6

This chapter examines the validity of several questions about academic challenge taken from the National Survey of Student Engagement. We compare student self-reports about the number of books assigned to the same number derived from course syllabi, finding little relationship between the two measures.

The Validity of Student Engagement Survey Questions: Can We Accurately Measure Academic Challenge?

Stephen R. Porter, Corey Rumann, and Jason Pontius

Survey data are widely used in higher education for purposes such as assessment and strategic planning. One of the most common ways of using surveys has been to assess student learning outcomes by means of proxy questions on a survey, assuming that students who engage in specific behaviors (called engagement) have learned more during college than other students.

The most well-known approach has been used by the National Survey of Student Engagement (NSSE), which asks students numerous questions about their academic experiences, such as number of times they did not prepare for class, how often they made a presentation in class, and the number of books assigned in their courses. Other national college student surveys ask similar questions. The Your First College Year Survey from the Higher Education Research Institute (HERI) at UCLA asks questions such as how often students revised papers to improve their writing, as well as questions about class preparation and presentations. The Community College Survey of Student Engagement also asks many questions about students' academic experiences.

As Porter (in press) has demonstrated, college student surveys in general, and the NSSE in particular, yield minimal evidence to support their validity. In part, this is due to the difficulty of collecting external data to use for validation studies; we usually ask survey questions to obtain data that are not easily available elsewhere. Many of the academic experience

New Directions for Institutional Research, no. 150, Summer 2011 © Wiley Periodicals, Inc.
Published online in Wiley Online Library (wileyonlinelibrary.com) • DOI: 10.1002/ir.391

questions used on higher education surveys ask about experiences that, in theory, could be tracked and measured (such as the number of class presentations) but that in practice are difficult to measure.

We developed a research design to investigate the validity of academic experience questions using students' course syllabi. Unlike questions about class preparation or other experiences, questions about the number of books and papers assigned to students (referred to by the NSSE as "academic challenge") do have a relatively accessible source of data to verify their responses, namely, course syllabi. By asking students about academic challenge, and then collecting syllabi for the courses they are taking at the time of the survey, we can compare student self-reports to the same information derived from course syllabi.

Survey Response and Academic Challenge Questions

The top panel of Table 6.1 shows the academic challenge questions from the 2011 version of the NSSE. The survey asks students to report the number of books and papers assigned in their classes for an entire academic year (that is, two semesters of classes) and further asks them to break down the number of papers by page length. The survey also asks students how many problem sets they are assigned during a typical week, broken down by whether the problem sets take more or less than an hour to complete.

The four-stage model of survey response developed by Tourangeau, Rips, and Rasinski (2000) is widely accepted in the field of survey methodology, and we use it to highlight several reasons the academic challenge questions asked on the NSSE and other surveys likely lack validity. He and his colleagues posit four components of the survey response process, not all of which are necessarily used by every respondent. First, the respondent must comprehend the survey question, not only in terms of understanding key terms but also in understanding exactly what information is being sought by the survey researcher. For factual questions, such as the ones under study here, the respondent must retrieve the requested information from his or her long-term memory. Once the relevant memories have been retrieved, the information from the memories must be combined or supplemented with additional information to form a judgment as to what the answer to the question is. Finally, respondents must report a response by mapping their answer to the response scale on the survey.

Comprehension. There are two reasons college students might be unable to fully comprehend questions about academic challenge. The first is the vagueness of the words in the questions. It is not entirely clear, for example, what is meant by a "written paper or report" or "problem set." As we explain below, the research team had trouble coding research papers and could not even agree on a definition of a problem set. If a faculty member and two doctoral students in a higher education administration

