

RUNNING HEAD: student survey validity

Do college student surveys have any validity?

Paper presented at the 2009 meeting of the Association for the Study of Higher Education

AUTHOR'S NOTE: If you come across any study that is pertinent to this paper, please let me know. I am citing research across several fields, and it is easy to miss relevant studies. Please do not cite without my permission, as this is a work in progress.

Version 1.2

Stephen R. Porter
Associate Professor of Research and Evaluation
Department of Educational Leadership and Policy Studies
N243 Lagomarcino Hall
Iowa State University
Ames, IA 50011
Phone: (515) 294-7635
Email: srporter@iastate.edu

Abstract

Using standards established for validation research, I review the theory and evidence underlying the validity argument of the National Survey of Student Engagement (NSSE) . I use the NSSE because it is the preeminent survey of college students, arguing that if it lacks validity, then so do almost all other college student surveys. I find that it fails to meet basic standards for validity and reliability, and recommend that higher education researchers initiate a new research agenda to develop valid college student surveys.

Do college student surveys have any validity?

Within the field of higher education, the majority of quantitative research focuses on college students. Given the limitations of institutional databases, surveys of college students have become one of our largest and most frequently used data sources. In addition, surveys of college students play an increasingly important role in evaluating college and university programs and policies. As such, having valid and reliable data about students is vital for both practitioners and scholars. Yet if our survey questions are not measuring what we think they are, then our knowledge of college students will be flawed.

In this paper, I argue that the typical college student survey question has minimal validity, and that our field requires an ambitious research program to reestablish the foundation of quantitative research on students. Our surveys lack validity because a) they assume that college students can easily report information about their behaviors and attitudes, when the standard model of human cognition and survey response clearly suggests they cannot, b) existing research using college students suggests they have problems correctly answering even simple questions about factual information, and c) much of the evidence that higher education scholars cite as evidence of validity and reliability actually demonstrates the opposite.

I choose the National Survey of Student Engagement (NSSE) for my critical examination of college student survey validity for several reasons. First, it is one of the most prominent surveys of student behavior and attitudes, and is widely used by researchers studying students, as well as institutions interested in assessment. Second, the NSSE survey serves as a model for surveys designed by other researchers and institutional assessment staff, precisely because of its prominence. Given its wide use by both practitioners and scholars, it is vital that we understand whether the NSSE can be considered a valid instrument. Finally, unlike many other college

student surveys, NSSE staff and researchers using the NSSE (henceforth collectively referred to as NSSE researchers) have gone to significant efforts to validate the survey through a variety of studies. Indeed, one could argue they have marshaled the best evidence to date as to the validity of common college student survey questions. So, if the NSSE cannot withstand scrutiny, it is likely that many, if not most, other college student surveys (such as those produced by the Higher Education Research Institute at UCLA) cannot either. And if the preeminent survey of college students lacks validity, then this calls into question much of what we think we know about college students.

What do we mean by ‘validity’?

Within the field of education, the definition of validity has changed greatly during the past century. Today, it encompasses a far broader meaning than most higher education researchers realize.

Originally, validity referred to criterion-related validity, in which a measure was compared against some external criterion (see Kane (2001) and Lissitz and Samuelsen (2007) for brief historical overviews of validity from the early 20th century to today). Early scholars also began to use content validity in their research, or the extent to which a measure or test¹ encompasses a specified content area. Later, and in part due to the obvious problem that many measures (such as attitudes) lack an external criterion with which to compare them, researchers turned to construct validity. Measures were now deemed valid depending on how they related to other constructs, with such relationships between constructs derived from theory postulating the existence and direction of these relationships.

¹ Much of the debate over validity uses the term “test”; here, I use the term measure, as more relevant to a discussion of survey validity.

In the 1980s, scholars began to develop a unified theory of validity, in which these different forms of validity were subsumed under a broader notion of validity (Messick, 1989). Although still debated, this unified notion of validity enjoys wide acceptance in the field of education. In part, this is due to discomfort over having several approaches to validity; with such an approach, researchers can pick and choose which approach makes their measure appear most valid. It is also due to the idea that measures cannot really be evaluated without some discussion as to the use to which they will be put. In other words, a strong correlation between a measure and a criterion may or may not be considered evidence of validity; it all depends on how the measure will be used.

I use Kane's (1992; 2001) argument-based approach to validity, for two reasons. First, his approach is typical of the unified theory of validity school of thought, and is firmly grounded in the current *Standards for Educational and Psychological Testing* (henceforth, the *Standards*), a manual jointly issued by the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education. Given the wide acceptance of these standards in the fields of education and psychology, any discussion of college student survey validity must at a minimum address these standards. Second, NSSE researchers have implicitly adopted his approach when defending the psychometric properties of the NSSE.

The *Standards* define validity as

... the degree to which evidence and theory support the interpretation of test² scores entailed by proposed uses of tests. Validity is, therefore, the most fundamental consideration in developing and evaluating tests. The process of

² The *Standards* use the word "test" as an overall term that clearly encompasses student surveys as used in most higher education contexts; see p. 3 of the *Standards*.

validation involves accumulating evidence to provide a sound scientific basis for the proposed score interpretations (p. 9).

As Kane (2001) notes, Cronbach suggested that researchers think of constructing a validity argument, rather than conduct validity research. We can think of validity as an argument, based on theory and evidence, rather than a simple correlation. Theory can range from descriptions of how measures and constructs should be related, theories about cognitive processes of those filling out surveys, as well as a variety of evidence, such as expert reviews of content and quantitative descriptions of how measures relate to other constructs on the instrument, as well as external data. Validity, then, is established by the combined strength of theory and evidence that is brought to bear to support a given interpretation of a measure, given the context in which it is used.

An examination of research and documents authored by NSSE researchers reveals the following argument as to the validity of the NSSE:

Argument One – Background. The NSSE “is specifically designed to assess the extent to which students are engaged in empirically derived good educational practices and what they gain from their college experience” (Kuh, 2004, p. 2). Survey data can be used by schools to improve the quality of the undergraduate experience, by states to measure institutional performance, and by the public as measures of college quality (Kuh et al., 2001).

In addition, items from the survey can be combined to create scales that are valid measures of student behavior (Kuh, 2004; Kuh et al., 2001), and individual items from the survey can also be used as valid measures of student behavior. Although the latter is not explicitly stated, it is clearly an implicit assumption underlying recommendations to institutions that they analyze individual items to better understand engagement at their institution (National

Survey of Student Engagement, 2007), numerous institutional reports prepared for NSSE users that show how their institution differs from various comparison groups in terms of effect size and statistical significance on an item-by-item basis (e.g., National Survey of Student Engagement, 2009a, 2009b), and the item analysis in the validation study by Carini, Kuh and Klein (2006).

Argument Two – Content. Questions in the survey consist of “items directly related to institutional contributions to student engagement, important college outcomes, and institutional quality. The NSSE design team selected items according to three general criteria: 1) is the item arguably related to student outcomes as shown by research? 2) Is the item useful to prospective students in choosing a college? 3) Is the item straightforward enough for its results to be readily interpreted by a lay audience with a minimum of analysis?” (Kuh et al., 2001, p. 3).

Argument Three – Response process. NSSE survey questions can be understood and accurately answered by college students (Kuh, 2004). More specifically, NSSE researchers assert that student self-reports of behavior can be considered accurate when the information is known to respondents, the questions are phrased clearly and unambiguously, and the questions refer to recent activities.

Argument Four – Internal structure. Items on the NSSE correlate with one another in such a way that items can be grouped into five constructs (level of academic challenge, active and collaborative learning, student-faculty interaction, enriching educational experiences, and supportive campus environment). These five constructs are conceptually distinct and the empirical evidence is strong enough that they can be referred to as “national benchmarks of effective educational practice” (Kuh, 2004; Kuh et al., 2001). In addition, individual items within the NSSE correlate with one another as theory suggests. For example, the number of hours spent studying is correlated with the number of assigned course readings at .25 (Kuh, 2004);

presumably, more assigned readings indicates more demanding coursework, which in turn necessitates more hours spent studying.

Argument Five – Relations to other variables. NSSE items and scales correlate with other data as predicted. For example, the NSSE psychometric properties report (Kuh, 2004) cites research that shows self-reported learning gains are correlated with achievement tests (Pike, 1995); research also shows that students in the natural sciences report larger gains in quantitative thinking than other students (Pace, 1985). NSSE researchers claim that the NSSE scales and items are correlated with external measures of student learning, such as tests of cognition and GRE writing scores (Carini et al., 2006). Most importantly, the NSSE “represents student behaviors that are *highly correlated* [emphasis added] with many desirable learning and personal development outcomes of college” (Kuh, 2004, p. 2).

Henceforth, I refer to these as Arguments One through Five. In this paper, I take Argument One as given, and evaluate the theory and evidence for the remaining arguments.

Taken together, these five points indicate that NSSE researchers have clearly articulated an argument for the validity of the NSSE. As the *Standards* note, “a sound validity argument integrates various strands of evidence into a coherent account of the degree to which existing evidence and theory support the intended interpretation of test scores for specific uses. It encompasses evidence gathered from new studies and evidence available from earlier reported research” (p. 17). NSSE researchers have clearly accumulated theory and multiple types of validation evidence to suggest the NSSE has validity, in terms of assessing the extent to which students engage in certain practices, and measuring institutional quality.

