

Advancing data collection in the digital age:
Methodological challenges and solutions in educational technology research.

Damian Bebell, Laura M. O'Dwyer, Mike Russell, & Tom Hoffman

Technology and Assessment Study Collaborative: Boston College

April 2007

Paper Presented at the: Annual Meeting of American Educational Research Association Meeting
Chicago, IL

Some of the research summarized in the current paper was supported and conducted under the Field Initiated Study Grant Program, PR/Award Number R305T010065, as administered by the Office of Educational Research and Improvement, U.S. Department of Education. The findings and opinions expressed in this report do not reflect the positions or policies of the Office of Educational Research and Improvement, or the U.S. Department of Education.

Introduction

“The great preponderance of technology evaluation studies to date depend upon relatively weak survey measures...” *Baker & Herman, 2000, p. 7*

There are few modern educational initiatives that have been as widespread, dramatic, and costly as the integration of computer technologies into American classrooms. Believing that increased use of computers will lead to improved teaching and learning, greater efficiency, and the development of important skills in students, educational leaders and policy makers have made multi-billion dollar investments in educational technologies such that the national ratio of students to computers has dropped from 125:1 in 1983 to 4:1 in 2006 (US Census Bureau, 2006). More recently, between 1997 and 2003, the percentage of American classrooms connected to the Internet grew from 27% to 93%. In 1997, 50% of schools used a dial-up connection to connect to the Internet and only 45% had a dedicated high-speed Internet line. By 2003, less than 5% of schools were still using dial-up connections while 95% reported broadband access. In a relatively short time period, computer based technologies have become commonplace across all levels of the American educational system. Given these substantial investments in educational technology, it is not surprising that there has been increasing calls for empirical, research-based evidence that these massive investments are affecting the lives of teachers and students (McNabb, Hawkes, & Rouk, 1999 – what about Cuban’s piece in Ed Week in 2006?).

As increased access and more powerful technology entered America’s classrooms, the variety of ways and the degree to which teachers and students applied this new technology increased exponentially. For example, several advances in computer-based technologies came together in the mid 1990’s that allowed teachers to use technology to support their teaching in an increasing variety of ways. Whereas instructional uses of computers had been limited largely to word processing and computer programming, teachers were now able to perform multi-media presentations and computer-based simulations. With the introduction of the Internet into the classroom, teachers were also able to incorporate activities that tapped the resources of the World Wide Web. Outside of class time, software for record keeping, grading and test development provided teachers with new ways of using computers to support their teaching. In addition, the Internet allowed teachers access to additional resources when planning lessons (Becker, 1999) and allowed teachers to use email to communicate with their colleagues, administrative leaders, students, and parents (Lerman, 1998).

Following the rise of educational technology resources, hundreds of studies have sought to examine instructional uses of technology across a wide variety of educational settings. Not surprisingly, in today’s zeitgeist of educational accountability the call for empirical, research-based evidence that these massive investments are affecting the lives of teachers and students has only intensified (McNabb, Hawkes, & Rouk, 1999). Despite the large number of studies, many researchers and decision makers argue that much of the current educational technology research is methodologically weak. Baker and Herman (2000), Waxman, Lin, and Michko (2003), Goldberg, Russell, and Cook (2003), and O’Dwyer, Russell, Bebell, and Seeley (2004) have all suggested that much of the educational technology literature suffers from one or more methodological shortcomings. Specifically, educational technology research is often limited by (1) limitations in the way students’ and teachers’ technology use is measured, (2) a lack of prior student achievement measures or comparison groups, (3) a reliance exclusively on paper based assessments in high-tech classroom environments, (4) poor alignment of measurement tools, and (5) a failure to account for the hierarchical nature of schools.

The collective weakness of educational technology research has created a challenging situation for educational leaders and policy makers who must use flawed or limited research evidence to make important decisions. Even in 2007, little empirical research exists to support many of the most cited claims on the effects of educational technology¹. For example, there has yet to be a definitive study that examines the effects of computer use in school on standardized measures of student achievement despite a generation of students being educated with technology. It is a growing problem that as an educational community, research and evaluation efforts have not really adequately addressed the short and long term effects of technology in the classroom. This situation forces decision makers to rely upon weak sources of evidence, if any at all, when allocating budgets and shaping future policy.

In the following pages, the authors detail five broad methodological concerns in educational technology research using examples from the current literature including some emerging examples from the authors' work. Specifically, the authors address how these concerns pose unique challenges to evaluating educational technology programs as well as detail potential techniques and methods for overcoming these challenges in future research and evaluation. Furthermore, as technology grows more ubiquitous for students and teachers, the authors suggest new methodological advances that capitalize on these new resources within a school.

Measuring Specific Technology Uses

Context and Survey Development:

While there is a strong desire to examine the impact of technology on student achievement and other outcome variables, the impacts on learning must first be placed in the context of technology use. In other words, before any outcomes of technology integration can be studied there must be (1) a clear understanding of *how* teachers and students are using technology, and (2) valid and reliable measures of these uses. These measures of technology serve as the independent variable in the research of the effects of educational technology. However, instead of developing measures of technology use, a great deal of literature only examines technology access, assuming that teachers' and students' access to technology is a proxy of their frequency of use.

When technology use is measured, researchers must invest time and effort in developing technology use instruments that exhibit high levels reliability and validity. Surveys, taken first on paper and increasingly through computers, provide the widely used means to measure the frequency of technology use. Survey development poses a particular challenge in technology research where new and novel uses are developed and implemented quickly. As technology uses may be divergent, the survey developer must consider a large number of potential uses for each technology tool to fully evaluate their effectiveness. The inclusion of the widest possible technology use spectrum can be insured through the survey development process. For example, student and teacher interviews, focus groups, shadowing of individual students and teachers, classroom observations, as well as reviews of the existing literature all help to provide adequate coverage in the creation of educational technology surveys. Iterative review and pilot testing are required to ensure

¹ Educational technology research is often divided into two broad categories: (1) research that focuses on effects *with* technology in the classroom and (2) research that focuses on the effects *of* technology integrated into the classroom and teacher practices (Salomon, Perkins & Globerson, 1991). Although not mutually exclusive, this categorization of research can be illuminating. Generally, research concerning the "effects with" technology focuses on the underlying evolution of the learning process with the introduction of technology. On the other hand, research concerning the "effect of" technology seeks to measure (via outcomes testing) the impacts of technology as an efficiency tool rather than focusing on the underlying processes. The current paper concentrated more on the latter category, the "effects of" technology.

both the reliability and validity of the measures. A step-by-step examination of such a survey development process can be found in the USEiT Study Technical Report (Russell, O'Dwyer, Bebell & Miranda; 2003).