Table 6.1. Question Wording of Academic Challenge Questions

2011 National Survey of Student Engagement:						
During the current school year, about how much reading and writing have you done?	*None*	*1–4*	*5–10*	*11–20*	*More than 20*	
Number of assigned textbooks, books, or book-length packs of course readings	○	○	○	○	○	
Number of books read on your own (not assigned) for personal enjoyment or academic enrichment	○	○	○	○	○	
Number of written papers or reports of 20 pages or more	○	○	○	○	○	
Number of written papers or reports between 5 and 19 pages	○	○	○	○	○	
Number of written papers or reports of fewer than 5 pages	○	○	○	○	○	
In a typical week, how many homework problem sets do you complete?	*None*	*1–2*	*3–4*	*5–6*	*More than 6*	
Number of problem sets that take you more than an hour to complete	○	○	○	○	○	
Number of problem sets that take you less than an hour to complete	○	○	○	○	○	
Validity Study Survey:						
During the current semester, about how much reading and writing have you done?	*None*	*1–2*	*3–4*	*5–6*	*7–8*	*More than 8*
Number of assigned textbooks, books, or book-length packs of course readings	○	○	○	○	○	○
Number of written papers or reports of 20 pages or more	○	○	○	○	○	○
Number of written papers or reports between 5 and 19 pages	○	○	○	○	○	○
Number of written papers or reports of fewer than 5 pages	○	○	○	○	○	○
In a typical week, how many homework problem sets do you complete?	*None*	*1–2*	*3–4*	*5–6*	*More than 6*	
Number of problem sets that take you more than an hour to complete	○	○	○	○	○	
Number of problem sets that take you less than an hour to complete	○	○	○	○	○	

program could not agree on these terms, it is unlikely that undergraduates across the country are using the same definitions in their minds when they read and respond to these questions.

The second problem with these questions is that they reflect a traditional view of postsecondary teaching and learning more suitable to the midtwentieth century than the twenty-first century. Consider the phrase "book-length packs of course readings." It is now common for faculty to post course material on course websites; one of the authors of this chapter has abandoned course packs completely and now supplies articles and chapters by way of an online course program. Do students think of these collections of online readings as a course pack? The same can be said for courses, especially online courses, which require periodic posting of online written responses. Twenty years ago, these postings would instead have been short papers handed into the professor at the beginning of class. Some students may be counting these when they respond to survey questions, while others may not. As technology changes, pedagogy changes as well, and the types of academic challenge questions asked on college student surveys must also change.

Recall. Even if all college students shared the same understanding of these survey questions, it is unlikely they could answer them reliably. On the NSSE, the questions ask students about their current academic year. Assuming the average student takes three to four courses per semester, this would require students to recall from memory information about six to eight courses. They would need to remember not only the number of books for each course (half of which, given the timing of the survey, were purchased at least six months previously) but the number of papers assigned, the page length of each assigned paper in every course, and the number of problem sets in each course and how long it took to complete every single problem set.

If we consider how much we are asking students to recall, and then think about our own memory and our ability to retrieve information that seems relatively insignificant (from the point of view of the student), it is difficult to believe that most students would be able to achieve this level of recall. Of course, students may be able to accurately recall *some* information about books or papers in their classes. If we asked students how much they paid for their books last semester (highly salient given the costs of college, and given that students usually purchase the books themselves, unlike tuition, which is often paid by their family), or whether they had to stay up all night to complete a paper (a relatively infrequent, unusual event for most students), many students would likely be able to recall this information. Yet academic challenge questions ask students for information that is neither salient nor unusual, so it is unlikely to be retained in memory.

Recall becomes more complicated when the respondents are students who have been in college for more than one year. Questions that ask about

experiences during a specific time period require respondents to not only retrieve memories but also to assign dates to their recollections. If respondents are uncertain about memories and time periods, events and experiences previous to the time period may be recalled and mistakenly assigned to the survey question time period (called forward telescoping; Tourangeau, Rips, and Rasinski, 2000). Unfortunately, research suggests that reports of frequency are correlated with the number of memories recalled, rather than the actual frequency (Bradburn, Rips, and Shevell, 1987). Given that seniors have several years of college memories from which to draw, telescoping is difficult to rule out, especially for the mundane information that academic challenge questions seek.

