred in the past—a difficult task that leads to inaccurate survey answers. There is considerable evidence of reporting errors when the reference period for recalling an event or behavior is long (e.g., Hilton, 1989; Lemmens, Knibbe, & Tan, 1988; Lemmens & Tan, 1992; Rubin & Baddeley, 1989). In research on principal leadership, it is common to survey principals during the spring of

the school year and ask them to recall behaviors and events that occurred across the span of the entire school year.

New data-collection strategies, such as daily logs and experience-sampling instruments, may allow researchers to capture more accurate large-scale evidence of principals' actions (Camburn, Spillane, & Sebastian, 2010; Gronn, 2003; Huff, 2006). There is considerable evidence that daily logs are more accurate than traditional surveys at measuring behavior because they are less susceptible to recall error (Hilton, 1989; Lemmens et al., 1988; Lemmens & Tan, 1992). Because of this greater accuracy, researchers 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) have relied on data from daily instruments as a benchmark for assessing the validity of annual surveys. However, the gains in accuracy of these newer methods come at a cost. Daily and experience-sampling instruments often require computer programming, and because these data-collection strategies are substantially more burdensome than surveys, it is necessary to provide larger incentive payments to respondents in order to achieve high response rates.

Because of their relative ease of use and lower administration costs, principal surveys will continue to figure prominently in principal leadership research for the foreseeable future. Therefore, in our view it is important to better understand the validity of these instruments. In this study we assess the validity of an annual principal survey by comparing measures of principal leadership practice from the survey with comparable measures from a daily log. Our assessment employs a commonly used strategy of evaluating data from a survey against evidence from a daily log that is viewed as a benchmark (Burstein et al., 1995; Hilton, 1989; Lemmens et al., 1988; Lemmens & Tan, 1992; Smithson & Porter, 1994).

First we situate our analysis within a broader discussion of validity as it relates to survey measurement, then we briefly review a small number of studies that have assessed the validity of principal surveys. Next we present the instruments used to examine the relationship between the logs and surveys, and present our findings. We then assess the validity of measures from the principal survey by constructing a multitrait-multimethod (MTMM) matrix. Based on the methods of Campbell and Fiske (1959), the matrix shows intercorrelations among the measures of four leadership practices that were captured using two different methods (the survey and daily log). In the final section we discuss the implications of the results for the measurement of principal leadership practice.

Background

Like all kinds of surveys, self-report surveys are affected by multiple sources of error. Groves et al. (2004) described measurement error as the failure of survey responses to accurately represent the intended construct, and they presented a model that describes how each step of the survey process can introduce measurement error. Two points in this process, operationalization—which Groves et al. (2004) called measurement—and survey response, are associated with the survey instrument itself and are thus relevant to this study. Measurement error can thus be introduced when (*a*) the survey does not represent the constructs well or (*b*) when respondents' answers inadequately represent the constructs.

There are various methods for evaluating survey items to minimize measurement error before a survey is administered. Having substantive experts review surveys can assess how well particular items represent the full range of topics or concepts associated with a construct of interest. Having survey experts review question wording, question order, question formatting, and other details can increase the likelihood that the survey will be clearly understood by respondents. Finally, focus groups and open-ended interviews conducted with members of the target population can shed light on what respondents know about survey topics, whether respondents are able to provide the necessary information to validly answer questions, and the language used by respondents to discuss the topics.

The second source of error of interest in this study is response error, which is introduced when respondents answer the survey instrument. In this case respondents may misunderstand questions and instructions, or their understanding of concepts referenced in survey questions may differ from that of researchers. Cognitive interviews are useful for finding out how individuals understand questions and arrive at their answers. In such interviews respondents typically answer survey questions verbally, and then interviewers ask probing questions in a conversational style. Cognitive interviews sometimes use think-aloud strategies, which are believed to capture respondents' cognitive processing while answering surveys.

A range of strategies can be employed to evaluate whether measurement error introduced either through faulty operationalization or response error has adversely affected a survey statistic. These strategies focus primarily on examining how well respondents' survey answers correspond to what the investigator is trying to measure, typically comparing a measurement from a survey to another measurement. The other measurement can come from the same survey, a different survey, or another source such as physical measurements or administrative records. Examples of such strategies include correlating two variables from the same survey that are hypothesized to be correlated with each other, correlating a measure of a construct from a survey with a measure of the same construct from a different data source, and using qualitative evidence on a construct to assess the validity of a survey measure of the construct.

There is a small yet growing body of research that assesses the validity of principal questionnaires using a number of the strategies just discussed. A handful of studies have documented strategies used by researchers in the survey development stage to reduce measurement error in principal surveys. Specific strategies used by these researchers include engaging substantive experts, survey experts, and members of the target population to review survey items, recommend constructs, and assess the feasibility of retrieving information requested by survey items (Dryden, 1995; Levine et al., 1998; Wildy & Clarke, 2009; Wildy, Forster, Louden, & Wallace, 2004).

A small number of studies have also used cognitive interviews to assess principals' understanding of the intended meaning of questions and survey instructions in an effort to reduce response error (Levine et al., 1998; Wildy & Clarke, 2009). For example, Wildy and Clarke (2009) used cognitive interviews to examine the degree to which principals shared researchers' understanding of terms used in survey items. These researchers identified a small number of items whose meanings were unclear and subsequently used this evidence to revise items. Cognitive interviews were also used to improve the collection of personnel data as part of the SASS. Levine et al. (1998) found that principal respondents frequently defined key concepts differently

than researchers. For example, principals' definition of the fundamental concept *school staff* was typically less inclusive than that of researchers. Cognitive interviews also revealed tremendous variation in the degree to which respondents paid attention to and understood directions on the survey. On the basis of the cognitive interview results, investigators proposed significant changes to SASS data-collection procedures.

A number of researchers have also conducted quantitative analyses to empirically test whether items validly measure constructs. Wildy et al. (2004) conducted extensive psychometric analyses to assess how well items intended to measure the same construct elicited internally consistent responses from principals. Other studies used factor analysis for empirically judging whether item clusters measured a construct well (e.g., Claudet & Ellett, 1993; LeSourd, Tracz, & Grady, 1990). Additional studies examined correlations between variables that were theoretically predicted to be correlated and interpreted correlations as evidence of convergent validity (Claudet & Ellett, 1993; Henderson & Hoy, 1982).

Finally, a number of researchers have attempted to gauge the validity of principal surveys by triangulating survey results against other sources of evidence. Levine et al. (1998) tested the accuracy of a survey on school staffing completed by principals and assistant principals against school district staffing records. Comparison of staffing lists produced by the two data sources revealed numerous discrepancies, and the investigators concluded that district records were more likely than survey responses to be the source of inaccuracy. The investigators concluded that collecting valid data on school staffing from district records was not feasible.