Defining Technology Use:

A short historical review of the literature reveals that the way “teachers’ use of technology” has been defined across different studies has resulted in different results. The very first large-scale investigation of educational technology occurred in 1986 when Congress asked the federal Office of Technology Assessment (OTA) to compile an assessment of technology use in American schools. Through a series of reports (OTA, 1988; 1989; 1995), national patterns of technology integration and use were documented. Ten years later Congress requested OTA “to revisit the issue of teachers and technology in K-12 schools in depth” (OTA, 1995). In a 1995 OTA report, the authors noted that previous research on teachers’ use of technology employed different definitions of what constituted technology use. In turn, these different definitions led to confusing and sometimes contradictory findings regarding teachers’ use of technology. For example, a 1992 International Association for the Evaluation of Educational Achievement (IEA) survey defined a “computer-using teacher” as someone who “sometimes” used computers with students. A year later, Becker (1994) employed a more explicit definition of a computer-using teacher for which at least 90 % of the teachers’ students were required to use a computer in their class in some way during the year. Thus, the IEA defined use of technology in terms of the teachers’ use for instructional delivery while Becker defined use in terms of the students’ use of technology during class time. Not surprisingly, these two different definitions of a “computer-using teacher” yielded different impressions of the technology use. In 1992, the IEA study classified 75 % of U.S. teachers as “computer-using teachers” while Becker’s criteria yielded about one third of that (approximately 25 %) (OTA, 1995). This confusion and inconsistency led the OTA to remark: “Thus, the percentage of teachers classified as computer-using teachers is quite variable and becomes smaller as definitions of use become more stringent” (p. 103).

It is clear, both in theoretical and investigative research, that defining and measuring teachers’ use of technology has only increased in complexity as technology has become more advanced, varied, and pervasive in the educational system. Today, several researchers and organizations have developed their own definitions and measures of technology use to examine the extent of technology use and to assess the impact of technology use on teaching and learning. Frequently these instruments collect information on a variety of different types of teachers’ technology use and then collapse the data into a single generic “technology use” variable. Unfortunately, the amalgamated measure may be inadequate both for understanding the extent to which technology is being used by teachers and for assessing the impact of technology on learning outcomes.

There is a strong likelihood that the school leaders who rely upon this information for decision-making will interpret findings in a number of different ways. For example, some may interpret one measure of teachers’ technology use solely as teachers’ use of technology for delivery, while others may view it as a generic measure of the collected technology skills and uses of a teacher. While defining technology use as a unitary dimension may simplify analyses, it complicates efforts by researchers and school leaders to do the following:

- provide valid measures of how technology is being used,
- interpret findings about the extent to which technology is used, and
- understand how to increase technology use.

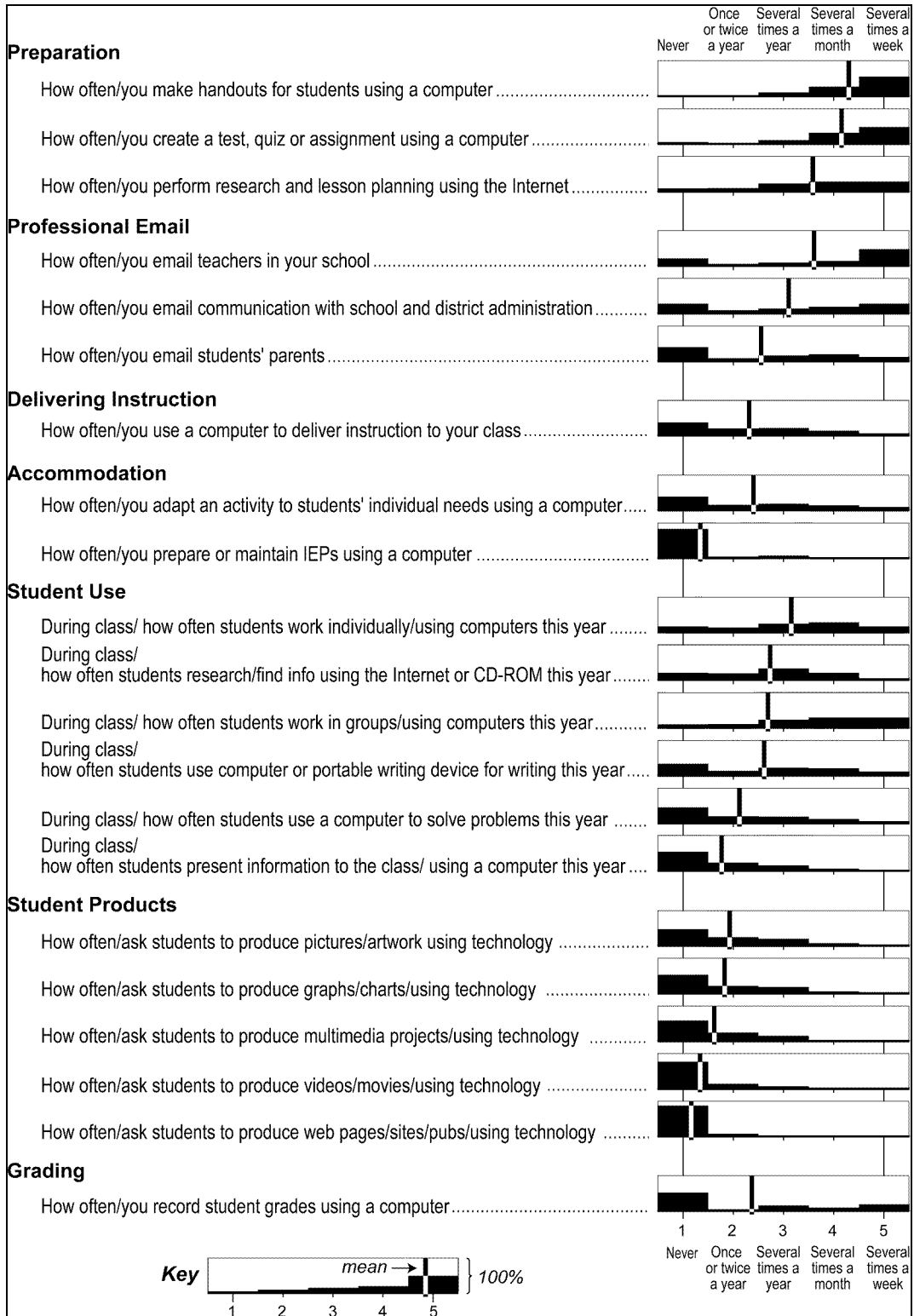
Recognizing the importance of how technology use is both defined and measured, the authors have applied an approach to measuring teacher technology use that involves examining the specific ways in which teachers make use of technology. In this case, multiple measures (i.e. scales) for the specific ways

