Self-Reported Learning Gains: A Theory and Test of College Student Survey Response

Stephen R. Porter

Abstract Recent studies have asserted that self-reported learning gains (SRLG) are valid measures of learning, because gains in specific content areas vary across academic disciplines as theoretically predicted. In contrast, other studies find no relationship between actual and self-reported gains in learning, calling into question the validity of SRLG. I reconcile these two divergent sets of literature by proposing a theory of college student survey response that relies on the belief-sampling model of attitude formation. This theoretical approach demonstrates how students can easily construct answers to SRLG questions that will result in theoretically consistent differences in gains across academic majors, while at the same time lacking the cognitive ability to accurately report their actual learning gains. Four predictions from the theory are tested, using data from the 2006–2009 Wabash National Study. Contrary to previous research, I find little evidence as to the construct and criterion validity of SRLG questions.

Keywords College students · Learning gains · Survey research · Validity

There is currently a vigorous debate over the validity of college student survey questions (Bowman 2010a; Campbell and Cabrera 2011; Ewell et al. 2011; McCormick and McClenney 2012; Porter 2011a). Critics have asserted a lack of content, construct, and criterion validity for college student survey questions in general, and for self-reported learning gains (SRLG) questions in particular. Of the numerous survey questions asked of college students, SRLG questions are clearly the most important, because student learning is at the very heart of the higher education enterprise. Thus, for both practitioners and scholars, the fundamental question is *can we measure learning by simply asking students how much they have learned*?

Looking across the higher education landscape, the implicit answer to this question appears to be positive. SRLG questions have been used extensively as dependent variables

S. R. Porter (✉)
Department of Leadership, Policy, and Adult and Higher Education, North Carolina State University, Box 7801, Raleigh, NC 27695, USA
e-mail: srporter@ncsu.edu

in higher education research (e.g., Kuh and Vesper 2001; Lambert et al. 2007; McCormick et al., 2009; Pike 2000; Zhao and Kuh 2004), and they remain on the revised version of the National Survey of Student Engagement that was just released. Recently, several scholars have asserted that SRLG are indeed valid measures of learning, showing that SRLG vary as theoretically predicted across academic major groupings, e.g. artistic majors report larger gains in artistic learning outcomes than students in other majors (Pike, 2011; Pike et al. 2011b). This is in stark contrast to research arguing that college students lack the cognitive ability to accurately report their learning gains while in college, and empirical findings of almost no relationship between self-reports and objective measures of learning (Bowman 2010a, b, 2011b; Porter 2011a).

The purpose of this paper is threefold. First, I seek to reconcile these two divergent sets of literature. If students do not have the cognitive ability to report how much they have learned in college, and there is no relationship between objective and subjective measures of learning gains, then what explains the robust finding that subjective measures of learning gains vary across academic majors as we would predict? Second, advocates for the validity of college student survey questions argue that if the critics are correct, and students lack the cognitive ability to accurately answer most survey questions, then the critics are, in essence, arguing that students must be generating random responses to survey questions (McCormick and McClenney 2012). Using the commonly accepted theory of attitude formation from the field of public opinion research, I develop a model of college student survey response for SRLG questions. This theoretical approach shows how students can easily construct answers to SRLG questions that will result in theoretically consistent differences across academic majors, while at the same time lacking the cognitive ability to accurately report their actual learning gains. Third, I test hypotheses derived from this model, using both SRLG questions and objective measures of student learning. Contrary to previous research, I find little evidence as to the construct and criterion validity of SRLG questions.

Literature Review

Any validity study must take first into account the purpose of the survey items being validated, because whether a survey question can be considered valid depends on how it will be used (American Educational Research Association et al. 1999). In general, SRLG questions have two purposes, one applied and one scholarly. First, these questions are used to provide information to practitioners about the state of learning on their campuses. For example, the most commonly used college student survey, the National Survey of Student Engagement (NSSE), provides institutions with point estimates for the SRLG questions on their instrument (Table 1 shows hows these questions are worded). In addition, schools are provided averages from similar institution types, as well as national averages, so that they can understand how much they differ from other schools (National Survey of Student Engagement 2012). Second, academic researchers use these questions in multivariate models to understand how gains in learning relate to other constructs, such as student engagement (see e.g., Laird et al. 2008; Pike et al. 2011a, 2012; Smart 2010). For both purposes, accurate self-reports of learning gains are vital; that is, self-reported gains should closely mirror actual gains. The entire premise of using SRLG questions is that they serve as excellent proxies for actual learning gains, obviating the need to measure student learning at entry and exit with multiple subject area tests (e.g., critical thinking, quantitative skills, writing skills, speaking skills, etc.). If these two sets of measures are not

Table 1 Wording of SRLG questions

To what extent has your experience at this institution contributed to your knowledge, skills, and personal development in the following areas?	Very much	Quite a bit	Some	Very little
Acquiring job or work-related knowledge and skills	☐	☐	☐	☐
Writing clearly and effectively	☐	☐	☐	☐
Speaking clearly and effectively	☐	☐	☐	☐
Thinking critically and analytically	☐	☐	☐	☐
Analyzing quantitative problems	☐	☐	☐	☐
Working effectively with others	☐	☐	☐	☐
Contributing to the welfare of your community	☐	☐	☐	☐
Understanding people of other racial and ethnic backgrounds	☐	☐	☐	☐
Understanding yourself	☐	☐	☐	☐

Source National Survey of Student Engagement, 2000–2012

highly correlated, it is not at all clear how we can use self-reports as proxies for actual learning when assessing institutional performance. Moreover, if students are misreporting their learning gains, and if the causes of this misreporting are not constant across institutions (e.g., due to student characteristics that vary across institutions, such as academic ability, and cultural and social capital), then any estimates or benchmarks will be misleading to institutional leaders.

If there is a low correlation between the two measures, then this begs the question of how students are constructing a response to SRLGs. If actual learning is not driving responses, and students are not randomly choosing answers to the questions, then other factors *must* be driving responses. Because these other factors may not be uniformly distributed across institutions and across student subgroups within an institution, any multivariate analysis trying to show relationships between school-level or student-level variables may be flawed, as these variables will be picking up the effects of these other factors driving student responses (see e.g., Astin and Lee 2003; Bowman 2011a; Pascarella and Padgett 2011). In sum, it is difficult to conceive of SRLG as valid measures of actual learning gains if (1) they are not highly correlated with actual learning gains and (2) factors other than actual learning drive student responses to these questions.