Desimone (2006) investigated how survey measures obtained from principals compared to measures of the same constructs obtained from other respondents (teachers and district administrators). The results showed that the responses of teachers, principals, and district administrators converged on questions about power and barriers to policy implementation but either diverged or were uncorrelated on questions regarding authority and the consequences of educational policy. Desimone (2006) acknowledged that the observed differences between different groups of respondents reflect differences in "the perceptions and experiences of respondents at a particular level" (p. 667). Thus it is not clear from this research whether this divergent evidence reflects survey invalidity or valid variation in the perceptions of different groups.

Measurement of Principal Practice

In this study we distinguish between two strategies that have been used to measure principal leadership practice. The first involves closed-ended, self-administered instruments that are implemented either once or given annually as part of a longitudinal study. The reference periods for these instruments are typically one school year. We refer to these instruments as surveys. The second strategy is to administer daily instruments. Daily instruments that have been completed by teachers in prior education studies are also self-administered surveys containing mostly closed-ended items, but the reference period for these instruments is typically a single school day. Like other studies, we refer to these instruments as logs (e.g., Rowan, Camburn, & Correnti, 2004; Smithson & Porter, 1994).

In this study we have chosen to use measures from a daily principal log as a benchmark against which to evaluate measures from an annual principal survey.

This choice has a basis in research on measurement error in surveys. Groves et al. (2004) discussed three ways that measurement error can arise in reports of behaviors in self-administered surveys. Each avenue for error reflects the fact that the survey response process involves an interaction between respondents and instruments. Consequently, measurement error can arise from respondents or survey instruments, or from interaction between the two. In order to report a behavior on a survey at some later date, respondents must first encode information about the behavior (such as its frequency, date, or context) into memory. Measurement error can arise when surveys ask respondents to report behaviors that have not been encoded into memory or when they seek information that respondents cannot accurately retrieve. Often in such cases respondents use estimation strategies that can be inaccurate. Second, measurement error can arise when respondents misunderstand survey items. Misunderstanding can be attributable to respondents (e.g., insufficient literacy skills or poor eyesight) or poorly worded questions. A third occasion for measurement error occurs when respondents translate the information about a behavior they have retrieved from memory into a survey's response choices. Rating scales such as those found in the principal survey appear to be associated with distinctive kinds of measurement error, such as an inclination to choose more positive response categories (Groves et al., 2004).

Survey reports of the kinds of behaviors that comprise principals' daily work are likely prone to all three kinds of error, though there are indications in the literature that measurement error for daily and annual instruments differs in important ways. One of the greatest differences between the two instruments is the way in which respondents must recall past behaviors from memory in order to report them on an instrument. In reporting behaviors on a survey, respondents must search their memories of past events. Such memories are either stored with contextual information that is specific to the event (who, what, where) or as generic "categories of events and stereotypical sequences of events" (Tourangeau et al., 2000, p. 69). Tulving (1983) contended that event-specific information is stored in *episodic memory* while generic information about events is stored in *semantic memory.* Responses to a survey that require respondents to recall events from episodic memory are believed to be more accurate than responses that require respondents to recall events from semantic memory (Menon, 1994). We posit that because the recall period for the daily logs is much shorter than the recall period for the surveys, principals' answers to the daily log will more likely be drawn from episodic memory and their answers to the annual survey from semantic memory. In light of this difference, we further posit that principals' log responses will tend to be more accurate than their survey responses.

The rating scales on the daily log and the survey differ substantially, and these differences may be associated with measurement error in either or both instruments. In general, respondents tend to gravitate to the positive ends of scales and tend to avoid scale extremes, either positive or negative. The number of categories in rating scales also appears to affect responses: Items with more categories tend to discriminate between respondents more than items with fewer categories (Krosnick & Fabrigar, 1997). Principals' tasks of translating their answers into rating scales were quite different on the survey than with the log. In the log, principals used a 4-point rating scale to report how much time they spent on a leadership function in a given hour. The scale categories corresponded to intuitive time blocks within an hour: 1–14 minutes, 15–29 minutes, 30–44 minutes, and 45–60 minutes. In contrast, on the principal survey,

respondents used a 6-point rating scale to indicate how frequently they worked on specific leadership tasks throughout the course of the school year: "Never," "A few times throughout the year," "A few times per month," "1–2 days per week," and "More than 2 days per week." Researchers using the same log data used here found that principals' emphasis on different leadership tasks can vary quite a bit from day to day (Camburn et al., 2010). The response categories on the principal survey imply a uniform distribution of emphasis on tasks over the course of the year that may not correspond well with a more variable emphasis common to some principals. Given that some principals may have had difficulty making a direct translation between their memories about their practice and these rating scale categories, the categories themselves may have been a source of measurement error.

The tasks of responding to the two instruments also differed in a second important way. In the log, single items were used to measure principals' emphasis in the four broad domains of leadership. In contrast, the principal survey measured principals' emphasis in these broad areas by asking principals how frequently they performed specific tasks within the broad areas. The list of tasks on which principals reported was intended to capture key exemplary tasks, but the list was not exhaustive. Principals' reports about specific tasks were subsequently combined to measure the overall emphasis in the broader leadership domains. When reporting their emphasis in a domain on the log, principals might not have considered specific tasks that were asked about in the survey, or they may have considered specific tasks that were not asked about in the survey. These differences in measurement strategies may have contributed to different kinds of measurement error in the two instruments and may also have had the effect of reducing correlations among log and survey measures.

In sum, both theoretical (Menon, 1994; Tulving, 1983) and empirical (Hilton, 1989; Lemmens et al., 1988; Lemmens & Tan, 1992) literature suggest a plausible explanation for observed differences in measures produced by the two instruments, that log measures are more accurate because recall is more accurate with logs. As the preceding discussion makes clear, however, recall errors are not the only potential explanation of differences between the two instruments. A significant limitation of our study design is that it does not allow us to definitively pinpoint which sources of measurement error have caused the differences that we observed. Therefore, we must allow for a range of possible explanations at this stage of our research.

Sample, Data, and Method

This study was conducted in a midsized urban school district as part of an evaluation of an executive training program for principals. At the time of the study there were approximately 55 schools in the district. Because the study was conducted in a single district, the results may not be broadly generalizable. However, we also view this condition as a strength in that it holds the district policy context constant, in effect controlling for key context variables.