In the remainder of the paper I take a similar approach, taking particular heed of the *Standards*' suggestion that one should develop rival hypotheses that may provide an alternative

explanation of specific validity evidence. I argue that NSSE researchers have misunderstood human cognition and survey response, and that college students can rarely report accurate information about their behaviors; that recent evidence indicates the NSSE educational benchmarks are different than asserted; and that a critical examination of how NSSE items and scales relate to external data indicates that these relationships are almost nonexistent. In sum, I assert that multiple strands of theory and evidence demonstrate that the NSSE has very limited validity for its intended purposes, and that researchers and institutions must adopt a new approach to surveying college students. I conclude the paper by outlining a research agenda for our field, so that we may begin to develop college student surveys that meet basic validity and reliability standards.

What do correlations tell us about validity?

The use of correlations has a long history in validation research. For this paper, three points are important. First, the *Standards* (Standard 1.15) have explicitly rejected the simple use of correlations in validation research in favor of more complex data analyses, both analytical (such as multiple regression) and descriptive.

Second, and related to the first point in terms of understanding measures and their validity, correlations can obscure as much as they illuminate. Consider asking students their height and weight with a survey, and then measuring the students to obtain their actual height and weight. We would find a strong correlation between the two (.90 to .93), and using the typical approach in higher education, conclude that self-reports of height and weight are valid measures. Yet a closer examination of the data would reveal that students over-report height (by 2.7 inches) and under-report weight (by 3.5 lbs), and that these errors are correlated with gender: females tend to report weighing less (Brener, McManus, Galuska, Lowry, & Wechsler, 2003). I

refer to this as the ‘correlation fallacy’ in validation research: relatively high correlations can mask large differences between measures. As survey methodologists have long argued, any quantitative description of associations in a validation argument should also consider issues of bias as well as association (Groves et al., 2004). Achen (1977; 1978) has an excellent discussion of the problems with using correlations to compare two supposedly similar measures.

Third, we must have some sort of criteria for declaring a correlation or effect size to be substantively significant. Cohen’s rule of thumb for effect sizes is now recognized as too rigid and for the most part useful only for power analyses (McCartney & Rosenthal, 2000); researchers instead rely on the research context to determine the substantive relevance of an effect size. As Kozak (2009) notes, whether a correlation should be considered weak or strong depends crucially on whether one *theoretically* expects no relationship or a strong relationship; if a strong relationship is posited, even correlations close to .90 could be considered weak, while with a non-existent relationship even modest correlations of .10 could be considered strong.

I believe there are two standards to be used for evaluating the empirical evidence in the NSSE validity argument. For areas where we would expect to see no differences, even an effect size of .10 should be considered troubling. For example, if we expect to find no differences in survey question understanding across student groups, because student understanding should not differ across groups due to clear and concise definitions in our survey questions, then almost any statistically significant association is problematic. Conversely, in areas where we expect to find strong associations, such as between NSSE scales and similar external measures, then correlations as large as .20 or .30 should be considered weak.

Sources of validity evidence for the NSSE

Evidence based on test content (Argument Two)

Is the NSSE a valid measure of student behavior and attitudes in terms of its content?

Looking at Argument Two, it is unclear what the content domain of the NSSE is, and why certain items are included. First, the domain is so widely defined that almost any student survey question could be included under the areas “engagement”, “student outcomes” and “institutional quality”.

Second, items are included because they are “arguably” (Argument Two) related to student outcomes as shown by research, yet I have been unable to locate such an explanation on an item-by-item basis. It is not clear, for example, what research underlies the inclusion of the frequency item “Used e-mail to communicate with an instructor”. Theoretically, it is not clear why schools with students using email on a frequent basis are of higher institutional quality, or would necessarily have better student outcomes, compared to schools where the majority of students do not use email to communicate with instructors. Similarly, items are included if students can use them in the college selection process, or if the results can be interpreted by a lay audience. Again, the standards here are not defined, nor is it made clear why these would be used as criteria for domain inclusion when trying to measure student behavior or institutional quality.

In general, the NSSE has an overly broad domain definition, and lacks an underlying theory specifying why *specific* items should be included in the instrument. This approach yields a content validity argument that is almost impossible to contradict: with no underlying justification for each item, and with such a broad content domain, the NSSE by definition must be valid, as almost anything can be included on the instrument. Moreover, validation research becomes difficult with such a broad domain, as virtually any student outcome, such as employment after college, can be used as evidence of validity, or a lack thereof (Gordon, Ludlum, & Hoey, 2008).

Evidence based on response processes (Argument Three)

An important part of any survey validity argument is a model of the response process, as well as empirical evidence supporting the proposed response process. In terms of surveys, we must be able to assert that respondents understand the survey question, that this understanding is constant across respondents (if it is not, then respondents are answering different questions), and that respondents can and will accurately answer the question. As described in Argument 3, NSSE researchers assert that their questions are phrased in precisely such a manner, and cite as evidence a handful of studies in the field of survey methodology. Notably, most of these citations are from the 1970s and 1980s (see e.g. Kuh, 2004).

This part of the NSSE validity argument stands in contrast to the current state of knowledge about human cognition and survey response, in terms of both theory and evidence. Most researchers in the field of survey methodology generally agree with Tourangeau et al.'s (2000) four-step theory of survey question understanding and response: comprehension, retrieval of information, judgment and/or estimation, and reporting of an answer.

When confronted with a question, a respondent must understand the individual terms and their combined meaning such that they can identify the information sought. For factual questions, they must have encoded memories about the topic and be able to retrieve these memories and attach dates to memories (if the question asks for reports for a given time period). Next, they must judge the quality of information from their memories (e.g., completeness and relevance), integrate the information from their memories to produce an answer, and if memories are incomplete, estimate a response. Finally, they must map their response onto the question's response scale, and do so accurately. As can be seen, the cognitive burden is substantial, and if the process breaks down at any stage, then validity becomes suspect. I examine the theory and

evidence for each of these four stages in terms how they apply to the NSSE in particular, and college student surveys in general.

Comprehension. One issue that college student survey researchers generally fail to consider is whether the typical college student can 1) understand the words and phrases we use, and 2) whether these understandings are similar across students. Consider the following questions from the 2009 NSSE (emphasis added):

How often have you:

Discussed grades or assignments with an *instructor*

Had *serious conversations* with students ...

To what extent does your institution emphasize each of the following?

Providing *the support you need to help you succeed academically*

To what extent has your experience at this institution contributed to your knowledge, skills, and personal development in the following areas?

Thinking critically and analytically

For the first question, it seems reasonable to conclude that there may be some confusion about who is an *instructor*. Does the term include just faculty, or does a graduate student listed as the instructor of record count? What about teaching assistants? Given that some research universities employ graduate students as instructors for the majority of their undergraduate classes, students at a research university and students at a liberal arts college may report the same answer, even though their level of faculty contact is very different.

For the second question, how does a student distinguish between *serious* and *frivolous* conversations? And what is a *conversation*? A chat in the bathroom? An hour-long bull session in a dorm room?

For the third question, what does the word *support* mean? It could refer to many different things, such as ready access to free tutoring supplied by their college, sufficient financial aid so that a student can spend time studying instead of working to pay tuition, or access to counselors to help with academic or emotional issues.

Finally, for the fourth question, what does *thinking critically* mean to students? This last question is a good example of how we let educational jargon creep into our surveys, and then assume that students understand what we mean. Recently, a graduate student interviewed me for a class she was taking about teaching, and asked me how I taught critical thinking in my classes. We then proceeded to have a discussion about what she meant by critical thinking, because I wasn't clear on what it meant, in terms of what she was asking. If I, as a higher education researcher, have trouble defining the phrase "critical thinking", how can we expect the average college student to understand the concept, much less ensure that this understanding is similar across college students?

There are of course many other examples in the NSSE and other college student surveys besides these. Given the lack of detail, and especially the lack of definitions, it is likely that students do not understand much of what we ask them. In part, these short, vague questions are probably due to an effort to reduce nonresponse bias, by reducing survey length through the use of short questions. But as recent research indicates, response rates cannot be viewed as a measure of survey data quality, and specifically, as a proxy for bias (Groves, 2006; Groves & Peytcheva, 2008). In trying to reduce nonresponse bias, NSSE and other researchers have inadvertently introduced additional substantial error into their survey data.

Consider the NSSE questions listed above, and then the question, "What is your income?" At first blush, this appears to be a reasonable question to pose, and its length and lack

of definitions is similar to many of the questions on the NSSE and other college student surveys. Now consider how one of the experts in the field of survey methodology, Floyd Fowler, thinks the question should be phrased to ensure common understanding across respondents (also note that he finds even this version somewhat lacking):

Next we need to get an estimate of the total income for you and family members living with you during 1993. When you calculate income, we would like you to include what you and other family members living with you made from jobs and also any income that you or other family members may have had from other sources, such as rents, welfare payments, social security, pensions, or even interest from stocks, bonds, or savings. So, including income from all sources, for you and family members living with you, how much was your total family income in 1993? (Fowler, 1995, p. 16)

Note how carefully the terms “your” and “income” are defined. It may seem silly to define a simple word like “your”, but astonishingly, research indicates that respondents may not even agree on what the word “you” means (e.g., the respondent, the respondent’s family, etc.) (Groves et al., 2004). Given the vagueness of the questions on the NSSE, as a matter of logic it seems clear that many respondents do not understand what they are being asked.