Judgment. Once the relevant memories are fully or partially retrieved, the respondent must use the information from memory and derive an answer to the survey question. In some cases the answer is quite simple; most college students have little trouble supplying the correct response for a question about their age or gender. The judgment component for academic challenge questions is more complicated.

First, because these questions ask about several courses, requiring the respondent to consider several numbers (for example, number of papers and page length), the cognitive ability to correctly combine them is not inconsequential. Second, from the discussion about recall, it is likely that only a few relevant memories will be retrieved. This leads to respondents using an estimation strategy to create an answer. Such an estimation strategy will not necessarily lead to an incorrect answer. If you are asked how often you visit the dentist and are unable to remember your dental visits (or assign dates to them), you might still be able to give a correct answer by remembering that you always follow your dentist's advice and schedule two visits a year.

The problem with many estimation strategies is that they do not always yield the correct answer. Respondents may have few memories on a topic, and thus infer that the topic happened infrequently. Or, as has been amply demonstrated in the literature (for example, Smyth, Dillman, and Christian, 2007), they may instead look at the response scale of a question and infer that the middle option is the "average" response and pick a response in relation to where they perceive themselves to be vis-à-vis the average person. In general, the more difficult it is to directly recall and tally information from memory, the more likely it is that the respondent will use an error-prone estimation strategy.

Response. Once the respondent has derived an answer, he or she must map the answer to the response scale given on the survey. Here, the academic challenge questions on the NSSE are well written in comparison to other NSSE questions and other college student surveys. Many researchers use a response scale of vague quantifiers, such as "often," "very often," etc. College students' understanding of the meaning of these terms varies widely in terms of how frequently an event must occur to be given the

label "often" (Porter, in press). In terms of the four-part model of survey response, academic challenge questions are much more problematic in terms of comprehension, recall, and judgment than response.

Methodology

To compare students' actual academic experiences with their reported experiences, we collected two sets of data. First, we surveyed a random sample of three thousand seniors at Iowa State University in November 2007. Second, we contacted their instructors and asked for a copy of their syllabi, which we then coded in terms of number of books, papers, and problem sets assigned.

Data Collection. On the web survey, we asked questions about academic challenge similar to the NSSE, but with one important difference (see Table 6.1). The NSSE asks students to report on reading and writing in all of their classes for an entire academic year (usually sometime in late spring). Given what we know about human cognition and the psychology of survey response, we believed that asking a similar question would undoubtedly result in a nonexistent relationship between reported and actual experiences. Instead, we altered the question and asked students to report on reading and writing during the current semester. We also reduced the distribution of the response scale from "none" to "more than 20" to the more limited range of "none" to "more than 8." Because we were altering the question wording, we also included an additional category so that each position on the response scale between "none" and "more than 8" consisted of two books or papers (1–2, 3–4, and so on). The problem set questions remained unchanged. The response rate to the student web survey was 31 percent, similar to many other college student web surveys. In 2010, the response rate for seniors taking the web version of the NSSE at institutions with a Carnegie Classification of Research Universities (very high research activity) was 30 percent (National Survey of Student Engagement, 2011). (Iowa State University, which was the location of the study, has the same classification.)

We note that by shortening the time period for the question, we are making it easier to validate the question wording of the academic challenge questions. If we find little relationship between reported and actual academic challenge data when asking about a semester's worth of classes, it is doubtful that we will find greater agreement in asking about academic challenge over an entire academic year, as the NSSE does.

Second, we used transcript data for students in our sample and contacted instructors of their classes at the same time as the student survey, with two email contacts. At the close of the student survey, we made additional follow-up efforts for instructors of students who had responded and who lacked only one syllabus for their classes. Finally, we searched the Iowa State website for syllabi that might have been posted by faculty.

Forty-one percent of faculty furnished the requested syllabi through this approach. This rate is in line with response rates of national faculty web surveys. The Faculty Survey of Student Engagement, for example, reports a response rate of 38 percent for Research Universities (very high research activity; Faculty Survey of Student Engagement, 2011).