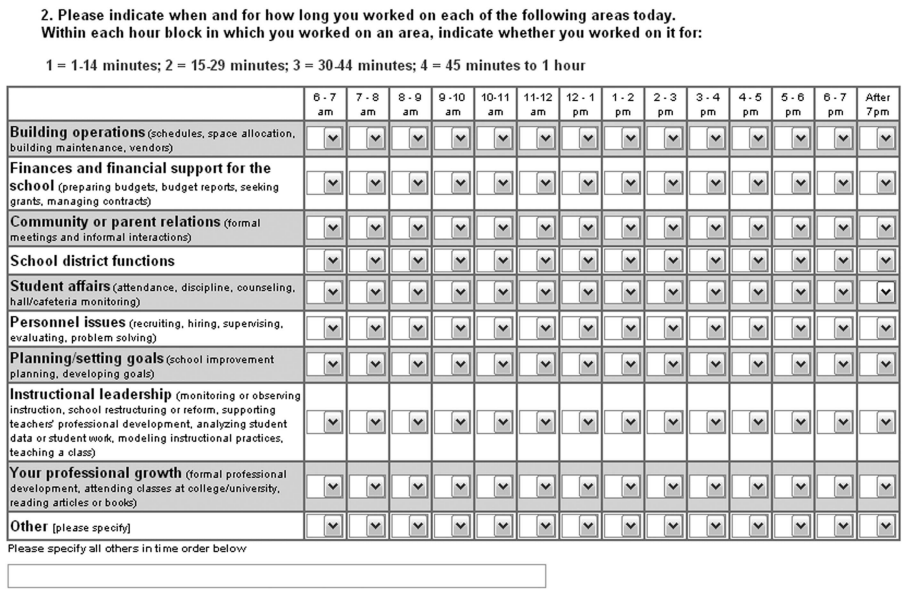
Sixty-five percent of the district's students are African American, and most of the remaining students are white. Fifty-five percent of the district's students receive free or reduced-price lunches. The sample for this study included 46 principals. Of these, 63% worked in elementary schools, 19% in middle schools, and 12% in high schools.

Slightly more than half (54%) of the principals in the sample were white, while most of the remaining principals were African American (44%). Principals in the sample had an average of 10 years of experience as building administrators.

The focus of this study is four dimensions of leadership conceptualized as part of an overarching construct we refer to as *principal leadership practice.* These dimensions of practice thus represent four constructs of interest. In their multitrait-multimethod framework, Campbell and Fiske (1959) described conceptual targets of empirical measures as traits. However, in order to avoid confusion with the literature on leadership traits, we have chosen to refer to the four leadership practices under investigation as constructs rather than traits.

We view principal leadership practice as actions taken by principals to influence people, processes, and organizational structures, and we view the influence of principal leadership practice as being exercised through multiple domains of responsibility (similar characterizations of principal leadership have a considerable history in the literature; see, e.g., Drake & Roe, 2003; Martin & Willower, 1981; Peterson, 1977). The four constructs assessed in this study thus correspond with four important domains of responsibility through which principals can exercise leadership through their actions. The two methods used for this study, annual surveys and daily logs, were designed to measure principals' emphasis on nine domains of leadership responsibility: (1) building operations, (2) finances, (3) community/parent relations, (4) school district functions, (5) student affairs, (6) personnel issues, (7) planning/setting goals, (8) instructional leadership, (9) professional growth. These domains were chosen after consulting a wide range of studies that developed comprehensive frameworks classifying the domains of responsibility of principals' work (Drake & Roe, 2003; Hallinger & Murphy, 1985; Heck & Marcoulides, 1992; Larsen & Hartry, 1987; Martin & Willower, 1981; Peterson, 1977; Pitner & Hocevar, 1987). The nine domains measured by the end-of-day log and annual principal survey are believed to exhaustively cover the range of principal responsibilities. The four constructs investigated for this study correspond with a subset of the nine domains: building operations, planning/setting goals, instructional leadership, and professional growth.

We chose the four constructs in part because they provide the strongest basis of comparison across the two measurement methods as the constructs were defined and measured in a similar fashion on the two instruments. The constructs were also chosen because they represent fundamental dimensions of principal practice that have a direct bearing on the work of students and staff in schools. A basic responsibility of principals is to manage the school's physical plant and staff in a way that supports students and the work of faculty and staff (Leithwood et al., 2004). Sufficient equipment, efficient schedules, safe public spaces, and the like combine to create conditions that can impact all facets of school life. Helping set and implement their schools' vision is a second fundamental way principals influence the work of students and teachers (Hallinger & Heck, 1998). A principal's actions in this domain include efforts to set the school's vision as well as long-term planning to guide the school's progress toward that vision. Thus, principals who actively shape their schools' vision not only plan and evaluate but also engage in actions that focus their faculties on these goals. Instructional leadership is a third critical domain in which principals can influence the teaching and learning that goes on in their buildings. Instructional leadership encompasses key subareas that include coordinating the school's curriculum, creating opportunities and conditions in which teachers can improve their teaching practice, and monitoring the quality of classroom instruction.



2. Please indicate when and for how long you worked on each of the following areas today. Within each hour block in which you worked on an area, indicate whether you worked on it for:

1 = 1-14 minutes; 2 = 15-29 minutes; 3 = 30-44 minutes; 4 = 45 minutes to 1 hour

	6 - 7 am	7 - 8 am	8 - 9 am	9 -10 am	10-11 am	11-12 am	12 - 1 pm	1 - 2 pm	2 - 3 pm	3 - 4 pm	4 - 5 pm	5 - 6 pm	6 - 7 pm	After 7pm
Building operations (schedules, space allocation, building maintenance, vendors)														
Finances and financial support for the school (preparing budgets, budget reports, seeking grants, managing contracts)														
Community or parent relations (formal meetings and informal interactions)														
School district functions														
Student affairs (attendance, discipline, counseling, hall/cafeteria monitoring)														
Personnel issues (recruiting, hiring, supervising, evaluating, problem solving)														
Planning/setting goals (school improvement planning, developing goals)														
Instructional leadership (monitoring or observing instruction, school restructuring or reform, supporting teachers' professional development, analyzing student data or student work, modeling instructional practices, teaching a class)														
Your professional growth (formal professional development, attending classes at college/university, reading articles or books)														
Other [please specify]														

Please specify all others in time order below

Figure 1. Screen shot of the principal daily Web log.

Leadership scholars have long believed that instructional leadership is one of the greatest sources of leverage principals have on student learning (Leithwood & Jantzi, 2000). Finally, developing new knowledge and skills through ongoing professional learning is an important way principals can improve their capacity to manage the school and effect improvements (Peterson & Kelley, 2002).