Unfortunately, there has been very little rigorous research on how students understand the questions we ask, and this is an area ripe for analysis. One study, however, did ask students what they meant when they used the vague response options, “occasionally”, “often” and “very often”,

by asking them the same question later in the survey, but this time providing a six-point response scale that ranged from “never” to “more than once a week” (Pace & Friedlander, 1982).³

Results from their study are presented in Table 1. Pace and Friedlander incorrectly concluded that questions using vague quantifiers in the College Student Experiences Questionnaire (the precursor to the NSSE) were valid measures of student frequency behavior, in part given that the response frequencies shifted upwards from the “occasionally” category to the “very often” category (the pattern highlighted in bold in the table). In other words, they make an argument based on the correlation between the two response options. But the bottom portion of the table, where students’ numeric responses have been translated to a common metric of number of times per year, reveals extensive disagreement. For example, one-fifth of students checking “very often” thought it meant around 60 times per year, one-third 32 times a year, another one-third 12 to 24 times per year, and one-tenth 1 to 6 times per year. In addition, there is substantial overlap of vague quantifier response by numeric category. By focusing on the overall relationship and ignoring other information, they inadvertently committed the correlation fallacy described above.

A better way to view their study is to ask the following question: if the meaning of “very often” was the same for all respondents, what should the distribution of responses by number of times per year look like? Because we can determine the number of times per year a student engages in an activity from their second set of responses, using the distributions in the Pace and Friedlander article we can estimate the standard deviation of the number of times per year for each of the three vague response categories. For example, if students in the “very often” category all had the same understanding of the term, they would all choose the same numeric response,

³ One issue with validation research is that any measure used to validate another measure must itself be valid. Here, Pace and Friedlander implicitly assume that the number of times response option is a valid measure, and I retain their assumption, although one could argue they both lack validity.

say, “about once a week”. The standard deviation would then be zero. If students perfectly disagreed (or randomly chose a numeric response), we would expect their responses to be evenly distributed across the six numeric categories, yielding a standard deviation of 13.7.

Figure 1 shows large disagreement among students as to what “occasionally”, “often” and “very often” mean in terms of actual frequency, so much so that their responses are similar to what we would see if there was perfect disagreement among students. Moreover, this disagreement is constant across a wide variety of topics. Such a result is not surprising, given other research that shows not only a lack of common understanding for vague quantifiers, but more troubling, variation in understanding across education and age groups (Tourangeau et al., 2000).

Figure 1 is important in two respects. First, it demonstrates that even for simple, everyday words, it is very easy for respondents to have a dissimilar understanding of terms. Second, the NSSE and many other student surveys use similar response options to question students about frequency of behavior, often for a majority of their questions. For example, the 2009 NSSE uses vague quantifiers for 67% of the items that do not ask about student background or demographics. If students differ about what these terms mean, then the validity of these survey questions is limited. Moreover, when vague wording leads to a lack of understanding about the information desired, respondents search for additional information to form an answer, and often use the distribution of the response scale to determine their answer, as explained below.

Retrieval. Most college student surveys implicitly view college students as having computer hard drives in their head. Do students really have accurate information in their brains about their behavior and attitudes that we desire, or are humans much more limited in the amount of information they store in their memories, and then provide to us via surveys? As Schwarz

(1990) notes, “Ideally, most researchers would like the respondent to scan the reference period, retrieve all instances that match the target behavior, and count them in order to determine the overall frequency of the behavior during the reference period. This, however, is the route respondents are least likely to take.” Memory research indicates that memories may not always be encoded; when they are, they may not be fully retrieved; and if they are retrieved, people have trouble assigning dates to memories (see Bradburn et al. (1987) for an outstanding brief overview of some of the problems in this area).

While there is debate as to how memories are encoded, it is clear that memory is not a film that we can rewind and play back at will. Instead, many researchers view memory as a sort of hierarchy, in which memories are grouped or stored in layers, according to attributes of the event. Memories may be linked through a series of indices, from more general groupings to more specific groupings, and are recalled by the way they are grouped (Tourangeau et al., 2000). For example, if memories of faculty contacts are encoded by class subject (English, chemistry), rather than access all the memories of faculty interactions encoded as a group, students might instead have to access the memory group of each class they have taken over the reference period, such as an academic year. If students first encode memories of events by semester, then by course, then by various aspects of the course (meetings with students in the course, meeting with that faculty instructor), the retrieval burden is substantial.

Research on memory and recall reveals several patterns (Tourangeau et al., 2000). First, recall fades with the passage of time. This vital fact is overlooked by many college student survey designers, who often ask questions not only about an academic year, but also about the entire academic experience during college in graduation exit surveys. Research suggests that the

ability of college students to recall even unique events fades after only a few weeks (Thompson, 1982).

Second, distinctive events are recalled more accurately than frequent and typical events. This can be seen most vividly in Garry et al.'s (2002) study of how accurately college students report frequency of sexual acts. Using a sample of students that had sex at least once a week, the researchers had students fill out detailed daily diaries about their sexual experiences for four weeks. Daily diaries are far more accurate than retrospective surveys, because respondents' memories of the day are fresh in their mind. Thus, they provide a way to validate frequency of behavior questions, which can be difficult to validate. Six to twelve months later, the students were given an unexpected follow-up survey testing their memory about their sexual activities for that month.

The results are unexpected: students in the follow-up survey over-reported the number of times having vaginal sex by almost 300% and oral sex by 100%-200% (see Figure 2). Social desirability bias should not affect the diaries, which were sent by email daily; students would not necessarily be aware they were reporting cumulatively large numbers of sexual acts, while for the follow-up survey they would be more likely to *reduce* reports of sexual activity due to social desirability bias. Most interestingly, anal sex was reported very infrequently, but memory reports of anal sex almost exactly matched diary reports. Their study suggests that unless students' contact with faculty and other academic experiences yield as vivid memories as having anal sex, it is unlikely they will be able to report accurately on them. If students are unable to accurately report the frequency of experiences such as vaginal and oral sex, which should be somewhat distinctive, how can we expect them to accurately report more mundane academic behaviors such as discussing grades with an instructor?

Third, besides basic issues of memory storage and retrieval, another major concern is the assigning of dates to memories. Research has shown that people have great difficulty assigning specific dates to memories: we may recall a doctor visit, but we may be unsure if it took place in the last six months, or the six months prior to the last six months. This causes problems for reporting activities for a specific time period, as telescoping can occur, in which more distant events are recalled as having occurred more recently. In addition, respondents report greater frequencies when they can recall many memories, regardless of the actual frequency of the event (referred to as the availability heuristic) (Bradburn et al., 1987).

Combined, these two problems in recall suggest that the longer a student has been in school, the more memories they will have, and they will thus report more frequent behavior. This could explain the common finding that seniors are more engaged than first-year students: they simply have more memories of academic experiences, and memories from sophomore and junior years are “leaking” into their recall because of an inability to assign dates to memories. Moreover, if students with higher cognitive ability are able to recall more memories, this suggests that findings related to SAT scores at both the individual and institutional levels may in part be driven by recall issues: more selective institutions show higher levels of engagement (Porter, 2006) due to the availability heuristic, not due to their selectivity.

Finally, research indicates that the more time respondents spend trying to recall information, the more accurately they report it. This leads to the recommendation that survey researchers use long introductions to questions as a way to increase recall effort (Tourangeau et al., 2000). Most college student surveys unfortunately take the opposite approach, and we can see another drawback to using short, vague questions like the NSSE does: besides increasing student misunderstanding of what information we want from them, it also allows students to

quickly answer the question without adequately recalling all relevant memories. Given how much we depend on student recall of events during college in our surveys, it is remarkable how little research has been conducted by the higher education research community on student memory and recall.

Judgment. Research indicates that respondents rarely keep a tally of events, nor are they generally able to recall and count each event to generate a frequency. From the memory discussion, it is clear that students cannot keep an exact tally of the majority of behaviors in the NSSE and other college surveys, especially given that the NSSE asks about behaviors over an entire academic year. Nor will they be able to recall all exact events and count them. Instead, students will resort to a variety of estimation strategies. They may recall a few events, estimate a rate based on these few events, and apply it over the reference period of the survey question. They may use some form of a recommended rate, or use a general impression to estimate an answer.

For example, most people cannot accurately recall the number of times they have visited a dentist during the year; even if they can retrieve memories, it can be difficult to determine which memories of visits occurred in which year. So many people instead report two times a year, knowing that they are supposed to visit a dentist every six months.

Assuming that students can accurately recall and report the frequency of events is probably the most fundamental theoretical flaw of the NSSE and other college student surveys. They assume that because students do not have a reason to falsify their responses (such as not wanting to answer questions about drug use), their reports must be accurate. Moreover, researchers fail to understand that most of what they ask on college student surveys is in the domain of mundane behaviors, which are notoriously difficult for respondents to recall. A

student may remember how many times they were hospitalized for surgery during the past year, but it is very unlikely that they can accurately recall how often they came to class without completing readings or how often they asked questions in class.