I have argued elsewhere that any validity argument for college student survey questions must provide both a theoretical model of college student cognition, as well as empirical evidence in support of validity (Porter 2011a). In the next two sections I review the theory and evidence for and against the validity of SRLG questions.

Arguments for Validity

Despite their widespread use in higher education, proponents of SRLG questions have not articulated a theory of cognition that explains how students are able to accurately answer these questions. Instead, proponents generally cite research showing that these questions vary across academic major groupings as one would expect.

Using SRLG questions from the College Student Experiences Questionnaire, Pace (1985) finds learning gains across majors that make intuitive sense. He finds that 92 % of

arts majors reported substantial gains in "developing an understanding and enjoyment of art, music, and drama," while the average percentage for all students was only 29 %. For "understanding the nature of science and experimentation," 85 % of biological sciences majors and 76 % of physical sciences majors reported substantial gains, compared to 36 % for all students. Using the same instrument, Pike and Killian (2001) found mixed results for differences in learning gains across Biglan (1973a, b) categories of pure versus applied majors, in which majors are classified into two categories based on the extent to which their disciplines emphasize application of knowledge. As expected, they find that students in applied disciplines had greater gains in vocational competence, but contrary to what Biglan's approach would predict, these students had lower general education gains compared to students in pure majors.

Perhaps the strongest empirical evidence supporting the validity of SRLG questions are two recent studies using gains questions from the NSSE (Pike 2011; Pike et al. 2011b); hereinafter, Pike et al. Using Holland's (1973) theory of person-environment fit, Pike et al. conclude that these questions have both construct and criterion validity. Holland proposes that individuals and environments can be classified into one or more of six types (Realistic, Investigative, Artistic, Social, Enterprising and Conventional), based on what members of these environments prefer, and how the environments in turn socialize people who enter a particular environment. Briefly, Realistic environments emphasize practical activities and include majors such as materials science and mechanical engineering. Investigative environments emphasize intellectual activities focused on knowledge and include majors such as the physical sciences and mathematics. Artistic environments emphasize unsystematized activities and include majors such as art and drama. Social environments emphasize manipulation of others to inform and enlighten them, and include majors such as elementary education and social work. Enterprising environments emphasize manipulation of others for economic and organizational gains, and include majors such as journalism and business administration. Finally, Conventional environments emphasize manipulation of data, and include majors such as accounting.

Building on work that shows students seek out majors that match their Holland type, and that Holland environments socialize students in different ways (e.g., Artistic environments emphasize Artistic endeavors and reward students for engaging in these endeavors) (Smart et al. 2000), Pike et al. make two main arguments. First, SRLG items from the NSSE should load onto four different factors matching Holland's categories of Investigative, Artistic, Social and Enterprising. For example, Investigative environments emphasize "analytical or intellectual activity aimed at trouble-shooting or creation and use of knowledge" (Gottfredson and Holland 1996), so the two items measuring gains in analyzing quantitative problems and thinking critically and analytically should both load onto the same factor. Because first-year students have not spent enough time in college to be socialized within a discipline, they analyze only data for seniors, who are surveyed near the end of their senior year, and find that the items load onto four factors as theoretically predicted (see top two panels of Table 2).

Second, they argue that due to the socialization process of academic disciplines, gains across the Holland major categories should vary by Holland environment. For example, students in Investigative majors such as biology, mathematics, and physics should report greater gains on the Investigative outcome factor than students majoring in Artistic, Social or Enterprising disciplines, while Artistic majors should report larger gains on the Artistic gains factor compared to students in the other Holland environments. Using seniors majoring in one of four Holland environments (Investigative, Artistic, Social and Enterprising), they find that the amount of gains vary as theoretically predicted. Their results are

Table 2 Factor analysis of NSSE SRLG questions

	Investigative	Artistic	Social	Enterprising
(a) *Pike (2011): End of senior year[a]*				
Analyzing quantitative problems	.86			
Thinking critically and analytically	.52	.49		
Writing clearly and effectively		.96		
Speaking clearly and effectively		.61		.34
Understanding yourself			.81	
Understanding people of other racial backgrounds			.79	
Contributing to the welfare of your community			.75	
Working effectively with others				.85
Acquiring work-related knowledge and skills				.67
(b) *Pike et al. (2011b): End of senior year[b]*				
Analyzing quantitative problems	.79			
Thinking critically and analytically	.69	.33		
Writing clearly and effectively		.96		
Speaking clearly and effectively		.57		.36
Understanding yourself			.91	
Understanding people of other racial backgrounds			.76	
Contributing to the welfare of your community			.78	
Working effectively with others				.81
Acquiring work-related knowledge and skills				.73
(c) *Wabash: end of freshman year*				
Analyzing quantitative problems	.93			
Thinking critically and analytically	.58	.59		
Writing clearly and effectively		.87		
Speaking clearly and effectively		.81		
Understanding yourself			.76	
Understanding people of other racial backgrounds			.76	
Contributing to the welfare of your community			.63	
Working effectively with others		.37	.43	.35
Acquiring work-related knowledge and skills				.94
(d) *Wabash: end of senior year*				
Analyzing quantitative problems	.94			
Thinking critically and analytically	.48	.64		
Writing clearly and effectively		.88		
Speaking clearly and effectively		.82		
Understanding yourself			.70	
Understanding people of other racial backgrounds			.81	
Contributing to the welfare of your community			.66	.38
Working effectively with others		.35		.57
Acquiring work-related knowledge and skills				.87

Note Only factor loadings >.30 are shown

[a] Table 3

[b] Table 2

reported in the top two panels of Table 4. These coefficients are taken from models using dummy variables to indicate the Holland environment of a student's major, with the requisite Holland category serving as the reference category. For example, the first column of numbers in the first panel shows that Artistic majors report Investigative learning gains about $\frac{1}{3}$ of a standard deviation less than Investigative majors, while Social and Enterprising majors report $\frac{1}{4}$ and $\frac{1}{5}$ of a SD less gains, respectively. Looking across the top two panels, we can see that all of the differences are statistically significant, and negative, as theoretically predicted. Students clearly appear to be reporting learning gains across majors as predicted.