We used two methods to measure the four constructs of principal leadership practice—a daily log instrument and an annual principal survey. Principals completed daily logs during three periods (fall, winter, and spring) during the 2006–2007 school year. During each period they completed one log per day for 5 consecutive school days. The log is a Web-based instrument that captured how principals allocated their time across the nine domains mentioned earlier. Using a calendar interface, principals reported how they allocated their time across different categories of leadership practice between the hours of 6 a.m. and 7 p.m. Figure 1 shows a screen shot of this calendar interface. The log also collected more specific information about how principals spent their time within these domains, such as recording the staff members principals worked with during each hour of the day. Overall, principals completed 78% of the log instruments they were asked to complete. The accuracy of the log instrument was demonstrated in a study that compared log estimates to those from an experience-sampling instrument (Camburn et al., 2010).

Measures of the four constructs that were based on log data were created by first calculating the number of minutes principals spent on a domain on a particular day and then averaging these daily estimates over the course of the school year. To calculate reliability statistics required for MTMM matrices, we viewed the daily estimates as multiple items used to calculate principals' scores. The log data thus have a nested structure—multiple daily reports (items) are nested within principals. Given the nested structure of the data, we used two-level hierarchical linear models (HLM) to estimate reliability statistics. Specifically, for each measure we fit a two-

level model nesting daily reports of the number of minutes spent on a leadership domain (level 1) within principals (level 2). Intercepts for these models indicate a principal's estimated annual score on a particular measure. Reliability statistics for these models thus represent the reliability for these annual scores (Raudenbush & Bryk, 2002).

The reliability statistics for variables measured with the principal log ranged from .31 to .77. A consideration of reliability statistics for two measures that fell below .60 (professional growth = .31, planning/setting goals = .57) is in order. With repeated measures such as these, reliability depends on the internal consistency of the instrument, variation in true score measurements across measurement occasions and across measurement objects (in this case, principals), and the number of measurement occasions. Using data from a daily teacher log, Rowan et al. (2004) demonstrated how reliability substantially decreases as the variation from one measurement occasion to the next increases. We observed the same pattern for the two log measures with low reliability. We found that 89% of the variation in the planning/goal setting variable was associated with day-to-day fluctuations in the amount of time principals spent on this leadership domain. Considerably more variation (96%) in the amount of time principals spent on professional growth was attributable to day-to-day variation. Clearly, principals' emphasis on these two domains of leadership fluctuates greatly over time. These results strike us as sensible because planning and professional development likely occur much less regularly and systematically than activities like managing building operations. Despite the plausibility of this evidence, it is nonetheless important to note that the low reliabilities of these two measures indicate that significantly more than 15 measurement occasions per principal would be needed to measure these two domains reliably.

The second method used in the MTMM analysis was a principal survey that was administered during the spring of 2007. Items for the survey came from two sources: the School Leader Questionnaire, administered as part of the Study of Instructional Improvement (Camburn, Rowan, & Taylor, 2003), and the School Leadership Self Inventory (National Policy Board for Educational Administration, 2000), a self-report inventory consisting of Likert scale items based on the Interstate School Leaders Licensure Consortium (ISLLC) standards of school leadership. The original inventory includes items relating to the content of each of the six ISLLC standards (e.g., "Articulates a vision of student learning for the school community," "Supports a school culture focused on student learning"). The reference period for this survey was the 2006–2007 school year, and the response rate for the principal survey was 75%.

The survey contained items measuring how often ("Never," "A few times throughout the year," "A few times per month," "1–2 days per week," or "More than 2 days per week") principals worked on building operations, instructional leadership, and planning and setting goals throughout the 2006–2007 school year. Each construct was measured with multiple survey items (App. A contains a list of principal survey items used to measure each construct). The measure of principals' emphasis on professional growth was based on a series of items on which principals reported the number of hours they spent in organized professional development activities (e.g., workshops, seminars, institutes, courses) in 2006–2007. Principals provided separate reports on the amount of time spent on professional growth planned and organized by six organizations: the school district, the state education

agency, professional associations, universities or colleges, school reform programs, and the school.

As mentioned previously, like the log the principal survey was designed to measure the four constructs under study. In developing measures of each construct from principal survey data, we first conducted exploratory factor analyses and reliability analyses to examine how well clusters of items worked together to measure the constructs. These analyses identified a small number of items that contributed weakly to measures, and these items were subsequently dropped. Alpha reliabilities for the final measures ranged from .64 to .86.

To investigate the validity of the principal survey, we used procedures outlined by Campbell and Fiske (1959) to create a multitrait-multimethod matrix. Generally speaking, MTMM matrices include intercorrelations for all construct/methods combinations and reliabilities for all measures. The matrix constructed for this study thus included correlations among eight variables: measures of the four constructs based on principal survey data and measures of the four constructs based on log data.

The correlations included in the matrix were adjusted for measurement error in the variables. All variables represent constructs with error. Measurement error has the effect of causing an underestimation, or attenuation, of the true relationship between variables. In order to gain a clearer picture of the relationships between constructs, we corrected for this attenuation using the following formula proposed by Allen and Yen (1979):

$$r'_{xy} = \frac{r_{xy}}{\sqrt{r_{xx}r_{yy}}},$$

where r_{xy} is the raw correlation between variables x and y, r_{xx} is the reliability of variable x, and r_{yy} is the reliability of variable y.

Investigators formulate hypotheses about how a construct is related to other constructs, and the classic strategy for assessing the validity of a measure is to examine whether hypothesized relationships between constructs are borne out empirically. When predicted patterns of relationships occur, the results are viewed as convergent validity. Typical tests for convergent validity include correlating measures of the same construct obtained through different methods, and correlating measures of different constructs that are hypothesized to be related to each other. In formulating MTMM analysis, Campbell and Fiske (1959) argued that convergent evidence was an insufficient means of establishing validity and that discriminant evidence was also needed. While convergent validation techniques seek evidence of expected correlations between constructs believed to be related, discriminant validation techniques seek evidence of the lack of correlation between constructs believed to be unrelated or less strongly related to each other.

Campbell and Fiske (1959) described four requirements for evidence to be interpreted as demonstrating the validity of a measure, two of which entail convergent evidence, and two discriminant evidence. Arguably, the most important evidence of convergent validity in an MTMM matrix are correlations between measures of the same construct produced by different methods. Campbell and Fiske (1959) contended that these correlations, which they referred to as the *validity diagonal,* should be "significantly different from zero and sufficiently large to encourage further examination of validity" (p. 82). A second source of convergent evidence

lies with the correlations between measures of different constructs. Campbell and Fiske (1959) argued that consistency in patterns of interrelationships between constructs that are produced by different methods combinations provide evidence of convergent validity.