that teachers use technology are constructed from related survey items. Both Mathews' (1996) and Beckers' (1999) research on teachers' technology use demonstrated that a new level of refinement in the measurement of *specific* technology uses. Similarly, Russell et al. (2003), used principal components analyses to develop seven separate scales that measure teachers' technology use across 2,628 classroom teacher surveys participating in the USEiT study. These seven scales include the following:

- Teachers' use of technology for class preparation (Preparation)
- Teachers' professional email use (Professional Email)
- Teachers' use of technology for delivering instruction (Delivering Instruction)
- Teachers' use of technology for accommodation (Accommodation)
- Teacher-directed student use of technology during class time (Student Use)
- Teacher-directed student use of technology to create products (Student Products)
- Teachers' use of technology for grading (Grading)

These seven categories of teacher technology examined in the USEiT study are displayed in Figure 1 along with the distribution and mean response for each of the items used to form each of the seven scales.

Figure 1: USEiT distribution and mean values across seven categories of teacher technology use



As shown in Figure 1, the number of survey items used to form each category of use ranges from one to five items². Also note that the distribution of responses and mean response varies considerably across the individual items. For example, the distribution of responses for the survey item asking teachers *how often they make handouts for students using computers* is negatively skewed, with the vast majority of teachers reporting that they do this several times a week or several times a month. While examining teacher responses at the item level is informative and may reveal interesting patterns across items, patterns generally become easier to identify when items that focus on related uses of technology are combined into a single measure.

Further analyses of the seven teacher technology use scales showed that each of the individual scales exhibited widely divergent frequency distributions (Bebell, Russell & O'Dwyer, 2004). For example, teachers' use of technology for preparation was strongly negatively skewed (skewness = -1.12) while the same teachers' use of technology for instruction is strongly positively skewed (1.09). Like instructional use, the distributions for assigning student products (1.15) and accommodation (1.04) were positively skewed. Using technology for grading also had a weak positive skew (0.60) while teacher-directed student use (0.11) was relatively normally distributed. Use of email however, presented a bi-modal distribution (≈ -0.18), with a large percentage of teachers reporting frequent use and a large portion of the sample reporting no use. If all of the survey items comprising these scales were summed to create a generic composite measure of technology use, the distribution closely approximates a normal distribution revealing none of the patterns observed in the specific technology use scales.

When compared to a single generic measure of technology use, multiple measures of specific technology use offer a more nuanced understanding of how teachers are using technology and how these uses vary among teachers. By developing separate measures of teachers' technology use the authors are not inferring that each individual measure is unrelated to the other technology use measures. Indeed, it would be reasonable to assume that all of the measures have some degree of relation to each other. The strength of the relationships among the seven technology uses from the USEiT study is examined via Pearson correlation coefficients which are presented in Table 1.

Table 1: Correlation Table of the Seven Specific Teacher Technology Measures

	Accommodation	Delivery	Prof. Email	Preparation	Student Use	Student Products	Grading
Accommodation	1.00						
Delivery	0.26	1.00					
Prof. Email	0.26	0.25	1.00				
Preparation	0.27	0.26	0.35	1.00			
Student Use	0.32	0.47	0.22	0.27	1.00		
Student Products	0.23	0.33	0.18	0.33	0.46	1.00	
Grading	0.11	0.17	0.15	0.24	0.07	0.00	1.00

Table 1 shows that the correlations among the seven teacher technology use measures are all positive, but generally indicate weak to moderate relationships. The positive inter-correlations suggest that teachers who use technology for one purpose are, on average, likely to use technology for other purposes. However, the moderate to weak correlations suggest that there is considerable variation between the

² The use of a scale to measure an attribute has a couple of advantages over the use of a single item. First, measurement relying on a single item will likely contain more error than measurement made with a scale containing multiple items. Second, the use of more than one item to measure an attribute has important consequences for the validity of the measure. A scale, through its individual items, can often represent the multiple aspects of a concept in a single measure, thus increasing the construct validity of the measure.

extent to which teachers use technology for one purpose and the extent to which they use technology for another purpose. The relatively weak to moderate provide evidence that: a) each measure does represent a separate and distinct category of technology use, and b) the frequency and distribution of technology use varies considerably across the seven measures. Research studies that have utilized this multi-faceted approach to measuring technology use have revealed that many illuminative patterns that were obscured when only general measures of use were employed (Mathews, 1996; Ravitz, Wong & Becker, 1999; Bebell, Russell, & O'Dwyer, 2004). For example the analysis of the USEiT teacher data indicated that the frequency of teachers' technology use for instruction and accommodating lessons was unrelated to the frequency of them asking students to use technology during class time. Similarly teachers' use of technology for grading operated independently of teachers' use of technology for preparation of their lessons (Ibid).