Given that our initial sample of students was 3,000, and we ended up with 925 respondents, it may seem somewhat strange that we have only 42 seniors in our analytic dataset. However, students take several courses, and we received syllabi for fewer than half of the courses seniors took. Given the response rate listed above, we can calculate the probability that an individual student in our random sample of 3,000 seniors has complete data, that is, responded to the student survey and also had syllabi for all of their classes in fall 2007. Assume a senior took four classes. The probability of complete data is

$$P(\textit{complete data}) = .31 \cdot .41 \cdot .41 \cdot .41 \cdot .41 = .0088$$

Taking our student sample size of 3,000 and multiplying it by .0088 yields 26 students. Our analytic sample size is larger because the average number of courses taken was not exactly 4.0, and because of our extra efforts to obtain syllabi for students who were missing only one syllabus for their set of fall classes.

Such a small sample size raises questions about the representativeness of our data. As we noted above, our student and faculty participation rates are in line with other higher education research projects. In addition, Groves (2006; Groves and Peytcheva, 2008) has convincingly demonstrated that there is little relationship between survey response rate and individual survey question bias: bias can be high for surveys with a high response rate, and vice versa. Thus there is no reason to expect our analytic sample to necessarily differ significantly from the respondent sample, unless only certain types of faculty were more or less likely to respond to our syllabi request.

Table 6.2 shows some descriptive statistics for the population of seniors, the student survey respondent sample, and the analytic sample (seniors for whom we have syllabi for all of their courses). The mean ACT score for the analytic sample is similar to the population, while the grade point average (actual, not self-reported) is somewhat lower at 2.78 versus 2.93. We attribute this to the exclusion of seniors from our analytic sample who took an independent study course. These courses lack syllabi and are generally taken by academically successful students. The distribution by college also appears fairly similar to the population.

The most important characteristic of the analytic sample is how representative it is in terms of accurate survey responses. If seniors for whom we have complete syllabi data are much better or worse reporters than the population, then this would call into question the external validity of our results. We do not have these data for the population, but we did

Table 6.2. Student Population and Sample Descriptives

	All Seniors (Population)	*Survey Respondents*	*Analytic Sample*
ACT: actual	24.6	25.6	25.1
ACT: reported		26.4	25.6
Percentage correct		68%	62%
Grade point average	2.93	3.01	2.78
Major college			
Agriculture and life sciences	13%	12%	21%
Design	9%	8%	2%
Engineering	22%	22%	2%
Human sciences	15%	13%	24%
Business	17%	16%	14%
Liberal arts and sciences	25%	30%	36%
	100%	100%	100%
N	6,276	925	42

ask students on the survey to report their ACT score and are thus able to compare this to their actual ACT score as listed in the university database. Comparing the percentage reporting correct scores for the respondent and analytic samples, these percentages are similar: 68 percent of the 925 survey respondents were able to correctly report their ACT score, while 62 percent of the 42 respondents for whom we have complete syllabi data were able to do so. The difference between actual and reported ACT scores was also smaller for the analytic sample, about .5 of a point, versus .8 for the respondent sample.

Coding of Syllabi. After data collection, we randomly selected thirty syllabi and developed a coding scheme for how to count the number of books, papers, and problem sets assigned. We then applied this to several additional syllabi, discussing how we coded and made additional adjustments. Finally, two of the authors separately coded fifty syllabi without consultation, and we calculated interrater reliability statistics.

By far the easiest type of academic challenge data to code was the number of books or book-length packs of course readings assigned. The two sets of numbers matched exactly for 89 percent of the test syllabi, with $r = .95$. The interrater reliability was much lower for the number of written papers or reports assigned, with 65 percent agreement and $r = .39$. Surprisingly, it was difficult to determine the number of papers assigned from a syllabus. For example, some classes had daily in-class writing assignments. Should these be counted as an assigned paper, or only assignments completed outside of class? One class had students develop a résumé as an assignment. This is a written work completed outside of class, but should it be considered a paper or report? Some syllabi also listed papers, but not their required page length. Because the interrater reliability statistics were so low, we decided these data were not reliable enough to analyze.