However, there are two problems with these studies. First, the models only include controls for gender, race/ethnicity, on-campus housing, first-generation college student status, whether the student is a transfer, and age. There are no controls for pre-college interest in academic disciplines. Such controls are crucial, as scholars using Holland's theory to study college students have argued that

> Self-selection is thus an important consideration in longitudinal efforts to study college outcomes and the extent to which patterns of student change and stability vary across disparate educational environments. This is so because the different academic environments (college majors) initially attract students with different interests and talents. Longitudinal studies of how academic environments contribute to differential patterns of change and stability in college students must take into account this "self-selection" of students to get a more accurate assessment of the actual influence of those environments on students. (Smart et al. 2000, p. 52)

Several studies have demonstrated that students who report stronger abilities and interest in a particular Holland environment tend to choose an academic major that matches that environment (Huang and Healy 1997; Porter and Umbach 2006; Smart et al. 2000). Without controls for these pre-college differences, estimates of the effect of Holland environments will be positively biased.

Second, it is not clear how large a difference has to exist between Holland major types in order to support the hypothesized differences derived from Holland's theory. The sample sizes in Pike (2011) and Pike et al. (2011b) are 20,000 students, so it is not surprising that all of the coefficients for the Holland major dummy variables are statistically significant. Given such a large sample size, the focus should be on substantive significance, not statistical significance. One approach is to use the typical effect size found in randomized interventions employed in primary and secondary education (Porter 2011b). These effect sizes typically range from .20 to .30, and it is common to use .20 when calculating power for a K-12 randomized trial. Given the arguments made for the strong socializing influence of Holland environments on college students (Smart et al. 2000), it is not unreasonable to assume that the effects of academic disciplines on learning gains after four to six years of college should at a minimum be as large as the average effect of a K-12 intervention that is typically implemented during a single year.

An effect size of .20 should be considered a conservative benchmark, given what we know of the growth in student learning over time. Because SLRG are used to measure growth in learning at the end of the college career, a reasonable benchmark is the typical academic growth we would expect to see during the same time period. Analyses of K-12 standardized tests suggest a gain of .44 SD for reading, .40 for mathematics, and .38 for science from the 9th grade to the 12th grade (Bloom et al. 2008). Analyses of gains in critical thinking using the Collegiate Learning Assessment and the Collegiate Assessment

of Academic Proficiency (CAAP) demonstrate .47 and .44 SD growth, respectively, from college entry to the end of the senior year (Pascarella et al. 2011). These studies suggest that .40 is a more appropriate effect size benchmark for learning growth.

With these benchmarks in mind, the results of Pike et al. shown in the top two panels of Table 4 are not as compelling as they might first appear. While all of the coefficients are statistically significant, and in the hypothesized direction, only 10 out of 24, or 42 %, are larger than .20. None are larger than .40.

Arguments Against Validity

One of the major problems with SRLGs is that no one has yet posited a credible theory as to how students can accurately report how much they have learned in college, either generally or in specific content areas such as critical thinking, analyzing quantitative problems, and writing. Porter (2011a) and Bowman (2010a) have argued that students simply lack the cognitive ability to produce this information.

Table 1 shows SRLG questions from the National Survey of Student Engagement, the survey used by Pike et al. in their validation studies.[1] A similar set of questions appears on the College Senior Survey, produced by the Higher Education Research Institute (2012), and SRLG questions appear on many institutional and consortia surveys, such as the senior survey that the Higher Education Data Sharing Consortium (2012a) uses for institutional analyses.

The current approach to survey and human cognition posits four steps in the thought processes of the respondent when asked autobiographical questions on a survey (Tourangeau et al. 2000). First, the respondent must understand the words and concepts within the question, and what information is being requested by the survey researcher (comprehension). Second, they must be able to retrieve the relevant memories from their mind that provide the requested information (retrieval). Third, they must assess and combine information from their memories to create an answer to the question (judgment). Finally, they must take their internal answer and determine how to map it onto the appropriate part of the response scale for the survey question (response). The cognitive burden can be substantial, particularly if the questions address subjects that college students may not think about on a regular basis, or even think about at all.

Keeping in mind how SRLG questions are typically worded, accurate reporting of learning gains requires the following steps to occur. Students must:

1. Comprehend the meaning of the content area in each question item. As Table 1 shows, these questions are always vaguely worded, and it is not at all clear that students understand what "thinking critically" is, or what "understanding" means. Students must share a common understanding of these content areas; if not, subgroups of students will in essence be responding to different questions.
2. Know the level of their knowledge at college entry, in many different content areas. Note that this level of knowledge must be placed on some sort of scale that distinguishes low levels of knowledge from high. What kind of scale(s) students are using is unknown.

---

[1] All of these items, except for "contributing to the welfare of your community" and "understanding yourself", appear on the revised version of the NSSE to be used in 2013. The quantitative item has been revised to "analyzing numerical and statistical information." Two other SRLG items are also included that do not fit within the Holland framework (Pike 2011; Pike et al. 2011b), and are not discussed here (solving complex problems and developing a personal code of ethics).

3. Encode the level for each content area in their memory. Even if students knew their level of knowledge at entry, if this is not encoded in their memory, then it cannot be retrieved when students are surveyed at the end of their freshman and senior years.
4. Retrieve each level of knowledge at entry roughly 8 months and 3–6 years later, depending on when they graduate.
5. Know the level of their knowledge in each content area when surveyed at the second time point. The scale used to rate their level of knowledge at Time 2 must match the scale used when determining their level of knowledge at entry.
6. Subtract the Time 1 level of knowledge from the Time 2 level to estimate the amount of gain during college.
7. Somehow map this amount to a vague response scale that ranges from "very much" to "very little". For comparable responses across students, all students must use the same internal knowledge scales and response mapping systems.

When viewed in the light of the current model of survey response, it seems highly unlikely that the majority of college students, or humans in general, have the cognitive ability to successfully navigate all seven steps.

If students lack the ability to accurately report their gains in learning during college, one implication is clear: self-reported gains should be unrelated to actual learning gains. Empirical findings to date suggest this is the case. In a series of studies, Bowman (2010a, 2010b, 2011b) compares self-reported gains in learning to actual gains in learning, using objective tests of critical thinking and moral reasoning measured at two time points in college. The average of his correlations is .05. Given that the square of a correlation is equal to the R^2 of a bivariate regression between the two variables, this indicates that less than 1 % of the variation in SRLG is explained by actual learning gains. One drawback to these studies is that they focus on gains during the first year of college; some scholars have argued that not enough time elapses during the first year of college for students to show gains (Pike 2011).