Campbell and Fiske (1959) further outlined two sources of discriminant evidence that are weighed against the validity diagonal of an MTMM matrix. They argued that a validity value for a variable (i.e., its correlation with a variable that measures the same construct but with a different method) should be higher than correlations between that variable and other variables "having neither trait nor method in common" (Campbell & Fiske, 1959, p. 82). They further stipulated that a validity value for a construct should be higher than the correlations between that construct and other constructs measured with the same method. It is clear from Campbell and Fiske's (1959) stipulations regarding evidence of discriminant validity that one gains considerable leverage in interpretations of discriminant evidence when hypotheses about relationships among constructs are strong. However, in our view, no strong theories nor empirical evidence exist on which to base clear predictions about the magnitude or direction of correlations between the four constructs.

Analyses of the MTMM matrix are organized around the requirements for evidence of validity just discussed. To facilitate discussion of the results, we frame the evidence requirements as questions: (1) What are the correlations of measures of the same construct produced by different methods? (2) What is the consistency of correlation patterns between constructs produced with different methods combinations? (3) Are validity values for a construct higher than correlations between that construct and others that are produced by a different method? (4) Are validity values for a construct higher than correlations between that construct and other constructs produced by the same method? Evidence bearing on each of these questions is used to gauge the validity of measures of principal leadership practice from the principal survey.

Results

The complete MTMM matrix is displayed in Table 1. To streamline the presentation of results, we also present separate tables of results that are excerpts from the full matrix that bear on each of the four questions just discussed.

We first examine correlations between measures of the same construct produced by the two methods under investigation. These so-called validity values are marked with the superscript *b* in Table 1. Recall that Campbell and Fiske's (1959) first criterion for convergent validity was that these correlations had to be significantly different from zero and sufficiently large. While it is not appropriate to test the statistical significance of correlations that have been corrected for attenuation (Muchinsky, 1996), we can get some understanding of whether Campbell and Fiske's (1959) first criterion is met by considering the statistical significance of the uncorrected coefficients. We found that for all constructs except professional growth, uncorrected correlation coefficients significantly exceeded zero as indicated by significance levels of two-tailed tests of .05 or less. The validity values for all four constructs exceeded .50, which is well within the range of validity values for many of the studies reported in Campbell and Fiske (1959). We

Table 1. Multitrait-Multimethod Matrix

	Principal Daily Logs				Principal Questionnaire			
	Building Operations	Instructional Leadership	Planning/Setting Goals	Professional Growth	Building Operations	Instructional Leadership	Planning/Setting Goals	Professional Growth
Principal daily logs:								
Building operations	.770[a]							
Instructional leadership	−.343	.756[a]						
Planning/setting goals	−.064	.680	.567[a]					
Professional growth	.173	.409	.534	.307[a]				
Principal questionnaire:								
Building operations	.513[b]	−.542	.129	.569	.695[a]			
Instructional leadership	−.323	.631[b]	.664	.173	.166	.860[a]		
Planning/setting goals	−.119	.513	.556[b]	.080	.128	.822	.858[a]	
Professional growth	.044	.250	.528	.666[b]	.096	.232	.057	.640[a]

[a] Reliability coefficient.

[b] Validity coefficient.

Table 2. Correlations between All Construct Combinations for All Methods Combinations

Construct 1	Construct 2	Methods Combinations[a] C1 = Survey C2 = Survey	C1 = Log C2 = Log	C1 = Survey C2 = Log	C1 = Log C2 = Survey
Building operations	Instructional leadership	.166	−.343	−.542	−.323
Building operations	Planning/setting goals	.128	−.064	.129	−.119
Building operations	Professional growth	.096	.173	.569	.044
Instructional leadership	Planning/setting goals	.822	.680	.664	.513
Instructional leadership	Professional growth	.232	.409	.173	.250
Planning/setting goals	Professional growth	.057	.534	.080	.528

[a] C1 = construct 1, C2 = construct 2.

interpret these correlations as quite substantial and conclude that they are sufficiently large to encourage further examination of validity.

We also computed separate validity coefficients for principals in elementary schools and those in other kinds of schools (middle schools and high schools). The results for these subgroups were very similar to correlations for the sample as a whole, with two exceptions. The validity coefficient for building operations for middle and high school principals was slightly higher than the coefficient for the entire sample (.635 vs. .513), and the validity coefficient for professional growth for elementary school principals was considerably lower than the overall coefficient (.326 vs. .666). We encourage caution in interpreting these results as the samples used to compute correlations for these subgroups were quite small.

We next take up the second question of convergent validity by examining how patterns of intercorrelation among constructs vary across different combinations of methods. Table 2 displays correlations for each unique pair of constructs. For each pair the table presents the following correlation coefficients: (1) the correlation between survey measures of the two constructs, (2) the correlation between log measures of the two constructs, (3) the correlation between a survey measure of construct 1 and a log measure of construct 2, and (4) the correlation between a log measure of construct 1 and a survey measure of construct 2. According to Campbell and Fiske (1959), correlations between two constructs should be similar for all methods combinations. We can gain an understanding of the degree to which this criterion is met by comparing correlations within the rows of Table 2.

Generally speaking, correlations between constructs are quite similar across methods combinations. Most of the correlations between building operations and instructional leadership are negative and fall within a narrow range: between −.542 and −.323. All correlations between building operations and planning/setting goals, and most of the correlations between building operations and professional growth, fall within a fairly narrow range: between −.119 and .173. Correlations between instructional leadership and planning/setting goals are all positive and quite strong, all exceeding .50. The correlations between instructional leadership and professional growth are also all positive and fall into a fairly narrow range, from .173 to .409.

However, Campbell and Fiske (1959) stipulated that for patterns such as these to be indicative of validity, they must be consistent across all methods combinations, and there are important exceptions to the general patterns just discussed. While three of the four correlations between building operations and instructional leadership were negative and fell within a fairly narrow range, when both constructs were measured with the principal

Table 3. Correlations between Measures of Different Constructs from Different Methods

	Principal Log			
Principal Survey	Building Operations	Instructional Leadership	Planning/Setting Goals	Professional Growth
Building operations	.513	−.542	.129	.569
Instructional leadership	−.323	.631	.664	.173
Planning/setting goals	−.119	.513	.556	.080
Professional growth	.044	.250	.528	.666

survey the correlation was .166. This is a significant departure from the general pattern, casting doubt on the validity of the survey. Similarly, there is a major exception to the pattern of generally low correlations between building operations and professional growth. When building operations was measured with the survey and professional growth was measured with the log, the correlation between the two measures was .569. It is not clear whether this departure reflects invalidity of the survey, the log, or both. Correlations between instructional leadership and planning/goal setting follow a similar pattern, with correlations for three of the four methods combinations falling between .50 and .68; however, the correlation between these two constructs when they are both measured by the log is substantially higher: .822. We are less troubled by the results for this construct combination than others because the general pattern is similar across methods combinations—all correlations are positive and quite strong.