If students do not have the requisite information, how do they answer the question? Due to the norm of helping tendencies, respondents almost always want to give an answer, even if they are unsure about the accuracy of their answer. So they rely on a variety of cues when developing an estimate, taken from themselves, their surroundings, and the survey. They may reason based on self-image – “I am a good student, and I know that good students meet with faculty, have serious conversations, etc., therefore I must be doing these activities on a frequent basis.” Research has indicated that halo effects, or responses to individual items being driven in part by general perceptions, are fairly strong for college students (Pike, 1999). This possibility is particularly problematic, as it explains much of the cross-sectional correlational research between student self-reports of engagement behavior and criterion such as grade-point average and achievement tests (e.g., Pike, 1995). Positive correlations occur because students may use their academic performance to infer an answer.

Students may also infer logically, by using a causal theory of what they think should be happening (Bowman, in press). When asked how much they have increased their quantitative skills, a student may reason that because they are a science major and have taken many science classes, they should have increased their quantitative skills, regardless of the actual change.

Extensive evidence documents how respondent answers are affected by the context, when asked questions that cannot be answered by a tally or recall and count strategy (Tourangeau et al., 2000). For example, a typical question on college student surveys, as well as the NSSE, asks students to provide the number of hours spent on various activities, such as studying. Given the

preceding discussion on memory, recall and judgment, it should be clear that it is impossible for the typical person to give an accurate answer to these types of questions, which in turn implies that context effects should be large.

One study asked college students to report the number of hours spent studying, with two different scales for two random samples: one scale ranged from “.5 hours or less” and ended at “more than 2.5 hours”, while the other started at “2.5 hours or less” and ended at “more than 4.5 hours”. This would appear to be a trivial difference in scaling, yet the proportion of students reporting more than 2.5 hours of studying was 30% for the first sample and 71% for the second, a rather stunning differential (Smyth, Dillman, & Christian, 2007). Why the large effect for such a small change in question wording? Given an inability to recall and report an answer based on memories, students turned to context help estimate an answer, and interpreted the middle of the scale as “normal”. They then responded based on whether they believed their study hours were normal or not.

This is another area of college student survey response that is ripe for exploration. While context effects on surveys are well-known, less is known about how student’s theories of causality and self-image affect survey response. We also do not know how much the college context affects survey response – do individual students at schools with a “grind” reputation report frequent studying because they actually spend a lot of time studying, or because that is their school’s reputation, and they thus make the logical inference that because they are a student there, they must study a lot? If context effects vary by student or college characteristics, then this raises serious questions about differentials found across schools.

Response. Finally, students may consciously alter their answer when choosing a response option after the estimation stage. Social desirability bias, or providing survey responses that

make the respondent look favorable to others, has been extensively documented in the field of survey methodology, but has been little studied in the area of college student survey response. The only study I have found to date suggests that social desirability bias, as measured by a well-known scale, is partially driving student responses to questions about learning gains (this is a study by Nick Bowman that is currently in progress). Again, this is an area ripe for analysis: we simply do not know how much social desirability bias affects college student survey response.

This brief review paints a different portrait of college student response behavior than the NSSE validity argument. If my counter-argument is true, an examination of validation evidence using college students should reveal differences between what students report and what actually occurs. I have reviewed the literature, and Table 2 summarizes the results of the student criterion studies that I have found to date. Because validity evidence that focuses on criterion measures is vulnerable to criticism because the criterion itself must be validated, I focus on studies where student self-reports are compared to measures drawn from school databases. While any database contains some error, institutional databases are probably the most error-free data source that we can use to study student self-reports.

Table 1 demonstrates that even with fairly accessible information, such as SAT score and grade point average, students are not accurate reporters. For example, Cole and Gonyea (in press) found a strong correlation between self-reported and actual SAT scores, but only 62%-70% of respondents could accurately report within +/- 20 points of their true scores. A meta-analysis of 37 samples found that only half of college students could accurately report their grade point average (Kuncel, Crede, & Thomas, 2005), while a study of high school seniors found that only three-quarters could accurately identify whether they had taken a U.S. history course in high school (Niemi & Smith, 2003). Surprisingly, less than a third of independent students (that is,

students supporting themselves and who should thus know their income) could accurately report their income (Olivas, 1986).

In addition, we can see that fairly strong correlations can be found, even if the accuracy rate is low. More troubling is the direction of the errors. If student reporting errors were relatively small and essentially random, one could argue that self-reports could be used as valid measures of student outcomes. But as the last column indicates, the reporting error is always in the same direction, in that errors are always in a positive direction for a student's self-image. Students over-report grade point averages, SAT scores, and taking certain types of courses, and they under-report failing classes, being on financial aid, and the amount of financial aid they receive. It is unclear whether these errors occur at the judgment stage, and reflect respondents' use of estimation strategies, or if they occur at the response stage, and reflect student unwillingness to accurately report negative information.

Finally, some studies have found that accuracy varies by student ability: low ability students are less likely to be accurate reporters (Cole & Gonyea, in press; Kuncel et al., 2005). This can be seen most clearly in Figure 3, taken from the Kuncel et al. meta-analysis. The correlation between self-reports and actual academic performance drops dramatically at lower levels of academic performance; they find a similar pattern using measures of cognitive ability. Such a finding is disconcerting, because the average level of student cognitive ability varies across colleges, implying that college comparisons using survey data may be flawed due to the correlation between cognitive ability and the accuracy of self-reports.

Evidence based on internal structure (Argument Four)

NSSE researchers assert that there are five engagement constructs in the NSSE: level of academic challenge, active and collaborative learning, student-faculty interaction, enriching

educational experiences, and supportive campus environment. These constructs comprise the five NSSE “Benchmarks of Effective Educational Practice”, and schools are rated on how they compare on each of these benchmarks. Schools are provided with their mean for each benchmark, as well as means for comparison groups, and schools may voluntarily release where they stand on the Benchmarks. These data are increasingly being used by as “a new arbiter of quality for higher education in America” (National Survey of Student Engagement, 2007, p. 5).

Given how much emphasis is placed on the Benchmarks, it is somewhat surprising to learn that other researchers have had difficulty replicating the five construct system used to measure student engagement. Using the NSSE, two sets of scholars (LaNasa, Cabrera, & Transgrud, 2009; LaNasa, Olson, & Alleman, 2007) have found eight separate constructs measuring student engagement at a single institution, while another found only three constructs at one institution (Swerdzewski, Miller, & Mitchell, 2007) (these are the only studies I have found that have attempted to independently confirm that student engagement is comprised of five distinct dimensions).

The fact that the NSSE’s conceptual structure could not be replicated is troubling, but not surprising; while “the benchmarks were created with a blend of theory and empirical analysis” (National Survey of Student Engagement, 2009d) it seems clear from a review of NSSE documents that the amount of theory is somewhat lacking in favor of empirical analysis. Why student engagement should consist of five distinct dimensions rather than four or six is never explained, yet a good conceptual framework would specify why this is the case. Instead, NSSE researchers seem to have relied on face validity and results of factor analyses to determine the internal structure (Kuh et al., 2001). Thus, it should come as no surprise that this internal structure is not replicated across institutions. Yet if the internal structure cannot be replicated,

this raises concerns about its validity, and most importantly, to its validity claim that it can be used to measure quality across different institutions.

Reliability. Although this paper is concerned with the validity argument of the NSSE, the reliability of the NSSE bears mentioning. Most researchers would argue that a valid yet unreliable instrument is not very useful, and many also consider reliability to be part of a broader definition of validity (Kane, 2006).

The measure most often reported by researchers using the NSSE is the reliability coefficient alpha. As with any rule of thumb, the minimum alpha for a scale varies from researcher to researcher, and ranges from .70 (DeVellis, 2003; Spector, 1992) to .80 and higher (Carmines & Zeller, 1979; Rosenthal & Rosnow, 1991). As Garson (2009) notes, “that .70 is as low as one may wish to go is reflected in the fact that when alpha is .70, the standard error of measurement will be over half (0.55) a standard deviation,” and Spector (1992) questions the use of scales with reliabilities less than .70.

Using the lower of proposed cutoffs, .70, most researchers would conclude that the NSSE scales are unreliable. Table 3 shows the distribution of Cronbach’s alpha for NSSE Benchmarks reported by the Center for Postsecondary Research at Indiana for the full set of participating NSSE schools. As can be seen, 40% of the Benchmark scales fail to meet the minimum required alpha of .70, and it is rare for the reliabilities to be higher than what many researchers consider to be the minimum, .80. A review of research publications using NSSE data in higher education journals reveals a similar pattern. With half of its scales failing to achieve widely accepted minimum reliabilities, the NSSE cannot be viewed as a very reliable instrument.

Another approach to reliability is the test-retest method; items are asked at two points in time, and the correlation or similar statistic for the two sets of answers is calculated. If responses

have not changed over the time period, then the expected correlation is 1.0. NSSE researchers have used two test-retest approaches (National Survey of Student Engagement, 2009c), both flawed, because they average out error and thus inflate the reported reliabilities.