A Theory of College Student Self-Reports

If existing theory and evidence suggests that students cannot accurately report how much they have learned, then how are students answering these questions? Given that several studies demonstrate consistent response differences across majors, it is clear that students are not simply generating random responses to these questions. In order to advance the field, it is vital that we develop a theory of college student survey response that yields testable predictions. An outline of the theoretical approach that I propose is as follows:

1. When tasked with a SRLG question, students use a belief-sampling approach to generate a response, rather than the seven-step recall and estimate approach described above.
2. Because SRLG questions ask students to report learning in reference to their experiences at their institution, their minds are flooded with considerations (beliefs, feelings, impressions and memories) related to their college experiences.
3. Many of these considerations will be unrelated to what they have actually learned, but correlated with student characteristics such as academic ability, interest in a content area, and experiences within their academic major.

If this is how college students respond to learning gains questions, then this would explain why there are weak relationships between actual versus self-reported gains in learning,

while self-reports vary across student characteristics as we might expect, if SRLG questions were indeed valid measures of learning.

The Belief-Sampling Model of Survey Response

Survey methodologists generally divide survey questions into two types. The first type of question is factual in nature, in that is has a correct answer. When asked an autobiographical question, for example, such as what is their grade-point average, or whether they have ever taken a service-learning course, a student can either give a correct or incorrect response. The second type of question, however, focuses on attitudes and subjective states, and has no answer that can be verified. If a student states that they are satisfied with their college education, there is no way we can independently verify their response, in contrast to their grade-point average or course-taking history. While scholars divide questions into these two groups, the division is usually based on expert judgment as to whether a question is objective or subjective. Generally, researchers assume respondents use the four-step response process (comprehension, retrieval, judgment, response) for factual questions, and the belief-sampling process for subjective questions.

The belief-sampling model of response (Tourangeau et al. 2000), as applied to attitudinal questions, posits a slightly different process during the retrieval and judgment stages of the response process.[2] Rather than retrieve actual memories and frequencies of events, retrieval instead "yields a haphazard assortment of beliefs, feelings, impression, general values, and prior judgments about an issue … (Tourangeau et al. 2000, p. 179). These are referred to collectively as "considerations." Importantly, what determines the exact set of considerations that come into someone's mind is accessibility; beliefs, feelings and related memories that are easily accessible are more likely to be retrieved. For any given topic, the number of considerations that are available will not all come to mind, and respondents instead unconsciously "sample" a set of considerations each time they are asked a question. This sampling process in part explains why responses to many attitudinal questions appear to be unstable and vary greatly over time. During the judgment stage, each consideration is combined to create a single response. As described by Tourangeau et al. (2000, p. 180), this is an

> … underlying process of successive adjustments. The respondent retrieves (or generates) a consideration and derives its implications for the question at hand; this serves as the initial judgment, which is adjusted in light of the next consideration that comes to mind; and so on … The formation of an attitude judgment is similar to the accretion of details and inferences that produces a frequency or temporal estimate.

The last sentence is important, because research on frequency estimates suggests that the more memories that are retrieved, the higher the estimated frequency of behavior (Bradburn et al. 1987). A similar process may be at work with college students, as students who retrieve many considerations related to learning in a specific content area may conclude they have learned a lot in that area during college.

Fundamental Assumption Underlying the Theory

The fundamental assumption of this theoretical approach to how students respond to SRLG questions is that in terms of cognitive response, students approach these questions as if they

---

[2] In Chapter 6 of their book, they review the evidence in favor of this model of survey response.

were attitudinal, rather than factual, questions. As is often the case, assumptions cannot be verified; they are simply assumed, as a basis for a theory that in turn will yield testable hypotheses. Here, this assumption is impossible to verify, because we cannot see into the brains of students to verify whether they perceive these questions as objective or subjective, and whether they actually try to retrieve their actual levels of learning at college entry and exit, or instead base their response on a sample of considerations.

However, we can argue that the assumption likely holds, as a matter of logic. First, note that while student survey response rates are typically low, it is clear that the students who decide to respond to the survey request want to provide answers to the survey questions; if not, then they would not be responding to the survey request. It is likely that these students are in part responding because of a helping norm, in response to the request for assistance from the survey researcher (Groves et al. 1992). Once they begin answering questions, students will want to help the researcher by providing a response to a question, even if the answer is not immediately obvious to them.

Second, as argued above, it is unlikely that students can successfully navigate the seven-step process that is necessary to accurately recall and report their gains in learning. Thus, students are in a position of (a) wanting to answer a question and (b) not being able to easily retrieve and report an answer. As Tourangeau et al. (2000) note, respondents will use different response strategies to respond to a question, and given that most students are cognitively unable to use the seven-step approach, it is likely they will shift to another response strategy. Given evidence about satisficing, where respondents seek to minimize the amount of work necessary to answer a question, it is also likely they will shift to a strategy that allows them to easily estimate an answer, such as the belief-sampling approach. One can also make a strong argument that the learning gains questions are impossible to answer on a factual basis; that is, given the seven-step process, students simply cannot retrieve the information requested. If so, these questions are best viewed as attitudinal questions rather than factual questions: they measure students' attitudes towards learning ("How much do I think I have grown") rather than objective growth in learning. If so, then it makes sense to adopt an attitudinal model of survey response to understand variation in student responses to these questions.

Considerations and Student Characteristics

Assuming that students use a belief-sampling approach when answering SRLG questions, what considerations come to mind when asked these questions? Consider a student in a quantitatively-oriented major who is asked how her college experiences have contributed to her development in analyzing quantitative problems. Multiple considerations then enter her mind: memories of lectures from a statistics class; memories of having possibly worked on problem sets with other groups of students; a general impression that she adept at math, based in part on her experiences in high school. These multiple, positive considerations then lead her to conclude that she has gained considerably in analyzing quantitative problems while in college. It is important to note that these considerations could easily be generated by a student, but that *none of them have anything to do with how much a student has learned while in college*. Because considerations that come into mind are a "haphazard assortment," it is clear that many, if not all, of the considerations that enter a student's mind will be related to their educational experiences, but not necessarily to how much they have actually learned in a specific content area. And because educational experiences are driven in large part by student choices, many of these considerations will also be related to student background characteristics.