By far the greatest problems we had involved coding problem sets. After much discussion, we abandoned this in the early stages of the research project. Despite our best efforts, after reviewing several syllabi we could not come up with a good definition for a problem set. For math classes this was relatively easy, but it became less so for classes in the sciences and other fields.

Finally, after coding the remaining syllabi, we reviewed the few syllabi where the coders differed on number of books assigned, discussing these syllabi until we reached agreement as to how many books had been assigned for that particular course.

Analysis. For each student we created a count of the number of books assigned by summing the syllabi numbers for each student, using student transcript data to link syllabi data to student data. To compare the actual and reported number of books assigned, we collapsed the actual number of books assigned to match the six response categories on the student survey. Table 6.3 shows the distribution of responses about number of books assigned from the student survey, as well as the distribution of actual books assigned, using syllabi data for each student. As can be seen, the two distributions are fairly similar, with similar means (3.6 books for the survey and 4.1 books for the syllabi). However, these simple distributions can be misleading, as shown in Table 6.4.

For each student we have two measures of books assigned, one self-reported and one derived from coding the syllabi of the classes in which the student was registered. Table 6.4 shows the cross-tabulation of these two sets of responses for the forty-two seniors for whom we have complete data. The bold diagonal highlights the cells that are correct responses—that is, the self-report matches the number of books from the syllabi. Responses below the diagonal are undercounts; for example, 2 percent of the students had five or six books assigned but reported none. Responses above the diagonal are overcounts, where students reported having more books assigned than they actually did.

As can be seen, there is little agreement between the two sets of data, even bearing in mind that the categories are such that a student could be

Table 6.3. Number of Books Assigned: Distributions

	Student Survey	*Syllabi Database*
None	4.8	0.0
1–2	23.8	16.7
3–4	23.8	16.7
5–6	21.4	33.3
7–8	9.5	11.9
More than 8	16.7	21.4
	100.0	100.0
Mean	3.6	4.1

Table 6.4. Relationship Between Student Responses and Syllabi Data for Number of Books Assigned

	Student Survey Responses					
Syllabi Data	*None*	*1–2*	*3–4*	*5–6*	*7–8*	*More Than 8*
None	**0**	**0**	0	0	0	0
1–2	0	**12**	5	0	0	0
3–4	0	5	**2**	2	0	7
5–6	2	4	14	**5**	2	5
7–8	2	2	0	5	**0**	2
More than 8	0	0	2	10	7	**2**

Note: cell entries are percentages and do not sum to 100 because of rounding.

Correlation	.38
Percentage correct	**21**
Percentage underreporting	55
Percentage overreporting	24

off by one but still give the correct answer. The correlation between the two measures is only .38, and looking at the bold diagonal we see that only 21 percent of students furnished a correct answer. To put this number in perspective, there are six categories in the response scale, so if students randomly chose a category they would be correct approximately 17 percent of the time (assuming a uniform distribution). Roughly two-thirds of the errors were underreports and about one-third were overreports.

Conclusion

The results presented here suggest that asking college students about their academic experiences is a difficult task. From the perspective of higher education researchers, we require detailed information about important aspects of the college experience: frequency of meetings with faculty and study groups, time spent studying, and the rigor of courses as measured by number of books, papers, and problem sets assigned. However, because these are frequent or mundane activities from the students' perspective, these experiences are probably not encoded in their memory, or if they are, only partially. Considering survey questions that use vague wording such as "problem sets," and the difficulty of correctly combining several numbers into specific answers about assignments and their length, cognitive theory predicts that most students are unable to correctly answer these types of questions.

Using a unique database of student transcripts and course syllabi, we were able to create an alternative measure of the number of books and

course packs assigned, and we compared this to self-reports of the number of books and course packs assigned. Even bearing in mind that there is some error in our syllabi measure, we found little correspondence between self-reported books and actual number of assigned books. Only 21 percent of students gave a correct answer, and the correlation between the two measures, which should be high even if students are off by only one or two books, is just .38. In addition, the two coders were unable to independently and reliably code the number of papers assigned, and the research team could not agree on a definition of a problem set to allow coding for the number of problem sets. This lack of agreement among trained coders who have expertise in the field implies that the variation in students' definitions of these items may be quite large. Together, these results suggest that students probably do not understand what is being asked of them in the area of academic challenge.