NSSE researchers have calculated averages at the institutional level for schools participating in the NSSE in consecutive years, and find correlations of .78 to .92 between two time periods. This result is not surprising; indeed, it is surprising that the correlation is not higher. Suppose the mean score on a benchmark for an institution is 5.0 for Time 1 as well as for Time 2. Next, assume that at Time 2, students take their Time 1 answer, randomly draw a number from -5 to +5 and add this to their Time 1 answer. The overall mean at Time 2 will remain 5.0, and match the Time 1 score, thus leading to a high reliability at the institutional level. But the correlation between Time 1 and Time 2 scores at the student level will be close to zero, which is why this empirical approach cannot be found in the literature on reliability.

In addition, NSSE researchers have looked at students who have accidentally filled out the NSSE twice during an administration, and compared their benchmark scores, finding correlations ranging from .69 to .78. Given that the NSSE asks for behavior for an entire academic year, and that the two reports are only a few weeks apart, the correlation here should be close to 1.0. However, the correlation is much lower because the individual items that make up the benchmarks fluctuate between the two time periods, reducing the correlation. The benchmark scores, in essence, average out the large error in the individual items and inflate the reliabilities. Given the relatively low reliabilities for the benchmark scores, the reliabilities for the individual items are probably much lower.

Evidence based on relations to other variables (Argument Five)

There have been several studies that measure the relationship between individual NSSE items and scales with other data, such as alternative measures of the item or measures of student learning and other student outcomes.

Relationships at the item level. One way to establish validity evidence is to demonstrate that an item varies convergently or divergently as theory would predict. Pace (1985) uses such an approach when he shows that arts and humanities majors report higher gains in an understanding of art and literature compared to the average student, while science majors report higher gains in understanding scientific developments and quantitative thinking. Kuh (2001) reports similar findings. Similar findings have been reported in terms of differences between institution types. As noted above, such findings can easily be explained by students reasoning what their gains should be, rather than accurately reporting their actual gains.

Two studies have assessed the relationship between items on the NSSE and alternative measures of the same items using external data. Porter et al. (2009) uses the NSSE question that asks students to report the number of books and coursepacks assigned in their courses during the academic year. Using registration data, they then contacted the instructors of respondents, and constructed an independent measure of the number of books and coursepacks by coding syllabi. The correlation between the two measures was .38, but they found that only 21% of students could accurately report the number of books assigned, while given their scale, 17% would have given a correct answer if they had randomly chosen a response.

In addition, they found a .21 correlation between the student self-report of books and number of hours spent studying, similar to the .25 correlation between books and studying on the NSSE reported by Kuh (2001), even though the student self-report is almost unrelated to the

actual number of books. As they note, this illustrates the danger of relying on relatively low correlations between survey items to establish validity.

Although this paper focuses on questions about factual data, questions about subjective data such as attitudes are even more difficult to collect. The problem is that most researchers adopt a file-drawer model of attitudes, in which attitudes exist in a respondent's head, and all the respondent has to do is reach in, read the file, and report an answer. Research on attitudes demonstrates the exact opposite: most attitudes are rarely fully formed until a respondent reads the question, and attitudes vary greatly over time, due to respondents forming and re-forming an attitude each time they are asked a question about that attitude (Tourangeau et al., 2000).

Given the nature of the learning gains questions, one can argue that these are really attitudinal questions rather than factual questions. If so, we would expect these responses to be rife with error, and to have almost no relationship to actual measures of student gains. Bowman (in press) has conducted an important study in which he compares actual student gains in critical thinking and moral reasoning, based on valid and reliable pre- and post tests administered at the beginning and end of the first year of college, and he compares these actual objective gains to self-reported gains in these areas using the NSSE. He finds no statistically significant correlations between the actual and reported learning gains, regardless of how the objective change is calculated.

Relationships at the scale level. Researchers have attempted to demonstrate that scales from the NSSE and similar surveys are correlated with student learning, as the NSSE validity argument argues it should be. Some of these studies use grade-point average as a measure of student learning, but GPA is flawed in many respects. Grading practices vary across institutions and within institutions by major, calling into question exactly what an A or B average represents.

More importantly, if students use estimation strategies to answer questions, as is likely given the time frame and content of typical student surveys, they may rely on self-image to estimate a response. Because academic self-image will be based on current student performance, we could find a positive correlation between GPA and behaviors simply due to response errors on the part of the respondent. Conversely, students generally have not taken the tests listed below and do not know how they will perform, and it is more difficult to argue that performance will drive their responses to the NSSE.

Thus, the best approach is to use actual measures of student learning that, unlike GPA, have demonstrated validity and reliability. Two such studies (Carini et al., 2006; Pascarella & Seifert, 2008) have used such measures in an attempt to show that NSSE scales are correlated with learning. They use a series of tests developed by RAND to measure critical thinking and performance (Klein, Kuh, Chun, Hamilton, & Shavelson, 2005), two writing tasks from the GRE asking students to critically analyze a topic, the critical thinking module from the Collegiate Assessment of Academic Proficiency, and the Defining Issues Test 2, which provides a measure of higher-order moral reasoning. Taken together, these two studies are probably the most sophisticated analyses of the relationship between the NSSE and student learning, especially the Pascarella and Seifert study, because their study also includes a pre-test measure of the two learning tests. In other words, they look at the effect of NSSE scales on gains in learning since the beginning of college, rather than a simple cross-sectional approach as used by Carini et al.

Both studies present relationships between several NSSE scales and the outcomes listed above, controlling for a host of other variables. I list these, rather than the simple relationships, as NSSE researchers have asserted that these are the appropriate estimates to use when validating the NSSE (Carini et al., 2006). Table 4 shows the distribution of the findings for the four

outcome measures, classified by the strength of the association (one study used partial correlations, while the other reports standardized regression coefficients, which express the change in the outcome variable in standard deviations, given a one standard deviation change in the NSSE scale). Statistically insignificant relationships are classified as zero, as they indicate no association between the NSSE scale and the outcome in the population.

As can clearly be seen, the number of zero or negative relationships ranges from 40% to 75%, and out of the 46 associations calculated in the two studies, *over half* indicate no relationship, and none yields a correlation or standardized regression coefficient larger than .20. Considering the Pascarella and Seifert study, which I consider the stronger study given their use of pre-measures of learning, only one association was larger than .10.

One concern I have raised is that higher ability students may report higher frequencies of student behavior, which yields a spurious correlation between student behavior scales (such as engagement) and single test measures of learning (that is, when we only have a measure of student learning at one point in time). If true, this implies that correlations between engagement and learning may drop in size or disappear when we control for academic ability.

The Carini et al. validation study provides a nice test of this hypothesis. Recall from Table 4 that roughly half of the correlations they estimated were statistically different from zero. If student ability is having an effect, then if the correlations are estimated on low ability students and high ability students separately, we would expect fewer significant correlations. With two groups, two measures of learning, and 15 scales, they estimate 60 partial correlations. Fifty of these, or fully 80%, are statistically insignificant. In other words, they find fewer relationships when analyzing low and high ability (measured by SAT) students separately.

Naturally, this comparison begs the question of what we should consider to be a large effect size; however, the NSSE validity argument clearly states that the NSSE scales are “highly correlated” with learning (Kuh et al., 2001, p. 2). To consider this another way, over half the associations that were found were statistically insignificant. Given this, how many associations must be zero before we as a field recognize that there is no overall association? I believe that when a validity argument states that measures should be highly correlated with learning outcomes, and we fail to find even modest associations for over half of the measures, then the evidence that the measures are valid is very limited.

Discussion

I have argued that the NSSE serves as a model for college student surveys, and that under close examination, their validity argument fails, thus calling into question most college student surveys used in the field of postsecondary research. First, the domain specification for the NSSE is overly broad and driven by empirical rather than theoretical concerns. Second, college students, as with all humans, have trouble encoding and accurately reporting on behavior and events, especially if they are mundane, and thus rely on a variety of estimation strategies that can result in large, systematic reporting errors. The few studies that have compared students self-reports to corresponding databases support this notion, and show that the errors are not random, but are instead biased in such a way that puts the student in a good light. The unnecessarily vague wording of questions on the NSSE also contributes to reporting errors, due to different understandings of question meaning. Third, the dimensional structure of the benchmarks proposed by the NSSE has not been replicated by other researchers, and many of their reliabilities fail to meet minimum standards. Fourth, studies measuring the association between NSSE items and scales and external data reveal limited associations, and research demonstrates

that scales derived from the NSSE are largely uncorrelated with objective measures of student learning.

Given that most student surveys in postsecondary research rely on the same assumption of human cognition as the NSSE, use the same types of vaguely worded questions, often have low reliabilities, and demonstrate limited associations with data external to the survey, it is clear that the grand statistical edifices that we have created rest on very shaky foundations. Based on this review of the literature, I call into question most of the research on student engagement, student development, and other postsecondary areas that rely on similar surveys of college students. I should note that I include my own research here as well. Looking over my research on students, I realize that I cannot produce any evidence that the survey questions I used were valid, except in terms of face validity: they seem plausible based on how they are worded. But as we have seen with the example of number of hours spent studying, even subtle changes in question wording can yield very large changes in response. For too long, we have relied on the notion of face validity to defend our survey questions, while the rest of the social sciences have long abandoned the entire notion of face validity in favor of more rigorous definitions of validity. While there are undoubtedly a few exceptions, particularly higher education surveys that use scales developed and validated by psychologists, when confronted with the question “What is the validity of your survey instrument?”, most higher education researchers studying college students would not have much to say.