Note that this approach does not preclude the use of relevant considerations when students construct a response. Students may also bring to mind considerations related to how much they have learned: memories of increasing grades on multiple papers submitted in a course, for example, or a comment made by an instructor about how much they have improved since the beginning of a semester. The crucial insight of the theory is that many other *non-relevant* considerations will also come to mind when students are answering questions about SRLG. These non-relevant considerations not only introduce considerable error into responses, but will also be correlated with student characteristics such as disciplinary interests, making the use of SRLG responses problematic.

Hypothesis 1 Higher ability students will report higher gains, because they have more academically oriented considerations.

Students who enter college with a strong academic background demonstrate higher levels of student engagement; they attend class more frequently, study more, etc. When queried as to how much they have learned in an area, they will be more likely to have multiple, related considerations come into their mind. A high school valedictorian queried about development in terms of writing clearly and effectively may recall the several English classes she took to improve her writing, and the multiple papers required in other courses she took because they had a reputation for being academically challenging. Another student, with only average high school grades, may have avoided English classes and sought courses known for multiple choice tests and "easy-A's". Perhaps no considerations about writing come to his mind. The two students report different amounts of gains, even though the first student may not have actually improved her ability to write, even after taking multiple courses that emphasize writing.

Note that this hypothesis is also consistent with self-reported learning gains as accurate proxies of actual learning. We would expect higher ability students to learn more, and thus report larger gains. Thus, in this paper I focus on deriving predictions from the next two hypotheses that yield empirical findings counter to what we should see if SRLG questions were valid measures of actual learning.

Hypothesis 2 Students with pre-college interests in an area will report higher gains, because their interests will cause them to seek out related educational experiences.

This hypothesis is based on the idea of students self-selecting into specific educational experiences due to their interests. Consider, for example, the SRLG content area on the NSSE, "contributing to the welfare of your community". A student who enters college with an interest in preserving the environment and volunteering may engage in many activities outside of class that will result in a flood of related considerations when asked this question, even if their skills in this area have not changed. Because students with these interests are also more likely to choose a major such as education or social work that emphasizes working with the community, large differences in reported gains across majors could occur. The implication of this hypothesis is that the large differences in self-reported gains that have been found in the literature are due in part to selection bias, and would be reduced when pre-college interests are taken into account.

Hypothesis 3 Students in academic majors that are congruent with SRLG questions will report higher gains due to considerations driven by educational experiences within their major.

A mathematics major asked to report on gains in analyzing quantitative problems will have a huge number of quantitative considerations come into their mind compared with someone majoring in art history, simply due to their experiences within their major. If the number and

variety of considerations are used by students to estimate their gains in a content area, some of the large differences in reported gains across majors are due to the size of the pool of considerations available to students in their minds. In other words, the mathematics major may have done quite poorly in mathematics courses, and not learned very much in terms of analyzing quantitative problems, while the art history major would not have learned much in this area due to their major focus. But given the difference in the size of the pool of their considerations, they would report very different learning gains. This suggests that using a measure of learning that is not affected by the survey response process should result in smaller differences between majors than a measure based on self-reported gains.

Predictions

This theory of college student survey response yields several empirical predictions about student survey responses. Because the two validation studies by Pike et al. are by far the stronger of the SRLG validity studies, largely because of their grounding in Holland's theory of person-environment fit, I base my predictions and empirical analyses on Holland's theoretical framework as well.

Prediction 1   The factor structure of SRLG for first-year students should be similar to the factor structure for seniors.

Pike et al. argue that SRLG questions have construct validity because the gains items from the NSSE cluster together as expected, into four factors that correspond to the four Holland environments of Investigative, Artistic, Social and Enterprising. Consider why we would expect student responses to items such as "working effectively with others" and "acquiring work-related knowledge and skills" to cluster together in a factor analysis to form an Enterprising factor; that is, why students who say they gained "very much" in terms of working effectively will also tend to choose "very much" for acquiring work-related knowledge. According to Pike et al.'s theoretical argument, students will gain in these two areas due to the socialization process of their major environment. Students in Enterprising majors, for example, will take many courses emphasizing these two content areas, with the result that these students will extensively develop in these areas, and then report to survey researchers that they have experienced substantial gains. Students in non-Enterprising majors, on the other hand, will take courses that place less emphasis on these two content areas, and subsequently report lower gains in these areas. The socialization process thus determines the factor structure that we observe.

Pike et al. argue that first-year students have not had enough time to be socialized within major environments. If true, then we would not expect the four-factor structure for SRLG questions that they propose and test on seniors to be replicated in a sample of first-year students: these students have not been in their majors long enough for responses to cluster together as theory would predict.

Consider a thought experiment, in which students are given a battery of SRLG questions during their first week of college. Almost everyone would choose "very little" or "some" as their response for each item, and the resulting factor analysis would probably yield a single factor, because students would not have time to achieve learning gains in related content areas. As time passes, and students take more courses during their college career, they then begin to gain in related content areas as Pike et al. would predict. Most of these related gains should take place later in the college career, as students take specialized courses within their major. Most first-year students take a variety of courses to satisfy general education requirements during their first year, so we would not necessarily expect

students who gain a lot in, for example, writing during their first year to also take a set of classes that emphasizes speaking and critical thinking. Conversely, if student responses to these questions are driven in part by the effect of their pre-college interests on subsequent student behavior during college, then we would expect the factor structure for first-year students to be similar to that of seniors.

Prediction 2 Substantively significant differences in learning gains between major groupings should exist for both first-year students and seniors.

The argument here is similar to Prediction 1. If first-year students do not have enough time to be socialized by their disciplines, then we should not see substantively significant differences across major groupings. However, if responses to SRLG are driven in part by their pre-college interests, and students select majors based on these interests, then we should see differences across major groupings similar to those reported by Pike et al.

Prediction 3 Differences between major groupings will decrease once pre-college interests are taken into account.

This is also derived from Hypothesis 2, and is a statistical argument about omitted variable bias. Controlling for content area interest at entry is essential, as Smart et al. (2000) have argued. When estimating differences in learning gains between different majors (or major groupings), we ideally wish to estimate the effect of academic disciplines on two identical students, who differ *only* in their choice of major. Such a comparison is generally only possible with randomization of treatment, but covariate adjustment within a regression framework is another approach that can yield plausible results, given a properly specified model. When specifying models with student self-reports as the dependent variable, taking into account not only demographic variables, but pre-college interests as well, is essential (Astin and Lee 2003).

Prediction 4 Differences between major groupings will decrease when objective measures of learning are used instead of subjective measures, such as SRLG questions.