The results presented here also suggest that relying on simple correlations between items and scales as a measure of validity may be problematic. In his discussion of the psychometric properties of the NSSE, Kuh (2004) states:

> Patterns of correlations among these items are consistent with what one would expect. For example, the item related to the number of hours spent preparing for class is positively related to several questions surrounding academic rigor such as the number of assigned course readings (.25) [p. 9].

Using our analytic sample, the correlation between hours spent preparing for class and number of assigned books (self-report) is similar, .21, and the correlation with hours spent relaxing and socializing is –.31. Yet these correlations, which Kuh (2004) uses as evidence of the validity of the NSSE items, are found using a self-reported measure of assigned books that is only moderately correlated with the actual number of books assigned. Clearly, it is not difficult to find moderate correlations among items on the same survey questionnaire, even if some of these items bear little resemblance to real-world data.

Finally, we recommend that efforts to collect these types of student data should be expanded beyond the typical cross-sectional student survey. Periodic surveys throughout the school year asking about student experiences during the previous week should yield better data; respondents generally are better able to report on recent, short time spans instead of long, distant time spans. Other possibilities for academic challenge data include university bookstore databases and the wealth of data about class activities that can be collected from courses using an online component. Although cross-sectional surveys are cheap and easy to administer, they yield data of low value. Only by expending greater resources on data collection can we hope to gather college student data with better reliability and validity.

References

Bradburn, N. M., Rips, L. J., and Shevell, S. K. "Answering Autobiographical Questions: The Impact of Memory and Inference on Surveys." *Science,* 1987, *236*(4798), 157–161.

Faculty Survey of Student Engagement. *FSSE 2010 and Overall Response Rate Averages by Carnegie Classification*. 2011. Retrieved January 24, 2011, from http://fsse.iub.edu/pdf/FSSE_IR_2010/FSSE10%20%20Overall%20-%20RR%20x%20Carnegie%20Class%20Chart.pdf.

Groves, R. M. "Nonresponse Rates and Nonresponse Bias in Household Surveys." *Public Opinion Quarterly*, *2006, 70*(5), 646–675.

Groves, R. M., and Peytcheva, E. "The Impact of Nonresponse Rates on Nonresponse Bias: A Meta-Analysis." *Public Opinion Quarterly*, 2008, *72*(2), 167–189.

Kuh, G. D. *The National Survey of Student Engagement: Conceptual Framework and Overview of Psychometric Properties*. Bloomington: Indiana University Center for Postsecondary Research and Planning, 2004.

National Survey of Student Engagement. *NSSE 2010 Response Rate Information*. 2011. Retrieved January 24, 2011, from http://nsse.iub.edu/NSSE_2010_Results/ResponseRates.cfm.

Porter, S. "Do College Student Surveys Have Any Validity?" *Review of Higher Education* (in press).

Smyth, J. D., Dillman, D. A., Christian, L. M. "Context Effects in Web Surveys: New Issues and Evidence." In A. Joinson, K. McKenna, T. Postmes, and U.-D. Reips (eds.) *Oxford Handbook of Internet Psychology*. New York: Oxford University Press, 2007.

Tourangeau, R., Rips, L., and Rasinski, K. *The Psychology of Survey Response*. Cambridge, UK: Cambridge University Press, 2000.

STEPHEN R. PORTER is associate professor of research and evaluation and higher education in the Department of Educational Leadership and Policy Studies at Iowa State University.

COREY RUMANN is an assistant professor in the Department of Counseling and Educational Psychology and director of the College Student Affairs program at University of West Georgia.

JASON PONTIUS is the coordinator of Continuous Academic Program Improvement at Iowa State University.