Moreover, if students cannot accurately report information about their academic experiences, then this raises questions of how we measure institutional performance, particularly in a comparative context. Evidence suggests that survey response behavior, in terms of understanding, recall, and accuracy in reporting, may vary by individual characteristics. If this is

true for college students, because our institutions vary in terms of student characteristics, many institutions may be unfairly labeled as underperformers in terms of engagement and other student outcomes, when the difference may simply be an artifact of a poorly worded survey. Given the intense desire in higher education to measure institutional performance, it is truly an indictment of our field that we have not conducted intensive research into variations in survey response behavior across student types. It is also troubling that the research that has taken place on college student memory and recall has almost entirely taken place outside the field of higher education. Within academia, we pride ourselves on being the preeminent scholars of college students, yet the most important research on college student survey response is conducted by economists, psychologists, and sociologists.

In terms of what surveys mean for institutions, a parallel here is high-stakes testing in the K-12 arena. Absent other ways to assess learning, college student surveys have become our own version of high-stakes testing. Institutions, programs, and departments are constantly being evaluated based on student survey data. Thus, we as a field must pay much more attention to the validity of the surveys we develop, use, and offer to others for their use.

How did we get here?

Having raised these issues, it would also seem relevant to offer an explanation as to why our field seems so dependent on poorly constructed surveys that assume students have almost superhuman powers of recall. To my mind, three trends are responsible: a lack of training, the demand for publications, and the demand for quick fixes to the problem of how we assess student learning.

First, it is notable that the vast majority of higher education programs do not offer courses on survey methodology in particular and measurement in general, and while some colleges of

education offer such courses, many scholars fail to avail themselves of this opportunity. It is somewhat ironic that we train generation after generation of scholars in how to use sophisticated quantitative methods to analyze survey data, but pay little attention to how these data are generated.

Second, as these scholars begin their careers, the publish-or-perish imperative of the tenure system takes over, pressuring scholars to quickly collect data and churn out articles. As I have hopefully made clear, collecting valid and reliable survey data is a resource-intensive process, in terms of both time and money. Faced with the choice of spending a year developing and validating a survey to be used for one article, versus quickly dashing off a survey that can always be defended with face validity, an assistant professor would be foolish to devote a year to one survey given the reward structures of most universities. If our field does not enforce high standards of validity and reliability, researchers will continue to take the easy way out during the quest for tenure and promotion.

Third, in terms of the NSSE, adequately assessing student learning is something that has been an issue for postsecondary education for some time. The promise of a survey instrument that can quickly and relatively cheaply provide an alternative to actually measuring learning has, not surprisingly, been alluring to many colleges. That an instrument that fails to meet basic standards of validity and reliability has been so quickly adopted by numerous institutions indicates the desire of many institutions for a solution to this issue.

Where do we go from here?

The issues raised in this paper lead to a series of brief recommendations for surveys of college students.

Design. Perhaps the most important change we can make in terms of how we design our surveys is to understand the limited cognitive ability of humans when confronted with a survey question. Contrary to popular belief, people have difficulty accurately reporting even simple information about themselves, especially after a short time period. Furthermore, we should understand that designing good questions about mundane behavior is very difficult, given memory and reporting issues.

The time frames of our questions should correspond to the item being asked. For example, asking students about their frequency of “coming to class without completing readings or assignments” requires a much shorter reference period than asking about the frequency of making a class presentation, as the former may occur every day, while the latter may occur only once a year. One of the primary sources used by sociologists and economists to understand time use is the Bureau of Labor Statistics’ American Time Use Survey, which asks respondents about their activities for only the previous 24 hours. The Berea Panel Study of college students, with its multiple surveys of the preceding 24 hours, is an excellent example of how we should be studying the effects of student behavior on academic outcomes (Stinebrickner & Stinebrickner, 2004).

When possible, researchers should use time-use diaries rather than surveys. This is a standard approach in other fields, because researchers understand the limited ability of humans to recall and accurately report on their daily activities even a few days after the events occur. This will require a serious re-orientation of how we study students, because time-use diaries are expensive: respondents must usually be paid a significant sum due to the amount of time spent filling them out, and converting the diaries to a usable dataset is also time consuming.

Finally, we as a field should abandon questions that appear to contradict theory and research on human cognition, and that research shows are not accurate representations of

behavior. The most obvious set of questions here are the self-reported learning gains that are used in so many college student surveys. It is clear that the majority of people cannot accurately report their learning gains, and recent research demonstrates this. Questions asking students to report the number of hours spent on various activities should also be abandoned, unless the reference period is quite short. Most importantly, we should recognize that for the majority of topics we are interested in, students can only accurately report on the previous week, and not the current semester or academic year.

Evaluation. In terms of evaluating surveys, we should understand that it is fairly easy to find small correlations between variables, and that correlations can be misleading without any additional analyses. We should also establish criteria for judging before beginning validation research; for example, what do we mean by “highly correlated”? Much of the current validation research in higher education appears to take what I think of as the “greater than zero” approach – if a correlation, standardized regression coefficient or reliability measure is greater than zero, then all is well. Clearly we as a field need to establish some commonly accepted minimum levels for judging relationships.

Researchers should also seek stronger evidence of validity than we have currently done in the past. For a given survey, it is relatively easy to look for convergent or divergent validity within the data. Consider again the example of asking students their height and weight in a survey. Higher education researchers might validate this by calculating the correlation between the reported number of desserts consumed per month and self-reported weight, or whether students on the basketball team report being taller than students on the chess team. Finding a positive correlation or difference, they would conclude that the questions are valid, missing

entirely the fact that students overreport their height and underreport their weight, and that these errors are correlated with student characteristics.

Finally, and most importantly, the tacit agreement in our field seems to be that validity is assumed until proven otherwise. Instead, we must establish standards such that a *lack* of validity is assumed until proven otherwise.

References

- Achen, C. H. (1977). Measuring representation: Perils of the correlation coefficient. *American Journal of Political Science*, 21(4), 805-815.
- Achen, C. H. (1978). Measuring representation. *American Journal of Political Science*, 22(3), 475-510.
- Bowman, N. A. (in press). Can first-year students accurately report their learning and development? *American Educational Research Journal*.
- Bradburn, N. M., Rips, L. J., & Shevell, S. K. (1987). Answering autobiographical questions: The impact of memory and inference on surveys. *Science*, 236(4798), 157-161.
- Brener, N. D., McManus, T., Galuska, D. A., Lowry, R., & Wechsler, H. (2003). Reliability and validity of self-reported height and weight among high school students. *Journal of Adolescent Health*, 32(4), 281-287.
- Carini, R., Kuh, G. D., & Klein, S. P. (2006). Student engagement and student learning: Testing the linkages. *Research in Higher Education*, 47(1), 1-32.
- Carmines, E. G., & Zeller, R. A. (1979). *Reliability and Validity Assessment*. Thousand Oaks, CA: Sage Publications.
- Cole, J. S., & Gonyea, R. M. (in press). Accuracy of self-reported SAT and ACT test scores: Implications for research. *Research in Higher Education*.
- DeVellis, R. F. (2003). *Scale Development: Theory and Applications*. Los Angeles: Sage Publications.
- Fowler, F. L. (1995). *Improving Survey Questions: Design and Evaluation*. Thousand Oaks: Sage Publications.

- Garry, M., Sharman, S. J., Feldman, J., Marlatt, G. A., & Loftus, E. F. (2002). Examining memory for heterosexual college students' sexual experiences using an electronic mail diary. *Health Psychology, 21*(6), 629-634.
- Garson, D. (2009). Scales and Standard Measures. Retrieved October 13, 2009., from <http://faculty.chass.ncsu.edu/garson/PA765/standard.htm>
- Gordon, J., Ludlum, J., & Hoey, J. J. (2008). Validating the NSSE against student outcomes: Are they related? *Research in Higher Education, 49*, 19-39.
- Groves, R. M. (2006). Nonresponse rates and nonresponse bias in household surveys. *Public Opinion Quarterly, 70*(5), 646-675.
- Groves, R. M., Fowler, F. L., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2004). *Survey Methodology*. Hoboken, NJ: Wiley-Interscience.
- Groves, R. M., & Peytcheva, E. (2008). The impact of nonresponse rates on nonresponse bias: A meta-analysis. *Public Opinion Quarterly, 72*(2), 167-189.
- Kane, M. T. (1992). An argument-based approach to validation. *Psychological Bulletin, 112*, 527-535.
- Kane, M. T. (2001). Current concerns in validity theory. *Journal of Educational Measurement, 38*(4), 319-342.
- Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational Measurement* (pp. 17-64). Westport: American Council on Education and Praeger.
- Klein, S. P., Kuh, G. D., Chun, M., Hamilton, L., & Shavelson, R. (2005). An approach to measuring cognitive outcomes across higher education institutions. *Research in Higher Education, 46*(3), 251-276.
- Kozak, M. (2009). What is a strong correlation? *Teaching Statistics, 31*(3), 85-86.