This is derived from Hypothesis 3, and is based on the idea that if educational experiences drive considerations, and considerations in turn drive responses to SRLG questions (instead of actual gains in learning), then a measure of learning that is *not* driven by considerations and the belief-sampling approach will yield smaller differences across majors. The large differences between majors found in the literature are thus in part an artifact of the SRLG response process. Objective measures of learning that actually measure student learning should yield smaller differences, because students cannot use the belief-sampling approach when answering questions on these instruments. In other words, if Pike et al. are correct, and students can accurately report their learning gains, then the effect sizes for Holland major categories across different areas of learning should be fairly similar for both self-reported gains and actual learning gains. But if students are unable to accurately report learning gains, and instead generate a response based on the belief-sampling approach, they will likely overestimate their gains. Thus, any effect sizes calculated across Holland categories for actual learning gains will be much smaller than self-reported gains.

Methodology

To test these predictions, I use the Wabash National Study, a unique longitudinal study from 2006 to 2010 of students at 19 colleges and universities. At entry, students were administered the CAAP test for critical thinking. This is the objective measure of learning

that has been used in several of Bowman's SRLG validity studies, and has substantial evidence as to its validity (Porter 2011b). Students were also asked questions about their background and interests in different areas such as making a contribution to the sciences and achieving professional success. The critical thinking test, as well as the NSSE, were administered to the same students at the end of their first and fourth years.

The Wabash study staff coded student majors using the 2000 Classification of Instructional Program's six-digit coding scheme. These were matched with the appropriate 1990 CIP code using the U.S. Department of Education crosswalk, and then coded to the appropriate Holland major category using Gottfredson and Holland's (1996) *Dictionary of Holland Occupational Codes*, which lists 1990 CIP majors and codes and their corresponding 3-letter Holland coding.[3]

In addition to the four predictions listed above, I also test whether Pike et al.'s findings hold when also including Realistic and Conventional majors in the analyses. Given Holland's theory, inclusion of Realistic and Conventional majors in the regression models should also yield negative differences as predicted, e.g., Realistic majors should report lower gains on the Investigative outcome compared with Investigative majors. If they do not, then this is further evidence against the validity of SRLG. Previous research has failed to include these Holland major types.[4] In the current study, 2 % of majors are Realistic and 4 % are Conventional, and as the results demonstrate, there are enough students to yield statistically significant results.

Dependent Variables

Two sets of dependent variables are used to evaluate the predictions of the proposed theory. First, factor scores for the Investigative, Artistic, Social, and Enterprising learning outcomes are used. Panel d of Table 2 shows the factor structure for Wabash seniors when the analysis is constrained to four factors, using the same confirmatory approach as Pike at al. use. As can be seen by comparing these results to the top two panels of the table (a and b), the factor structures are almost identical, which is not surprising given that both studies use the same instrument on samples of students from multiple institutions.

The second set consists of a single variable, the CAAP test of critical thinking. This is an objective measure of learning that can be used to evaluate Prediction 4.

Independent Variables

Consistent with Pike et al.'s models, control variables in all models include gender, race/ethnicity, and whether the student lived on campus, was a first-generation college student,

---

[3] Several majors not listed in the *Dictionary* were coded as follows: Pre-Medicine Studies (51.1102) as Investigative per Smart et al. (2000); Gay/Lesbian Studies (05.0208), German Studies (05.0125), Italian Studies (05.0126) and Japanese Studies (05.0127) as Social, similar to other area/ethnic studies; Polish Language and Literature (16.0407) and English Language and Literature/Letters, Other (23.9999) as Artistic, similar to other language and literature majors; Early Childhood Education and Teaching (13.1210) as Social, similar to elementary education; Cell/Cellular and Molecular Biology (26.0406) as Investigative, similar to biology and biochemistry.

[4] It is not clear why previous researchers do not include these environments in their analyses. Pike (2011) states that "too few seniors were majoring in Realistic and Conventional disciplines to permit stable estimates of learning outcomes" (p. 49), but he does not recall the number of these majors in the NSSE that were available (Pike 2012). At least one Conventional major (Accounting) and two Realistic majors (Materials Engineering and Mechanical Engineering) are collected and coded by the NSSE. Given the large number of students in the NSSE response pool (approximately 416,000 in 2011), there should be enough students in these majors to include them in a statistical analysis.

or was a transfer student. All students in the Wabash sample were of traditional age. A second set of control variables is used to take into account student attitudes at college entry (see Table 3). The Wabash study asked students a series of questions for four scales that closely match the Holland learning outcomes:

- Importance of making a contribution to the sciences (Investigative)
- Importance of making a contribution to the arts and humanities (Artistic)
- Importance of political and social involvement (Social)
- Importance of professional success (Enterprising)

Supplementing these four scales, other pre-college measures that might affect learning gains and included as control variables are whether the student has a positive attitude towards literacy, their need for cognition, their academic motivation, and their academic ability as measured by ACT score. In addition to student-level variables, I also include school-level fixed effects that control for all differences between institutions, such as selectivity, size, mission, etc.

Models

Two sets of regression models are estimated, with standard errors that take into account the clustering of students within schools. The first set uses the four Holland learning outcome factor scores from the factor analysis of NSSE learning gains items, taken from NSSE administrations at the end of the first and fourth years. These dependent variables are similar to those created by Pike et al. The second set uses tests of critical thinking at the end of the first and fourth years as the dependent variable, with test scores at entry as a control variable. This approach to studying learning has been advocated by Pascarella and Wolniak (2004), and is similar to the value-added models currently being used in much of K-12 education (Rothstein 2009). Because there is some debate as to whether this is the best approach, I also estimate these models using difference scores (end of first year score/fourth year score minus entry score) and residual change scores.

In each of the multivariate models, dummy variables for the Holland major groupings are included, and tested against the relevant reference category. For example, in the models using the Holland Investigative outcome as the dependent variable, having a Holland Investigative academic major is the reference category, and dummies are included for majors in other Holland categories. Given Pike et al.'s theoretical argument, all of the dummy variables should be negative and statistically significant, indicating that students in non-Investigative majors score less on the Investigative outcome than Investigative majors. A similar approach is used for the model using critical thinking as the dependent variable. Because Investigative and Artistic majors should see the largest gains in critical thinking skills, there should be large differences between these majors and majors in other Holland categories.

Results

Prediction 1

Table 2 shows the factor structures from the Pike et al. studies and the Wabash data. As stated previously, the results for Wabash seniors are very similar to those of Pike et al.