To summarize, how technology is defined and measured (if measured at all) plays a substantial, but often overlooked, role in educational technology research. Using NAEP data, Wenglinski (1998) employed two measures of technology use in a study on the effects of educational technology on student learning. The first measure focused specifically on use of technology for simulation and higher-order problem solving and found a positive relationship between use and achievement. The second measure employed a broader definition of technology use and found a negative relationship between use and achievement. Thus, depending how one measures use, the relationship between technology use and achievement seems to differ. Such differences may account for some of the complexity policy makers and educational leaders confront when interpreting educational technology research.

Computer Adaptive Surveying:

As a frequent measurement tool in educational research, the questionnaire or survey has been widely applied in studies of educational technology. Paper and pencil administrations of student and teacher surveys are common in the literature, however, an increasing number of researchers are utilizing Internet-based tools to collect their survey data (Bebell, 2007; MEPRI, 2003). These web-based surveys are particularly advantageous in settings where technology is easily accessible, which is increasingly the case in educational technology research settings. In addition, the resulting data from a computer-based survey can be accessed and analyzed nearly instantly, streamlining the entire data collection process. However, the constraints and limitations of paper-based surveys have been rarely improved upon in their evolution to computer-based administration. The current authors believe that paper-based surveys are unnecessarily limited by their mode of administration and improvements in the delivery and measurement capabilities of surveys can be developed in the age of digital data collection. In an effort to more accurately measure specific facets of students' use, researchers at the Technology and Assessment Study Collaborative have recently developed two advancements in research instrumentation, the computer adaptive survey and the Flash sliding scale.

The computer adaptive survey (CAS) represents a new development in the advancement of the state-of-the-art in research instrumentation. In contrast to the current web-based surveys used to evaluate technology use, which present all students with a limited set of items in a linear manner, the CAS adapts the presentation of items based upon individual student's responses to each item. Computer adaptive surveying builds upon the theory and design of adaptive achievement tests and computer adaptive testing which can more efficiently provide an accurate estimate of an examinee's ability than a comparable paper-based test (Wainer, 1990). Similarly, a computer adaptive survey can tailor the survey items presented to a given student or teacher in order to probe the specific details of a general phenomenon.

Take, for example, an evaluation of how middle school students use computers in a multi-school 1:1 laptop program. Since students across multiple settings may potentially use laptop computers in very

different ways, an adaptive survey approach will enable the research team to probe the specific ways in which students use technology without requiring them to respond to sets of items that are unrelated to the ways in which they use computers. At the present time, the authors are employing a limited form of CAS to measure the variety and frequency of student and teachers technology across seven Massachusetts middle schools in the evaluation of a 1:1 student laptop program (Bebell, 2007). Through an adaptive survey, a more complete and accurate descriptive understanding of a given phenomenon will be acquired. Moreover, due to the adaptive nature of the survey, students and teachers will not be presented with sets of items that are unrelated to the ways in which they use laptops, thus decreasing time required to complete the survey, decreasing fatigue, and increasing the accuracy of information provided for items that are related to their use of laptops. A summary of the development and validation process for a computer adaptive survey example is presented in Appendix A.

Flash Slider:

The Flash sliding scale represents another recent advancement in the development of technology that uses continuous scaling choices to more accurately measures educational phenomenon in a web-based survey (Russell, Auhor, & Hoffman, in review). Traditionally, most paper and pencil surveys employ fixed, close-ended response options for the respondent to select their responses from. For example, when measuring the frequency of teachers' use of technology, teachers may be asked to select from a discrete number of responses presented in a survey:

During the last school year, how often did you use a computer to deliver instruction to your class?

- Never*
- Once or twice a year*
- Several times a year*
- Once a month*
- Several times a month*
- Once a week*
- Several times a week*
- Everyday*

In this example responding teachers simply select the response option that best describes the frequency of using a computer to deliver instruction. To enable the statistical analyses of the results, the researcher must assign a numeric value to each of the potential response options. Using the current example, the number assigned to each response option would correspond linearly with increasingly frequent technology use:

Ex. 1: Assigning Linear Values

Response Option	Assigned Value
<i>Never</i>	0
<i>Once or twice a year</i>	1
<i>Several times a year</i>	2
<i>Once a month</i>	3
<i>Several times a month</i>	4
<i>Once a week</i>	5
<i>Several times a week</i>	6
<i>Everyday</i>	7

In Example 1, an eight point scale (0-7) differentiates how frequently each teacher uses technology for instruction over the course of a given year. These linear values (0-7) are assigned to each teacher response creating an eight-point scale of technology use. Once quantified, the frequency of teachers' use of technology can be represented using an eight-step scale ranging from *never* (=0) to *everyday* (=7). By quantifying the responses numerically a wide variety of arithmetic and statistical analyses may be performed.

In measurement theory, the greater number of response options provides greater mathematical differentiation of a given phenomenon, in this case the frequency of technology use. However, a large number of response options may lead to difficulty for respondents forced to select their response from tedious and lengthy lists. In addition, computer and paper-based surveys impart their own inherent space and layout limitations. Conversely, using fewer response options provides less differentiation of a phenomenon and less information to the research team. Compromising these concerns, five to seven point scales are most typically applied in survey research where the detail of measurement is balanced with the ease of administration (Dillman, 2000; Nunnally, 1978).

However, the careful interpreter of research will recognize limitations to this widely employed approach. Using the current example, the response options are assigned using linear 1-step values, while the original response options actually describe non-linear frequencies. Thus, the linear 1-step values presented in Example 1 result in an ordinal measurement scale, "where values do not indicate absolute qualities, nor do they indicate the intervals between the numbers are equal" (Kerlinger, 1986, p. 400). From a measurement point of view, the values assigned in the preceding example are actually arbitrary (with the exception of zero which indicates that a teacher *never* uses technology). Although this type of scale serves to differentiate degrees of teachers' technology use, the values used to describe the frequency of use are unrelated to the original scale easily leading misinterpretation. Consider that this survey question was administered to a sample of middle school teachers at the beginning and again near the end of the school year. The average value calculated across all teachers during the first administration was 2.5 indicating that, on average, teachers used technology for instruction between *several times a year* and *once a month*. The average value calculated across the teachers during the second administration was 5.1, or about *once per week*. Arithmetically, it appears that technology use has actually doubled in the interval between the survey administrations, but in relation to the original survey scale, teachers use of technology had actually increased substantially more than that.