- Kreuter, F., Presser, S., & Tourangeau, R. (2008). Social desirability bias in CATI, IVR, and web surveys. *Public Opinion Quarterly*, 72(5), 847-865.
- Kuh, G. D. (2004). *The National Survey of Student Engagement: Conceptual Framework and Overview of Psychometric Properties*. Bloomington: Indiana University Center for Postsecondary Research
- Kuh, G. D., Hayek, J. C., Carini, R. M., Ouimet, J. A., Gonyea, R. M., & Kennedy, J. (2001). *NSSE Technical and Norms Report*. Bloomington: Indiana University Center for Postsecondary Research and Planning.
- Kuncel, N. R., Crede, M., & Thomas, L. L. (2005). The validity of self-reported grade point averages, class ranks, and test scores: A meta-analysis and review of the literature. *Review of Educational Research*, 75(1), 63-82.
- LaNasa, S. M., Cabrera, A. F., & Transgrud, H. (2009). The construct validity of student engagement: A confirmatory factor analysis approach. *Research in Higher Education*, 50, 315-332.
- LaNasa, S. M., Olson, E., & Alleman, N. (2007). The impact of on-campus student growth on first-year student engagement and success. *Research in Higher Education*, 48(8), 941-966.
- Lissitz, R. W., & Samuelsen, K. (2007). A Suggested Change in Terminology and Emphasis Regarding Validity and Education. *Educational Researcher*, 36(8), 437-448.
- McCartney, K., & Rosenthal, R. (2000). Effect size, practical importance, and social policy for children. *Child Development*, 71(1), 173-180.
- Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational Measurement* (pp. 13-103). New York: Macmillan.

- National Survey of Student Engagement. (2007). *Experiences That Matter: Enhancing Student Learning and Success*. Bloomington: Indiana University Center for Postsecondary Research.
- National Survey of Student Engagement. (2009a). NSSEville State University - Frequency Distributions. Retrieved October 13, 2009
- National Survey of Student Engagement. (2009b). NSSEville State University - Mean Comparisons. Retrieved October 13, 2009
- National Survey of Student Engagement. (2009c). Reliability. Retrieved October 13, 2009, from http://nsse.iub.edu/html/PsychometricPortfolio_Reliability.cfm
- National Survey of Student Engagement. (2009d). Survey Development. Retrieved October 13, 2009, from http://nsse.iub.edu/html/PsychometricPortfolio_SurveyDevelopment.cfm
- Niemi, R. G., & Smith, J. (2003). The accuracy of students' reports of course taking in the 1994 National Assessment of Educational Progress. *Educational Measurement: Issues and Practice*, 22(1), 15-21.
- Olivas, M. A. (1986). Financial aid and self-reports by disadvantaged students: The importance of being earnest. *Research in Higher Education*, 25(3), 245-252.
- Pace, R. C. (1985). *The credibility of student self-reports*. Los Angeles: Center for the Study of Evaluation, University of California Los Angeles.
- Pace, R. C., & Friedlander, J. (1982). The meaning of response categories: How often is "occasionally," "often," and "very often"? *Research in Higher Education*, 17(3), 267-281.
- Pascarella, E. T., & Seifert, T. A. (2008). *Validation of the NSSE benchmarks and deep approaches to learning against liberal arts outcomes*. Paper presented at the Association for the Study of Higher Education.

- Pike, G. R. (1995). The relationships between self reports of college experiences and achievement test scores. *Research in Higher Education, 36*, 1-22.
- Pike, G. R. (1999). The constant error of the halo in educational outcomes research. *Research in Higher Education, 40*(1), 61-86.
- Porter, S. R. (2006). Institutional structures and student engagement. *Research in Higher Education, 47*(5), 521-558.
- Porter, S. R. (2009). *Validity and response bias in student engagement survey questions: Can we accurately measure academic challenge?* Unpublished manuscript, Ames.
- Rosenthal, R., & Rosnow, R. L. (1991). *Essentials of Behavioral Research: Methods and Data Analysis*. Boston: McGraw-Hill.
- Schwarz, N. (1990). Assessing frequency reports of mundane behaviors. In C. Hendrick & M. S. Clark (Eds.), *Research Methods in Personality and Social Psychology*. Newbury Park: Sage Publications.
- Smyth, J. D., Dillman, D. A., & Christian, L. M. (2007). Context effects in web surveys: New issues and evidence. In A. Joinson, K. McKenna, T. Postmes & U. Reips (Eds.), *Oxford Handbook of Internet Psychology* (pp. 427-443). New York: Oxford University Press.
- Spector, P. E. (1992). *Summated Rating Scale Construction: An Introduction*. Newbury Park: Sage Publications.
- Stinebrickner, R., & Stinebrickner, T. R. (2004). Time-use and college outcomes. *Journal of Econometrics, 121*, 243-269.
- Swerdzewski, P., Miller, B. J., & Mitchell, R. (2007). *Investigating the validity of the National Survey of Student Engagement*. Paper presented at the Northeastern Educational Research Association, Rocky Hill, CT.

Thompson, C. P. (1982). Memory for unique personal events: The roommate study. *Memory and Cognition*, 10(4), 324-332.

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

Trusheim, D. (1994). How valid is self-reported financial aid information? *Research in Higher Education*, 35(3), 335-348.

Table 1. Consistency of the Meaning of Occasionally, Often, and Very Often (Make an Appointment to See a Faculty Member)

		<i>Among students who had checked:</i>		
<i>% defining their activity as:</i>		Occasionally	Often	Very often
Calculated as a percentage distribution	Never	4%	0%	0%
	Once or twice a year	37%	6%	2%
	3 to 6 times a year	38%	24%	9%
	1 or 2 times a month	18%	45%	33%
	About once a week	3%	21%	35%
	More than once a week	<u>0%</u>	<u>4%</u>	<u>21%</u>
		100%	100%	100%
Calculated as the number of times per year	Never	0	0	0
	Once or twice a year	1 to 2	1 to 2	1 to 2
	3 to 6 times a year	3 to 6	3 to 6	3 to 6
	1 or 2 times a month	12 to 24	12 to 24	12 to 24
	About once a week	32	32	32
	More than once a week	64 or more	64 or more	64 or more

Note: Top half of table taken from Table 1 of Pace and Friedlander (1982); bolding replaces their brackets.

Table 2. Student Reporting Accuracy, Selected Studies.

Study	Survey item	External data	Findings		
			Accuracy rate	<i>r</i>	Direction of bias
Cole & Gonyea (in press) ^a	SAT scores for reading, math and writing	College databases	62% - 70%	.86 to .88	Over
	ACT score	College databases	89%	.95	Over
Kuncel et al. (2005)	Total SAT score	College databases	36%	.82	Over
	College GPA	College databases	54%	.90	Over
Niemi & Smith (2003)	H.S. seniors: took U.S. history course in past	High school transcripts	69% ^b	NA	Over
	H.S. seniors: currently taking U.S. history course	High school transcripts	73% ^b	NA	Over
Kreuter et al. (2008) ^c	Alumni: at least one D or F	College databases	FN: 20% - 33%	NA	Under
	Alumni: academic probation	College databases	FN: 25% - 33%	NA	Under
	Alumni: received academic honors	College databases	FP: 5% - 6%	NA	Over
	Alumni: donated to school	College databases	FP: 8% - 11%	NA	Over
Olivas (1986) ^d	Dependent students: family income	College databases	22%	NA	Over
	Independent students: family income	College databases	29%	NA	Over
Trusheim (1994)	Received financial aid (yes/no)	College databases	91% - 93%	NA	Under
	Dollar amount of aid received - student loans	College databases	NA	.46 to .76	Under
	Dollar amount of aid received - work study	College databases	NA	.28 to .68	Under
	Dollar amount of aid received - state aid	College databases	NA	.45 to .65	Under

Note: FN indicates false negative, FP false positive; percentage accurate was not given. NA indicates not available.

^a Accuracy rate defined as within +/- 20 points.

^b Calculations by the author based on their Table 2.

^c Alumni survey conducted 3 to 16 years after graduation; thus some error here is due to the length of time between events and the reporting of events.

^d Accuracy rate defined as within +/- \$500 in 1979 dollars; this is almost +/- \$1,500 in 2008 dollars.