A final inconsistent pattern was observed for correlations between planning/setting goals and professional growth. While the correlations between these constructs were near zero when each was measured by the survey and when planning/setting goals was measured by the survey and professional growth was measured by the log, the correlations associated with the other two methods combinations did not follow this pattern. Indeed, when both constructs were measured with the log, and when planning/setting goals was measured with the log and professional growth was measured with the survey, correlations between the two constructs slightly exceeded .50. It was not clear to us whether this pattern reflected negatively on the validity of the survey or the log.

We next sought to investigate two forms of evidence of discriminant validity. We first examined whether correlations between constructs measured with the principal survey and those measured with the log exceeded a construct's validity value. In this case, the validity of a measure based on survey data is indicated when correlations with other constructs measured by the log are lower than the validity value. Table 3 displays correlations that bear on this issue. Validity values are presented in the diagonal of the table. Comparing these validity values to the other correlations within a row indicates the degree to which Campbell and Fiske's (1959) second criterion was met—that correlations between measures of the same trait captured by different methods should exceed correlations of different traits captured by different methods.

The results in Table 3 indicate that these criteria were met for two of the four constructs. The validity value of .556 for planning/setting goals exceeds the correlations between the survey measures of planning/setting goals and log-based measures of the other three constructs. The same pattern was observed for professional growth. For instructional leadership, the validity value is the second highest corre-

Table 4. Comparison of Validity Values with Correlations among Variables Measured with the Principal Survey

Construct	Validity Value	Building Operations	Instructional Leadership	Planning/Setting Goals	Professional Growth
Building operations	.513	1	.166	.128	.096
Instructional leadership	.631	.166	1	.822	.232
Planning/setting goals	.556	.128	.822	1	.057
Professional growth	.666	.096	.232	.057	1

lation by a relatively small margin. The validity value for building operations exceeds a positive correlation between that construct and planning/setting goals but fails to exceed the correlation between building operations and professional growth. In contrast, the log-based measure of instructional leadership is negatively associated with the survey-based measure of building operations, and the magnitude of that correlation exceeds the validity value. Thus, strictly speaking, this mixed evidence for building operations cannot be interpreted as reinforcing the validity of the measurement of this construct using the principal survey.

An interesting result emerging from this analysis was the relatively strong correlations between instructional leadership and planning/setting goals across different combinations of methods. This finding may be a reflection of the sample that was used for this study. The executive training program in which some principals participated placed a strong emphasis on the use of planning for facilitating instructional improvement. Thus the relatively strong correlations between planning and instructional leadership may reflect a connection between these two leadership domains that the program was attempting to make. Given that there were forces in the district to make this connection, we view the pattern of correlations between these two constructs as convergent evidence for the validity of measures of these constructs.

A final analysis examines a second form of discriminant validity—evidence that the validity value for a construct exceeds correlations of that construct with other measures captured using the same instrument, which is Campbell and Fiske's (1959) third criterion. Table 4 presents correlations that address this validation strategy. The left-hand side of Table 4 contains the validity value for each construct and the main body of the table contains the correlation matrix for all variables measured with the principal survey. As with previous analyses, our focus is within table rows, and we are examining whether the validity value for a construct exceeds correlations with other constructs.

As Table 4 indicates, we find that Campbell and Fiske's (1959) second criterion for discriminant validity is met for two of the four constructs. The validity values for building operations and professional growth both exceed the correlations between survey-based measures of these two constructs and survey-based measures of other constructs. Evidence for the instructional leadership construct failed to meet this validity criterion, as the correlation of .822 between the survey-based measures of instructional leadership and planning/setting goals was higher than the .631 validity value for instructional leadership. The strong correlation between instructional leadership and planning/setting goals also exceeded the validity value for planning/setting goals. As discussed earlier, the strong correlation between survey-based measures of instructional leadership and planning/setting goals may indicate the influence of the executive training program operating in

the district at the time of the study, which emphasized the use of planning as a vehicle for instructional improvement.

Discussion

Campbell and Fiske (1959) framed their MTMM criteria as requirements that must be met in order to establish validity. Like most MTMM studies, our results did not meet all required criteria. Thirty-three years after publishing their original 1959 paper, Campbell and Fiske acknowledged that meeting all MTMM criteria is rare and that even when this occurs, "validity coefficients were typically .30 to .50, disappointingly low" (Fiske & Campbell, 1992, p. 393). Rather than using the results to render a summary judgment, we think it is more informative to use the richness of the MTMM results to shed more nuanced light on the strengths and weaknesses of the principal survey.

In our view, the MTMM analysis provided considerable evidence attesting to the validity of the principal survey. Using a common strategy of benchmarking annual surveys against a daily instrument, we found that validity values for all four constructs, which represent four important aspects of principals' work in schools, exceeded .50. This is a slightly stronger result than most of the studies reported in Campbell and Fiske (1959). Uncorrected validity values for all constructs except one (professional growth) met Campbell and Fiske's (1959) requirement that validity values should differ significantly from zero. Campbell and Fiske's (1959) third criterion—validity values for a construct should exceed correlations between that construct and another when measures come from the same method—was met for two of the four constructs. For the other two constructs, the criterion was not met because of a high correlation between instructional leadership and planning/setting goals, a result that might be unique to the sample used for this study because of the influence of a principal development program operating in the district under study. Measures from the survey also appear not to have been deeply affected by mono-method bias (unusually strong correlations among variables measured with the same method), a problem experienced in many other MTMM studies, and a concern in fields like education in which studies often rely exclusively on survey data (Donaldson et al., 2000; Mersman & Donaldson, 2000). Campbell and Fiske's (1959) fourth criterion, that correlations between constructs should be similar regardless of the combination of methods used, was also largely met. Correlation patterns for three of the six construct combinations were generally consistent across methods combinations. Correlation patterns for two of the six construct combinations were also generally consistent, but a single significant deviation from these patterns for these two construct pairs meant that Campbell and Fiske's (1959) criterion was not met.

Our results also suggest that annual surveys may be better suited for measuring some dimensions of principal leadership practice than others. Evidence from the daily log suggests that principals engage in planning and personal professional development activities less frequently than building operations and instructional leadership, and that their engagement in the former two activities occurs at irregular time intervals. Irregular patterns in practice are likely difficult for principals to remember and accurately report on in a survey, and consequently, investigators may obtain more valid measures of constructs like these from daily instruments. In contrast, less costly and less burdensome annual surveys may provide sufficient measurement of

practices that occur with greater regularity, such as building operations and instructional leadership. As our evidence illustrates, however, measuring irregularly occurring activities accurately comes at an additional cost, as reliably measuring such activities requires that measurements be obtained on a greater number of occasions.