In example 2, the assigned values for each response option are designed to reflect the actual frequency of how often teachers could deliver instruction over the course of a 180 day school year:

Ex. 2: Assigning "Real" Values

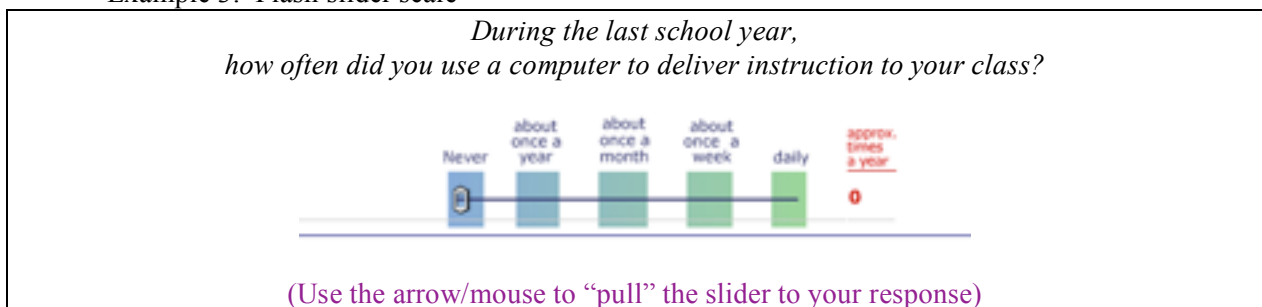
Response Option	Assigned Value
<i>Never</i>	0
<i>Once or twice a year</i>	2
<i>Several times a year</i>	6
<i>Once a month</i>	9
<i>Several times a month</i>	27
<i>Once a week</i>	36
<i>Several times a week</i>	108
<i>Everyday</i>	180

In Example 2, the same survey question and response options are presented, however the researcher assigns values to each response choice that represent “real” values. Assuming the school year equals 180 days, or 9 months, or 36 weeks, the analyst assigns values to each response option that mirror the estimated frequency of use. This approach results in a 180 point scale where 0 represents a teacher *never* using technology and 180 represents *everyday* use of technology. This approach provides easier interpretation and presentation of summary data because the difference between the numbers actually reflects an equal difference in the amount of attribute measured (Glass & Hopkins, 1996).

In the current example, the resulting survey data takes on qualities of an interval measurement scale whereby “equal differences in the numbers correspond to equal differences in the amounts the attributes measure” (Ibid, p. 8). In other words, rather than the 8 step scale presented in Example 1, the 181 step scale resulting from Example 2 offers a clearer and more tangible interpretation of teachers’ technology use. The number of times a teacher may use technology may occur at any interval on a continuous scale between 0 and 180, however, in the current example teachers responding to the item were still provided with only 8 discrete response options in the original survey question. The small number of response options typically employed in survey research forces survey respondents to choose a response option answer that best approximates their situation. For example, a teacher may use technology somewhat more than *once a week* but not quite *several times per week*. Faced with inadequate response options, the teacher must choose between the two options. In this scenario, the survey respondent is forced to choose one of the two available options both yielding imprecise, and ultimately, inaccurate data. If the teacher selects both options, the analyst typically must discard the data or be forced to subjectively assign a value to the response.

Recognizing the measurement limitations of limited response options, the current authors have developed a survey presentation method where survey items are presented with continuous true-to-life scales allowing the survey user to select exactly the best response for a given question rather than relying on a limited number of fixed, closed-ended response options (Russell, Author, & Hoffman, in review; Bebell & Rusell, 2006). Through the use of Macromedia Flash slider scale, survey respondents may select from a limitless series of response options in a computer-based survey environment. This advancement in technology allows the same survey item to be measured using a true ratio scale where the entire range of potential use (with every available increment present) is presented to teachers. In Example 3, teachers are presented the technology use survey question with the Flash sliding scale:

Example 3: Flash slider scale



Each respondent uses their mouse to select their response on the sliding scale. Although the interactive nature of the Flash sliding scale is challenging to demonstrate on paper, the program is designed to help respondents quickly and accurately place themselves on the scale. In the current example, the teachers’ response is displayed for them (in red) under the heading “approximate number of times per year”. So, as a survey respondent pulls the sliding scale across the response options on the horizontal line, the “approximate number of times per year” field displays their response in real time. Thus, a teacher can

move the slider to any response option between 0 (never) and 180 (daily). In addition, the descriptions above the horizontal slider provide some parameters for teachers in quickly selecting the appropriate response. Solving many of the limitations of traditional categorical survey response options, the Flash sliding scale, although still being pilot tested, offers improved measurement capabilities. Russell, Author, and Hoffman (in review), have recently investigated the measurement properties of six different versions of Flash sliding scales. In addition, Flash sliding scales have been recently used to measure teachers' use of technology across five schools participating in 1:1 computer initiatives (Bebell & Russell, 2005).

Lack of Prior Achievement Measures or Comparison Groups:

In addition to focusing on the improvement of how technology use is conceived of and measured, a great number of educational technology studies also suffer from weak research design. As new types of educational technology applications continue to be developed, it is typically the goal for the corresponding research to demonstrate how teaching and learning may be affected. When examining the impact of new educational technology initiatives on student learning, it is critical that measures of students' prior achievement are collected or data are similarly obtained from students in comparable settings. The importance of using measures of student prior achievement is especially relevant when comparisons are made among students who are not randomly assigned to experimental or control groups. Without measures of prior achievement, policy makers cannot develop a clear understanding of whether differences in student achievement are a result of participation in a technology program or simply an artifact of selection bias due to non-random assignment.