Table 3. Reliabilities of NSSE Benchmarks

	First-year students	Seniors	Total
Less than .60	20%	0%	10%
.60 to .69	20%	40%	30%
.70 to .79	60%	40%	50%
.80 and greater	0%	20%	10%
N scales	5	5	10

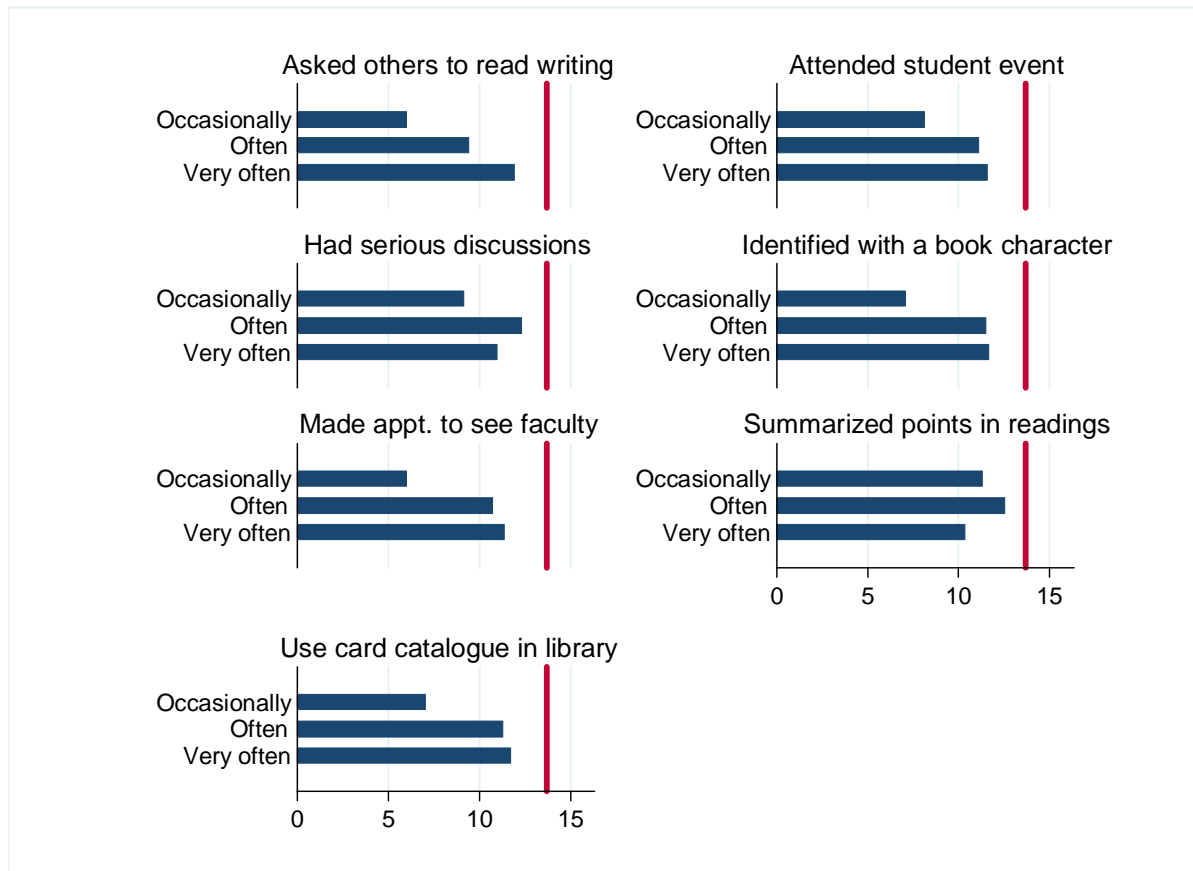
Source: NSSE (2009d)

Table 4. Distribution of Measures of Association between NSSE Scales and External Measures of Student Learning

Size of relationship	Correlations		Standardized regression coefficients		Total
	RAND	GRE	CAAP	DIT2	
-.01 to -.09	0%	0%	13%	0%	2%
0	40%	67%	63%	50%	54%
.01 to .09	20%	20%	25%	38%	24%
.10 to .19	40%	13%	0%	13%	20%
.20 and greater	<u>0%</u>	<u>0%</u>	<u>0%</u>	<u>0%</u>	<u>0%</u>
	100%	100%	100%	100%	100%
N scales	15	15	8	8	46

Source: Author's calculations from Table 2 of Carini et al. (2006) and Tables 5 and 6 of Pascarella and Seifert (2008).

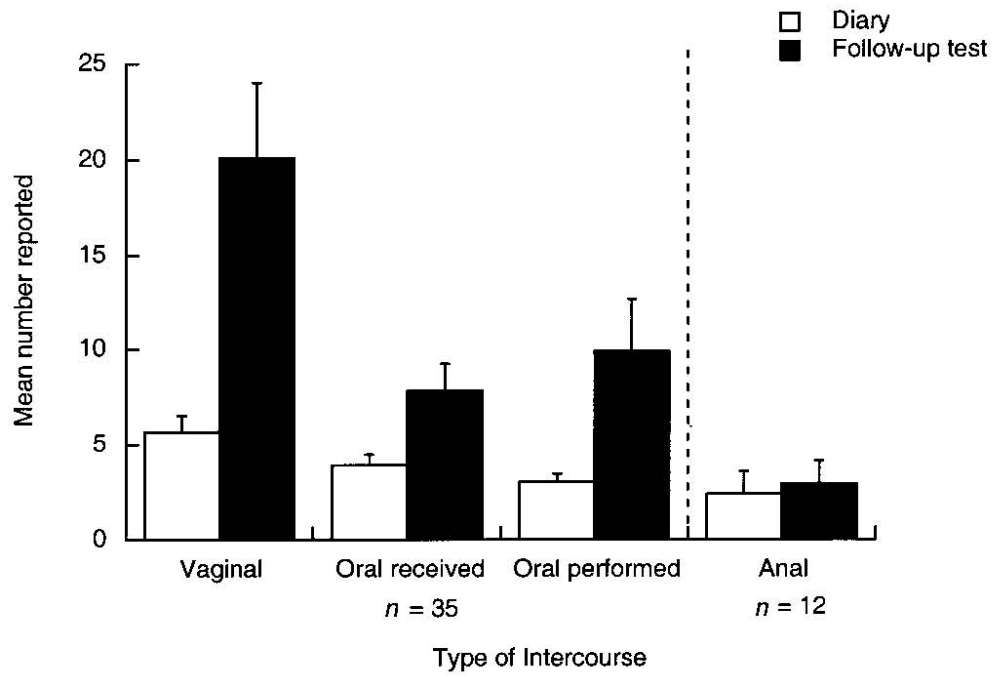
Figure 1. Standard Deviation of Numeric Responses (Converted to Number of Times per Academic Year) by Topic and Vague Quantifier



Source: Author's calculations based on Table 1 of Pace and Friedlander (1982).

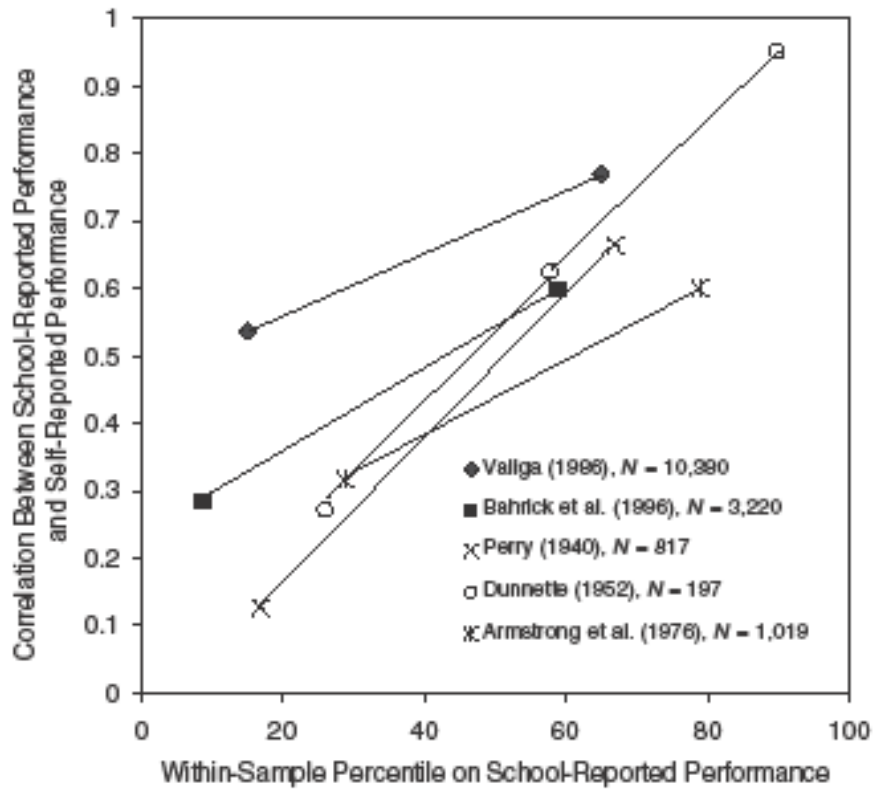
Note: Perfect agreement would result in a standard deviation equal to 0; perfect disagreement, with students equally distributed across numeric categories, would result in a standard deviation equal to 13.7, indicated by the thick vertical line. Uses the following values to determine s , based on an academic year from mid-August to mid-December and mid-January to mid-May: Never (0), Once or twice a year (1.5), 3 to 6 times a year (4.5), 1 or 2 times a month (12), About once a week (32), More than once a week (33). I use 33 instead of 64 because "more than once a week" is somewhat unclear; it could be twice week or more, or between once and twice a week. 33 is a more conservative approach that favors finding consistency of responses.

Figure 2. Frequency of Intercourse: Diary Reports and Follow-Up Survey Self-Reports



Source: Garry et al. (2002).

Figure 3. Relationship between Academic Performance and Accuracy of Self-Reports



Source: Kuncel et al. (2005).