While we found considerable evidence supporting the validity of the principal survey, we believe there is still ample room for improving survey measurement of principal leadership practice. One potential avenue for improvement is empirical research that furthers our understanding of how principals allocate their time across different leadership domains. Two recent studies have used latent class models and cluster analysis techniques to produce profiles that classify principals in this regard (Barnes, Camburn, Sanders, & Sebastian, 2010; Goldring, Huff, May, & Camburn, 2009). The studies, which use log data from the same sample of principals studied here, found that all principals appear to consistently spend considerable time running their buildings, that a subset of principals focus significantly more time on instructional leadership, and that this latter group places slightly less emphasis on building operations. We believe that more studies such as these can support the evaluation of the validity of survey-based measures of principal leadership practice by yielding insight into patterns of intercorrelation among principals' emphasis on different leadership domains. Having greater knowledge of such patterns will bolster investigators' ability to draw stronger inferences in validity studies such as this one.

The marked lack of research on the validity of principal surveys also points to a strong need for more and different kinds of research on this topic. Patterns in our evidence led us to believe that the relationship between instructional leadership and planning/setting goals may have reflected idiosyncrasies in the study's sample. In light of this limitation, validity studies based on larger samples of principals collected across a broader range of settings are sorely needed. We also believe that mixed-method validity studies could shed important light on the validity of principal surveys. Having observers complete self-report surveys during the same time frame as respondents and conducting postsurvey follow-up interviews have been shown to be useful strategies for assessing the validity of self-report instruments (e.g., Camburn & Barnes, 2004). Finally, we believe that studies that assess the predictive validity of principal surveys are needed. Such studies would conduct parallel sets of analyses where independent measures of principal leadership practice obtained from a survey and some other method would be used to predict the same outcome variables hypothesized to be related to principal leadership (e.g., teacher motivation or student achievement). Patterns of observed relationships for the two analyses could be compared and, if similar, would be interpreted as evidence of the validity of independent variables measured with the principal survey.

The ubiquity of one-time self-report surveys in research on principal leadership merits a clear understanding of their validity. The use of such surveys is understandable. They can validly measure important dimensions of leadership practice and do so at a lower cost and lower respondent burden than other measurement strategies. Like others, we conclude that principal surveys have a useful place in the methodological tool kits of researchers studying leadership and organizations (Howard, 1994; Spector, 1994). However, we also conclude that much more evidence about the validity of this tool is needed so that the evidence it produces can be better understood and improved over time. The hope is that having an improved understanding of this tool will help researchers gain much needed insight into the ways principal leadership practice impacts student learning, teaching, and school operations.

Appendix A

Table A1. Items Used to Measure Traits

Trait	Daily Log	Principal Survey Questionnaire Items
Building operations	Reliability = .77 Building operations (schedules, space allocation, building maintenance, vendors, student affairs)	Alpha coefficient for reliability = .70 Supervise clerical, cafeteria, and maintenance staff Monitor public spaces, such as the cafeteria, hallways, playgrounds, etc. Deal with emergencies and other unplanned circumstances Work with students and their parents on discipline/attendance issues Complete routine paperwork (such as reports and record keeping)
Instructional leadership	Reliability = .76 Instructional leadership (monitoring or observing instruction, school restructuring or reform, supporting teachers' professional development, analyzing student data or student work, modeling instructional practices, teaching a class)	Alpha coefficient for reliability = .86 Demonstrate instructional practices and/or the use of curricular materials in a classroom Observe a teacher who was trying new instructional practices or using new curricular materials Examine and discuss what students were working on during a teacher's lesson Examine and discuss standardized test results of students from a teacher's class Develop the staff development program in the school Personally provide staff development Troubleshoot or support the implementation of school improvement efforts Monitor the curriculum used in classrooms to see that it reflects the school's improvement efforts Monitor classroom instructional practices to see if they reflect the school's improvement efforts Frequency with which data are used for identifying and correcting gaps in the curriculum for all students Frequency with which data are used for identifying areas where teachers need to strengthen their content knowledge or teaching skills Frequency with which data are used for determining topics for professional development
Planning/setting goals	Reliability = .57 Planning/setting goals (school improvement planning, developing goals)	Alpha coefficient for reliability = .86 Frame and communicate broad goals for school improvement Examine the school's overall progress toward its school improvement goals Set explicit timelines for instructional improvement Clarify expectations or standards for students' academic performance Work on plans to improve the teaching of specific curricular units or objectives
Professional growth	Reliability = .31 Your professional growth (formal professional development, attending classes at college/university, reading articles or books)	Alpha coefficient for reliability = .64 This year, how much time did you spend in PD activities organized by . . . Your school district The state education agency A professional association A school reform program Your school Other

Note

This research was funded by the Institute of Education Sciences, U.S. Department of Education (grant R305E040085). Please address correspondence regarding this article to Eric Camburn, 270E Education Building, 1000 Bascom Mall, Madison, WI 53706-1326. E-mail: camburn@wisc.edu.

References

Allen, M. M., & Yen, W. M. (1979). *Introduction to measurement theory.* Long Grove, IL: Waveland.

Barnes, C. A., Camburn, E. M., Sanders, B. R., & Sebastian, J. (2010). School leaders as learners: Acquiring expertise for improving teaching and learning. *Educational Administration Quarterly,* **46**(2), 241–279.

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

Burstein, L., McDonnell, L., Van Winkle, J., Ormseth, T., Mirocha, J., & Guitton, G. (1995). *Validating national curriculum indicators* (MR-658-NSF). Santa Monica, CA: RAND.

Camburn, E., & Barnes, C. (2004). Assessing the validity of a language arts instruction log through triangulation. *Elementary School Journal,* **105**, 49–74.

Camburn, E. M., & Han, S. W. (2011). Two decades of generalizable evidence on U.S. instruction from national surveys. *Teachers College Record.* Retrieved from http://www.tcrecord.org/content.asp?contentid=16064

Camburn, E., Rowan, B., & Taylor, J. (2003). Distributed leadership in schools: The case of elementary schools adopting comprehensive school reform models. *Educational Evaluation and Policy Analysis,* **25**(4), 347–373.

Camburn, E. M., Spillane, J. P., & Sebastian, J. (2010). Assessing the utility of a daily log for measuring principal leadership practice. *Educational Administration Quarterly.* Prepublished September, 2010; DOI: 10.1177/0123456789123456

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin,* **56**, 81–105.

Claudet, J. G., & Ellett, C. D. (1993). *Developing, measuring and testing an organizational model of instructional supervision: Implications for administrative leadership.* Paper presented at the annual meeting of the University Council for Educational Administration, Denver.