Table 3 Wabash study: attitudinal scales

Contribution to the sciences ($\alpha = .73$)

Making a theoretical contribution to science

Working to find a cure for a disease or illness

Contribution to the arts and humanities ($\alpha = .67$)

Becoming accomplished in one of the performing arts (e.g., acting, dancing, singing, etc.)

Creating artistic work (e.g., painting, sculpture, film, etc.)

Writing original works (e.g., poems, novels, short stories, etc.)

Political and social involvement ($\alpha = .80$)

Becoming a community leader

Becoming involved in activities that preserve and enrich the environment

Helping others who are in difficulty

Improving my understanding of other countries and cultures

Keeping up to date with political affairs

Developing a meaningful philosophy of life

Helping to promote racial understanding

Influencing social values

Influencing the political structure

Integrating spirituality into my life

Volunteering in my community

Professional success ($\alpha = .77$)

Obtaining recognition from my colleagues for contributions to my field of expertise

Having administrative responsibility for the work of others

Working in a prestigious occupation

Making a lot of money

Becoming successful in a business of my own

Positive attitude toward literacy ($\alpha = .66$)

I enjoy reading poetry and literature.

I enjoy reading about science.

I enjoy reading about history.

I enjoy expressing my ideas in writing.

After I write about something, I see that subject differently.

If I have something good to read, I'm never bored.

Academic motivation ($\alpha = .67$)

I am willing to work hard in a course to learn the material even if it won't lead to a higher grade.

When I do well on a test, it is usually because I am well-prepared, not because the test is easy.

I frequently do more reading in a class than is required simply because it interests me.

I frequently talk to faculty outside of class about ideas presented during class.

Getting the best grades I can is very important to me.

I enjoy the challenge of learning complicated new material.

My academic experiences (i.e., courses, labs, studying, discussions with faculty) will be the most important part of college.

My academic experiences (i.e., courses, labs, studying, discussions with faculty) will be the most enjoyable part of college.

Need for cognition ($\alpha = .85$)

I would prefer complex to simple problems.

Table 3 continued

I like to have the responsibility of handling a situation that requires a lot of thinking.

Thinking is not my idea of fun.*

I would rather do something that requires little thought than something that is sure to challenge my thinking abilities.*

I try to anticipate and avoid situations where there is likely a chance I will have to think in depth about something.*

I find satisfaction in deliberating hard and for long hours.

I only think as hard as I have to.*

I prefer to think about small, daily projects to long-term ones.*

I like tasks that require little thought once I've learned them.*

The idea of relying on thought to make my way to the top appeals to me.

I really enjoy a task that involves coming up with new solutions to problems.

Learning new ways to think doesn't excite me very much.*

I prefer my life to be filled with puzzles that I must solve.

The notion of thinking abstractly is appealing to me.

I would prefer a task that is intellectual, difficult, and important to one that is somewhat important but does not require much thought.

I feel relief rather than satisfaction after completing a task that required a lot of mental effort.*

It's enough for me that something gets the job done; I don't care how or why it works.*

I usually end up deliberating about issues even when they do not affect me personally.

Note Students surveyed at college entry

Source http://www.liberalarts.wabash.edu/study-overview

* Item is reversed-worded

In terms of evaluating the first prediction, we can see that the factor structure for the Wabash students when they were surveyed at the end of their first year is almost identical to the factor structure for when they were surveyed at the end of their fourth year of college. Not only do items load onto the same factors, but the values of the factor loadings are very similar. This result implies a similar response process on the part of first-years and seniors. Given that students have not had enough time to be socialized by their academic disciplines, this is turn implies that something other than actual learning gains is driving responses. My argument is that student background and pre-college interests are determining the considerations that comes into students' minds when answering these questions, causing first-year students to exhibit a pattern of responses similar to seniors.

Prediction 2

Table 4 shows the results of models testing for differences in Holland learning outcomes between Holland major types, controlling for the independent variables used by Pike et al. in their validation studies. The top two panels show the coefficients from their results, while the bottom four panels show the results from the Wabash study dataset, first estimating differences without Realistic and Conventional majors to make the models as comparable as possible to Pike et al., and then including them to see whether the results change. Because the dependent variables are factor scores, the coefficients show differences in gains between Holland major groupings in terms of standard deviations.

Several patterns are evident. First, comparing Wabash seniors to the Pike et al. results (panel d vs. a and b), we can see that qualitatively the results are similar. All of the statistically significant Wabash coefficients are negative, following the pattern found by Pike et al. in their studies. Second, the effect sizes on average are much larger using the Wabash data. This is possibly due to differences and errors in coding. Pike et al. used the academic major titles provided by NSSE, and tried to match them by name using the same *Dictionary of Holland Occupational Codes* used in the current study. Such matching by name may lead to errors. Many schools in the Wabash study provided Wabash with the CIP codes for their majors, which allows for an exact match between a school's academic major CIP code and Holland environment using the *Dictionary*. Third, substantial differences can be seen between Holland major groupings for the Wabash first-year students (panel c). This is contrary to what the Holland approach to SRLG would predict, because first-year students have not been in college long enough for academic disciplines to begin affecting how much they have learned. The effect sizes for the Investigative learning outcome are substantial, over half a standard deviation, and are consistent with the idea that student self-selection into majors is partially responsible for student response differentials across majors.

Fourth, including Realistic and Conventional majors in the sample yields results inconsistent with the Holland theoretical approach. Theory would predict, for example, that Realistic majors should report less gains in the Investigative outcome compared with Investigative majors. The bottom panel of Table 4 indicates the opposite: Realistic majors report over half a standard deviation more gains than Investigative majors. A similar positive result occurs for Conventional majors; they are much more enterprising than Enterprising majors. While not related directly to the theoretical approach used in this paper, these results do call into question the validity results of Pike et al. using their validation methodology, because some differences in SRLG are diametrically opposite of what Holland's theory would predict.

Prediction 3

Table 5 shows the results for models controlling for students' ACT score, pre-college interests in Investigative, Artistic, Social, and Enterprising outcomes, their attitudes toward literacy and need for cognition, their academic motivation, and school-level fixed effects. Collectively, these variables partially control for student self-selection into institutions and academic majors. Comparing panel b of Table 5 to panel d of Table 4, we can see that the effects of Holland major groupings on seniors' SRLG are generally reduced when controlling for student self-selection, consistent with Prediction 3. All three statistically significant coefficients for the Investigative learning outcome decrease. The significant coefficients for the Artistic outcome also decrease, but the changes here are very small. One coefficient for the Social learning outcome does become more negative, and statistically significant, but the one negative coefficient for the Enterprising outcome drops by almost half. Contrary to Holland's theory, after controlling for student and school characteristics, there is now a positive difference between Social and Enterprising majors; Social majors are more enterprising by about one-third of a standard deviation.