Paper Versus Computer Based Assessments:

The results of a series of experimental studies by Russell and colleagues strongly suggest that many states' paper-based standardized achievement tests may be underestimating the performance of students who are accustomed to working with technology simply because they do not allow students to use these technologies when being tested (Russell, 1999; Russell & Haney, 1997; Russell & Plati, 2001). Through a series of randomized experiments, Russell and his colleagues provide empirical evidence that students who are accustomed to writing with computers in the classroom perform between 0.4 and 1.1 standard deviations higher when they are allowed to use a computer when performing tests that require them to compose written responses (Russell, 1999; Russell & Haney, 1997; Russell & Plati, 2001). These results suggest that researchers studying the effects of educational technology are particularly at risk for underestimating the ability of technology savvy students when they rely on paper-based assessment instruments as their outcome measures.

Becker (as cited in Kirkpatrick & Cuban, 1998) also discusses the struggle faced by educational technology researchers in interpreting achievement tests data stating, "When standardized test scores rise, it's difficult to discern whether the rise was due to the students' work with computers or another influence" (p. 6).³ Together, the shortcomings of using standardized tests confound efforts to examine the direct effects of technology use on student learning. Thus, when using standardized tests, or any measure of student achievement, it is important that the constructs measured by the instrument are aligned with the constructs developed through students' uses of technology and that the form of the test (e.g., paper- or computer-based) does not add construct-irrelevant variance. Given this evidence concerning the underestimation of high-tech students on low-tech tests (i.e., paper tests), the authors currently utilize in

³ Pedagogical beliefs and practices have been shown to interact with technology use to affect student learning. Thus, it is also important to consider both pedagogical practices and specific technology use when examining the relationship between technology use and learning outcomes. For this reason we advocate measuring teachers' pedagogical beliefs and practices with previously validated sets of survey items (Becker, 1999; Russell, O'Dwyer, Bebell, & Miranda, 2003).

their own research a self designed web-based writing assessment whereby students voluntarily choose to complete the assessment using their laptops or paper and pencil. (Bebell & Russell, 2006).

Alignment of Measurement Tools:

A similar challenge that must be addressed to adequately determine the impacts and effects of educational technology on student learning relates to defining and aligning educational outcomes. By federal law, all states now administer grade level tests to students in grades 3-8. In addition, states administer tests to some high school grades and/or end of course tests for high school students. For many observers of educational technology programs, the state test results provide easily accessible educational outcomes. However, given the intent of state tests to sample the state standards, these tests do not necessarily provide valid measures of the types of learning that may occur when computers are used by students and/or their teachers.

When examining the impact of 1:1 technology use on student learning, it is critical that the outcome measures actually assess the types of learning that may occur as a result of technology use and that those measures are sensitive enough to detect potential changes in learning that may occur. Since most standardized tests attempt to measure a domain broadly, standardized test scores often do not provide measures that are aligned with the learning that may occur when technology is used to develop specific skills or knowledge. As an example, computers may be used extensively in mathematics classes to develop students' understanding of graphing and spatial relationships, but not for other concepts. While the state mathematics test may contain some items related to graphing and spatial relationships, it is likely that these two concepts represent only a small portion of the assessment and are tested with a limited number of items. As a result, the total state test score is unlikely to be sensitive to any effects of computer use on these two concepts. While analyses could focus on the subset of items that focus on these concepts, the small number of items is likely to be insufficient to provide a reliable estimate of student achievement in these areas. Rather than employing state test results, an alternate strategy is to develop customized tests that contain a larger number of items specific to these concepts. Although it can be difficult to convince teachers and/or schools to administer an additional test, aligned assessments will result in more reliable scores and provide increased validity for inferences about the impacts of computer use on these concepts.

Hierarchical Structure of Schools:

Another challenge to evaluating the effects of educational technology programs on teaching and learning is the inherent hierarchical structure of schools that is rarely taken into account by researchers, evaluators and school leaders as they evaluate the impact of technology programs. As a consequence, many studies of educational technology fail to properly account for the organizational processes that mediate the relationship between technology use and achievement. Over the past two decades, researchers have become increasingly aware of the problems associated with examining educational data using traditional analyses such as ordinary least squares analysis or analysis of variance. Since educational systems are typically organized in a hierarchical fashion, with students nested in classrooms, classrooms nested in schools, and schools nested within districts, a hierarchical or multilevel approach to data analysis is often required (Robinson, 1950; Cronbach, 1976; Haney, 1980; Burstein, 1980; Raudenbush & Bryk, 1992; Kreft & de Leeuw, 1998).

A hierarchical approach is recommended because education systems are typically organized in a hierarchical fashion; students are nested within classrooms, classrooms within schools, and schools within districts. At each level in an educational system's hierarchy, events take place and decisions are made that potentially impede or assist the events that occur at the next level. For example, decisions made at the

district level may have profound effects on the technology resources available for teaching and learning in the classroom. A hierarchical approach is well suited to analyzing the relationship between technology use and student achievement and requires the analysis of individuals within classrooms and where possible, classrooms within schools. This approach has three distinct advantages over traditional analyses: (1) the approach allows for the examination of the relationship between technology use and achievement to vary as a function of classroom, teacher, school, and district characteristics; (2) the approach allows the relationship between technology use and achievement to vary across schools; and (3) differences among students in a classroom and differences among teachers can be explored at the same time therefore producing a more accurate representation of the ways in which technology use may be impacting student learning (Raudenbush & Bryk, 1992; Goldstein, 1995; Kreft & de Leeuw, 1998).

To date, very little published research in educational technology has applied a hierarchical approach. However, in 2005, O'Dwyer and colleagues published results of a 2-level hierarchical model was applied to study of the effects of students technology use on student achievement across 55 intact fourth grade classrooms where both teacher and student data was collected (O'Dwyer, Russell, Bebell, Tucker-Seeley; 2005).