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**(5), 439–461.

Desimone, L. M. (2006). Consider the source: Response differences among teachers, principals, and districts on survey questions about their education policy environment. *Education Policy,* **20**(4), 640–677.

Donaldson, S. I., Ensher, E. A., & Grant-Vallone, E. J. (2000). Longitudinal examination of mentoring relationships on organizational commitment and citizenship behavior. *Journal of Career Development,* **26**(4), 233–349.

Drake, T. L., & Roe, W. I. H. (2003). *The principalship.* Upper Saddle River, NJ: Merrill/Prentice-Hall.

Dryden, J. (1995). *Design of a quality schools survey instrument through the use of a four-cycle action research process.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Fiske, D. W., & Campbell, D. T. (1992). Citations do not solve problems. *Psychological Bulletin,* **112**(3), 393–395.

Goldring, E. B., Huff, J. T., May, H., & Camburn, E. (2009). School context and individual characteristics: What influences principal practices? *Journal of Educational Administration,* **46**(3), 332–352.

Gronn, P. (2003). *The new work of educational leaders: Changing leadership practice in an era of school reform.* London: Sage/Paul Chapman.

Groves, R. M., Fowler, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2004). *Survey methodology.* Hoboken, NJ: Wiley.

Hallinger, P., & Heck, R. (1996). Reassessing the principal's role in school effectiveness: A review of empirical research, 1980–1995. *Educational Administration Quarterly,* **32**, 5–44.

Hallinger, P., & Heck, R. H. (1998). Exploring the principal's contribution to school effectiveness: 1980–1995. *School Effectiveness and School Improvement,* **9**(2), 57–91.

Hallinger, P., & Murphy, J. (1985). Assessing the instructional management behavior of principals. *Elementary School Journal,* **86**(2), 217–247.

Heck, R. H., & Marcoulides, G. A. (1992). Principal assessment: Conceptual problem, methodological, or both? *Peabody Journal of Education,* **68**(1), 124–144.

Henderson, J. E., & Hoy, W. K. (1982). *Leadership authenticity: The development and test of an operational measure.* Paper presented at the annual meeting of the American Educational Research Association, New York.

Hilton, M. E. (1989). A comparison of a prospective diary and two summary recall techniques for recording alcohol consumption. *British Journal of Addiction,* **84**(9), 1085–1092.

Howard, G. S. (1994). Why do people say nasty things about self-reports? *Journal of Organizational Behavior,* **15**(5), 399–404.

Huff, J. T. (2006). *Measuring a leader's practice: Past efforts and present opportunities to capture what educational leaders do.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Krosnick, J., & Fabrigar, L. (1997). Designing rating scales for effective measurement in surveys. In L. Lyberg, P. Biemer, M. Collins, E. de Leeuw, C. Dippo, N. Schwarz, & D. Trewin (Eds.), *Survey measurement and process quality* (pp. 141–164). New York: Wiley.

Larsen, T. J., & Hartry, A. (1987). *Principal/teacher perceptual discrepancy: Instructional leadership in high and low-achieving California schools.* Paper presented at the annual meeting of the American Educational Research Association, Washington, DC.

Leithwood, K., & Duke, D. L. (1999). A century's quest to understand school leadership. In J. Murphy & K. Seashore Louis (Eds.). *Handbook of research on educational administration* (pp. 45–72). San Francisco: Jossey-Bass.

Leithwood, K., & Jantzi, D. (2000). The effects of transformational leadership on organizational conditions and student engagement with school. *Journal of Educational Administration,* **38**(2), 112–126.

Leithwood, K., Seashore Louis, K., Anderson, S., & Wahlstrom, K. (2004). *How leadership influences student learning: A review of research.* New York: Wallace Foundation.

Lemmens, P., Knibbe, R., & 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**(5), 476.

LeSourd, S. J., Tracz, S., & Grady, M. L. (1990). *Validation of a visionary leadership attitude instrument using factor analysis.* Paper presented at the Annual Meeting of the Mid-Western Educational Research Association, Chicago.

Levine, R. E., Chambers, J. G., Ixtlac, E., & Hikido, C. S. (1998). *Improving the measurement of staffing resources at the school level: The development of recommendations for NCES for the Schools and Staffing Surveys (SASS).* Washington, DC: American Institutes for Research.

Martin, W. J., & Willower, D. J. (1981). The managerial behavior of high school principals. *Educational Administration Quarterly,* **17**, 69–90.

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

Mersman, J. L., & Donaldson, S. I. (2000). Factors affecting the convergence of self-peer ratings on contextual and task performance. *Human Performance,* **13**(3), 299–322.

Muchinsky, P. M. (1996). The correction for attenuation. *Educational and Psychological Measurement,* **56**, 63–75.

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 field test results to improve item construction.* Washington, DC: U.S. Department of Education.

National Policy Board for Educational Administration. (2000). *Collaborative professional development process for school leaders.* Interstate School Leaders Licensure Consortium. Washington, DC: Council of Chief State School Officers.

Peterson, K. D. (1977). The principal's tasks. *Administrator's Notebook,* **26**(8), 1–4.

Peterson, K. D., & Kelley, C. (2002). Principal in-service programs: A portrait of diversity and promise. In M. Tucker & J. Codding (Eds.), *The principal challenge: Leading and managing schools in an era of accountability* (pp. 313–346). San Francisco: Jossey-Bass.

Pitner, N. J., & Hocevar, D. (1987). An empirical comparison of two-factor versus multifactor theories of principal leadership: Implications for the evaluation of school principals. *Journal of Personnel Evaluation in Education,* **1**(1), 93–109.

Pounder, D. G., Ogawa, R. T., & Adams, E. A. (1995). Leadership as an organization-wide phenomenon: Its impact on school performance. *Educational Administration Quarterly,* **31**(4), 564–588.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage.

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

Rubin, D. C., & Baddeley, A. D. (1989). Telescoping is not time compression: A model of the dating of autobiographical events. *Memory & Cognition,* **17**(6), 653–661.

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 No. 31). Madison: University of Wisconsin, Consortium for Policy Research in Education.

Spector, P. E. (1994). Using self-report questionnaires in OB research: A comment on the use of a controversial method. *Journal of Organizational Behavior,* **15**(5), 385–392.

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

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

Wildy, H., & Clarke, S. (2009). Using cognitive interviews to pilot an international survey of principal preparation: A Western Australia perspective. *Educational Assessment,* **21**(2), 105–117.

Wildy, H., Forster, P., Louden, W., & Wallace, J. (2004). The international study of leadership in education: Monitoring decision making by school leaders. *Journal of Educational Administration,* **42**(4), 416–430.