Discussion/Conclusions:

The current paper explores common methodological limitations in the field of educational technology research. To summarize, the individual limitations presented pose significant challenges to the current status of educational technology research. Collectively, these limitations suggest that the current status of educational technology research is plagued by a number of methodological shortcomings that has stymied the potential growth and development of educational technology programs. This lack of methodological precision and validity is particularly concerning in light of the considerable federal, state and local investments in school-based technologies. Many of these limitations are likely responsible for the major shortage of high quality empirical research studies addressing the impacts of technology in schools. Thus, education decision-makers contemplating the merits of educational technology are forced to make decisions about the expenditure of millions of dollars with only weak and limited evidence on the effects of such expenditures on instructional practices and student learning.

With the rising interest in expanding educational technology access, particularly 1:1 laptop initiatives, the psychometric and methodological weaknesses inherent in the current generation of research results in studies that (1) fail to capture the nuanced ways in which laptops are being used in schools, and (2) fail to align learning outcome measures with the measures of student learning. Beyond documenting that use of technology increases when laptops are provided at a 1:1 ratio, the current research tools used to study such programs fail to provide the important detailed information about the extent to which technology is used across the curriculum and how these uses may be impacting student learning.

Although the current paper outlined a number of common methodological weaknesses in educational technology research, the current lack of high-quality research is undoubtedly a reflection of a general lack of support provided to research and evaluate technology in schools. Producing high-quality research is an expensive and time-consuming undertaking that is often beyond the resources of most schools and individual school districts. At the state and federal level, vast amounts of funding are expended annually on educational technology and related professional development, yet few if any funds are earmarked to research the effects of these massive investments. For example, the state of Maine used a \$37.2 million dollar budget surplus to provide all 7th and 8th grade students and teachers with laptop computers. Despite the fact that Maine was the first state to ever implement such an innovative and far-reaching program, approximately \$200, 000, or about one half of one percent (0.005%) of the overall budget, was budgeted for research and evaluation. Similarly, in Henrico County, VA, a similar initiative invested \$24.2 million

dollars to provide laptop computers to over 23,000 students and teachers. During the first four years of the program's implementation, no significant research or evaluation efforts were undertaken or funded. Similarly, technology companies such as Apple, Dell, and Gateway invest little, if anything, on measuring the impacts and effect of technology in education despite the obvious benefits that positive results would generate for their products and services.

Recognizing that collecting research in educational settings will always involve compromises and limitations imparted by scarce resources, the authors propose that many opportunities exist to improve data collection and analysis within the structure of existing methodological designs. Just as technology has transformed the efficiency of commerce and communication, we feel that technology may provide many opportunities to advance the art and science of educational research and measurement. Growing frustrated with the status quo, we have begun to address these methodological weaknesses and employ our improved alternatives in our own research and evaluation efforts as well as offer them to the larger research community. The authors feel these small methodological improvements represent only a sampling of the potential advancements for data collection in the digital age.

References

- Baker, E. L., & Herman, J. L. (2000). *New models of technology sensitive evaluation: Giving up old program evaluation ideas*. SRI International: Menlo Park, CA. Retrieved January 10, 2003 from, <http://www.sri.com/policy/designkt/found.html>
- Bebell, D. (2007). 1 to 1 Computing: Year One Results from the Berkshire Wireless Learning Initiative Evaluation. Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, IL.
- Bebell, D. & Russell, M. (2006). Revised Evaluation Plan for Berkshire Wireless learning Initiative. Boston, MA: Boston College, Technology and Assessment Study Collaborative. page 16 FIX
- Bebell, D., Russell, M., & O'Dwyer, L.M. (2004). Measuring teachers' technology uses: Why multiple measures are more revealing. *Journal of Research on Technology in Education*, 37(1), 45-63
- Becker, H. (1994). *Analysis and trends of school use of new information technologies*. Washington, DC: Office of Technology Assessment.
- Becker, H. (1999). *Internet use by teachers: Conditions of professional use and teacher-directed student use*. Irvine, CA: Center for Research on Information Technology and Organizations.
- Burstein, L. (1980). The analysis of multi-level data in educational research and evaluation. In D.C. Berliner (Ed.), *Review of research in education* (Vol.8, pp. 158- 233). Washington, DC: American Educational Research Association.
- Cronbach, L.J. (1976). Research on classrooms and schools: Formulation of questions, design, and analysis. (Occasional paper). Stanford, CA: Stanford Evaluation Consortium, Stanford University.
- Dillman, 2000
- Glass, G. V and Hopkins (1996)
- Goldberg, A., Russell, M., & Cook, A. (2003). The effect of computers on student writing: A meta-analysis of studies from 1992 to 2002. *Journal of Technology, Learning, and Assessment*, 2(1).
- Goldstein, H. (1995). Multilevel statistical models. London: Edward Arnold.
- Haney, W. (1980). Units and levels of analysis in large-scale evaluation. New Directions for Methodology of Social and Behavioral Sciences, 6, 1-15.
- Kerlinger, 1986
- Kirkpatrick, H. & Cuban, L. (1998). Computers make kids smarter—right? [Electronic version]. Technos Quarterly, v7, No. 2. Retrieved November 12, 2003, from http://www.technos.net/tq_07/2cuban.htm
- Kreft, I., & de Leeuw, J. (1998). Introducing multilevel modeling. Thousand Oaks, CA: SAGE.

- Lerman, J. (1998). *You've got mail: 10 nifty ways teachers can use e-mail to extend kids' learning*. Retrieved January 10, 2003 from <http://www.electronic-school.com/0398f5.html>.
- Mathews, J. (1996, October). *Predicting teacher perceived technology use: needs assessment model for small rural schools*. Paper presented at the Annual Meeting of the National Rural Education Association, San Antonio, TX.
- McNabb, M., Hawkes, M., & Rouk, U. (1999). *Critical issues in evaluating the effectiveness of technology*. Proceedings of the Secretary's Conference on Educational Technology: Evaluating the Effectiveness of Technology. Retrieved January 10, 2003 from, <http://www.ed.gov/Technology/TechConf/1999/confsum.html>.
- Nunally, J. C. (1978). *Psychometric theory*. New York, NY: McGraw-Hill Book Company.
- O'Dwyer, L. M., Russell, M., Bebell, D., & Tucker-Seeley, K. R. (2005). Examining the relationship between home and school computer use and students' English/language arts test scores. *Journal of Technology, Learning, and Assessment*, 3(3). Available from <http://www.jtla.org>.
- Office of Technology Assessment. (1995). *Teachers and technology: Making the connection*, OTA-EHR-616. Washington, DC: U.S. Government Printing Office.
- Office of Technology Assessment. (1989). *Linking and Learning: A new course for education*. Washington, DC: U.S. Government Printing Office.
- Office of Technology Assessment. (1988). *Power On! New tools for teaching and learning*. Washington, DC: U.S. Government Printing Office.
- Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods*. Sage Publications.
- Ravitz, J.; Wong, Y., & Becker, H. (1999). *Teacher and teacher directed student use of computers and software*. Irvine, CA: Center for research on information technology and organizations.
- Robinson, W.S. (1950). Ecological correlations and the behavior of individuals. *American Sociological Review*, 15, 351-357.
- Russell, M. & Haney, W. (1997). Testing writing on computers: An experiment comparing student performance on tests conducted via computer and via paper-and-pencil. *Educational Policy Analysis Archives*. Vol. 5 No. 3
- Russell, M. (1999). Testing on Computers: A Follow-up Study Comparing Performance on Computer and On Paper. *Education Policy Analysis Archives*. Vol. 7 No. 20.
- Russell, M. & Plati, T. (2001). Effects of Computer Versus Paper Administration of a State-Mandated Writing Assessment. *Teachers College Record*.
- Russell, 2002
- Russell, M., O'Dwyer, L., Bebell, D., & Miranda, H. (2003) Technical report for the USEIT study. Boston, MA: Boston College, Technology and Assessment Study Collaborative.

Russell, M., Bebell, D., & Higgins, J. (2004). Laptop learning: A comparison of teaching and learning in upper elementary classrooms equipped with shared carts of laptops and permanent 1:1 laptops. *Journal of Educational Computing Research*. Vol. 30 No. 3.

Russell, M. Auhtor2, & Hoffman, T. (in press). Slider Analysis Paper.

Salomon, G, Perkins, D., & Globerson, T. (1991). Partners in cognition: extending human intelligence with intelligent technologies. *Educational Researcher*, 20, 2-9.

Wainer, H. (1990). *Computerized Adaptive Testing: A Primer*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Waxman, H. C., Lin, M., & Michko, G. M., (2003). A meta-analysis of the effectiveness of teaching and learning with technology on student outcomes. Naperville, IL: Learning Point Associates. Retrieved April 22, 2004, from <http://www.ncrel.org/tech/effects2/>

Wenglinsky, H. (1998). Does it compute? The relationship between educational technology and student achievement in mathematics. Princeton, NJ: Educational Testing Service Policy Information Center.

Appendix A: Summary example of developing and assembling computer adaptive survey for measuring technology use

Assembling the technology use item sets into a comprehensive computer adaptive survey will mirror the steps of computer adaptive test development (Wainer, 1990). Specifically, previously developed survey item sets can be incorporated into a logic tree that reflects the complete domain of technology use in science and mathematics. The logic tree begins with survey items that measure general technology use (e.g., how often do you use laptops during mathematics class, how often do you use laptops outside of class to work on mathematics problems, how often do you use laptops in science class). The next layer of items will ask more specific questions about technology use for each subject area (e.g., how often do you use laptops to solve problems in mathematics, how often do you use laptops to write during mathematics class, how often do you use the Internet during mathematics class, how often do you use spreadsheets or graphing programs during mathematics class). The third layer of survey items will ask even more specific questions about use within each of the preceding areas (e.g., how often do you use the Internet during class to access data sets, to access information about mathematicians, to access information about mathematical theorems or concepts, to access applets). Based on the information garnered during the item development process, additional layers may be required to fully quantify and measure the extent of technology use.

Once developed, the adaptive technology use surveys would need to be piloted in a sample of comparable classrooms to the studies sample. The purpose of this piloting is to focus specifically on the functionality and usability of the instrument. Additionally, the pilot test of the adaptive surveys allows for the examination of whether students and teachers are administered those survey items that were designed to capture the specific technology practices as expected through the item development process.

Once the adaptive surveys have been piloted, a formal validity study should be conducted using the same sample of classrooms as before. Additionally, another sample of classrooms should be recruited from comparable settings. Specifically, the validity study will examine the (1) psychometric qualities of the survey instrument, (2) students' interpretation of the survey items, as well as (3) a comparison of the survey results and practices as documented through other forms of inquiry including student interviews, focus groups of students, and classroom observations.

Within all participating classrooms the adaptive surveys would then be administered. Shortly following the completion of the surveys, a sample of survey participants will be selected to participate in focus groups. Student focus groups will be used to collect detailed information about the variety of ways in which laptops are used inside and outside the classroom for mathematics and science learning. Data from the adaptive surveys will be aggregated to the classroom level and then compared to data collected through the teacher interviews and student focus groups.

Traditional psychometric analyses of scale validity and reliability will also be performed. These analyses include principal component analysis of survey items believed to measure closely-related technology uses as well as analyses of internal consistency and item to scale correlations. In addition, classroom survey data will be used to examine the extent to which reported use varies among students within classrooms. If the survey functions well, one would expect within-classroom variation to be relatively small compared to between-classroom variation.

Finally, to examine the extent to which students' interpretations of item prompts are consistent with the intended interpretation, a talk-aloud study will be conducted with a sub-sample of 20 students.

Specifically, as students included in this component of the validity study perform the adaptive survey, we would expect them to explain their reasoning for each item response as they complete the survey. The students stated reasoning will be compared to the intended interpretation of the given survey item. In cases where student reasoning is not consistent with the intended interpretation, the item will either be rewritten or removed.

Based on the results of the initial validity work, the computer adaptive technology use survey may be further revised as necessary. In the event that revisions result in significant changes, validation should be repeated with a new sample of 10-15 classrooms.