The bottom panel of Table 5 is perhaps the most appropriate test of the validation approach for SRLG advocated by Pike et al., because it includes all six Holland major types, not just four, and also controls for student and school characteristics. Using the criterion of .20 for effect size (and ignoring statistical significance), only 11 out of the 20

Table 4 Effects of Holland disciplinary groupings on self-reported learning gains, with and without realistic and Conventional majors

		Investigative outcome factor	Artistic outcome factor	Social outcome factor	Enterprising outcome factor
(a)	*Pike (2011): End of senior year*[a]				
	Investigative major		−0.25***	−0.18***	−0.18***
	Artistic major	−0.33***		−0.10***	−0.20***
	Social major	−0.24***	−0.07***		−0.12***
	Enterprising major	−0.21***	−0.16***	−0.12***	
(b)	*Pike et al. (2011b): End of senior year*[b]				
	Investigative major		−0.25***	−0.18***	−0.18***
	Artistic major	−0.33***		−0.10***	−0.20***
	Social major	−0.24***	−0.07***		−0.12
	Enterprising major	−0.22***	−0.15***	−0.12***	
(c)	*Wabash: end of freshman year*[c]				
	Investigative major		−0.25*	−0.13	0.04
	Artistic major	−0.56***		0.05	−0.18
	Social major	−0.56***	0.00		0.19**
	Enterprising major	−0.59***	0.14	0.03	
(d)	*Wabash: end of senior year*[d]				
	Investigative major		−0.61***	−0.17	−0.15
	Artistic major	−0.89***		0.09	−0.40*
	Social major	−0.79***	−0.37*		0.39
	Enterprising major	−0.66***	−0.08	−0.04	
(e)	*Wabash: end of freshman year*[e]				
	Investigative major		−0.25*	−0.13	0.04
	Artistic major	−0.56***		0.05	−0.19
	Social major	−0.56***	−0.01		0.18*
	Enterprising major	−0.59***	0.14	0.02	
	Realistic major	0.06	−0.32	−0.20	0.27
	Conventional major	−0.33**	0.04	−0.09	0.23
(f)	*Wabash: end of senior year*[f]				
	Investigative major		−0.61***	−0.17	−0.15
	Artistic major	−0.89***		0.09	−0.40*
	Social major	−0.79***	−0.37*		0.39
	Enterprising major	−0.66***	−0.08	−0.04	
	Realistic major	0.54***	−0.61**	−0.27	0.12
	Conventional major	0.00	−0.43	−0.23	0.64***

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

[a] Table 3.2; controls for gender, race, on-campus housing, first-generation, transfer student and traditional age; $N = 20{,}000$

[b] Table 3; controls for gender, race, on-campus housing, first-generation, transfer student, and traditional age; $N = 20{,}000$

[c] Controls for gender, race, on-campus housing and first-generation; $N = 1{,}259$

[d] Controls for gender, race, on-campus housing and first-generation; $N = 895$

[e] Controls for gender, race, on-campus housing, first-generation, and includes Realistic and Conventional majors; $N = 1{,}352$

[f] Controls for gender, race, on-campus housing, first-generation, and includes Realistic and Conventional majors; $N = 965$

Table 5 Effects of Holland disciplinary groupings on self-reported learning gains, controlling for environmental self-selection

		Investigative outcome factor	Artistic outcome factor	Social outcome factor	Enterprising outcome factor
(a)	*Wabash: end of freshman year*[a]				
	Investigative major		−0.24	−0.05	0.07
	Artistic major	−0.25*		−0.01	−0.15
	Social major	−0.23	0.06		0.19*
	Enterprising major	−0.32**	0.10	−0.05	
(b)	*Wabash: end of senior year*[b]				
	Investigative major		−0.57**	−0.27*	−0.05
	Artistic major	−0.63***		0.06	−0.22
	Social major	−0.60***	−0.36*		0.35*
	Enterprising major	−0.53***	−0.07	−0.05	
(c)	*Wabash: end of freshman year*[c]				
	Investigative major		−0.23	−0.08	0.07
	Artistic major	−0.26*		−0.01	−0.15
	Social major	−0.22	0.07		0.18*
	Enterprising major	−0.31**	0.11	−0.06	
	Realistic major	−0.03	−0.28	−0.17	0.27
	Conventional major	−0.14	0.08	−0.09	0.14
(d)	*Wabash: end of senior year*[d]				
	Investigative major		−0.57***	−0.30**	−0.05
	Artistic major	−0.64***		0.04	−0.21
	Social major	−0.61***	−0.35*		0.36*
	Enterprising major	−0.53***	−0.08	−0.05	
	Realistic major	0.46***	−0.53*	−0.39	−0.01
	Conventional major	0.02	−0.35	−0.21	0.36**

Note Controls for gender, race, on-campus housing, first-generation, ACT and pre-college academic interests and motivation (see Table 3)

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

[a] $N = 1,259$

[b] $N = 895$

[c] $N = 1,352$; includes Realistic and Conventional majors

[d] $N = 965$; includes Realistic and Conventional majors

coefficients are negative and have an absolute value greater than .20; only 6 of the 20 are larger than .40. With only slightly more than half of the differences between major groupings larger than .20, it is difficult to conclude that these results support the idea that SRLG vary as theoretically predicted between Holland major groups. Note that this proportion is the same if we use a lower effect size cutoff of .10. More importantly, three of the differences in this panel are positive, statistically significant, and substantively large, contrary to expectations. In sum, this modeling approach provides a better test of whether SRLG vary across majors as expected, produces results inconsistent with Pike at al.'s findings shown in panels a and b of Table 4, and does not support their theoretical claims.

Prediction 4

Table 6 shows the results for the same set of models, but now using the CAAP critical thinking test score as the dependent variable. The first three columns use the critical thinking test score from the end of the first year of college, and the last three columns use the critical thinking test score from the end of the fourth year of college. Each column estimates a different model: the first replicates the models from the Pike et al. validation study using their control variables, the second includes controls for student characteristics at entry and school fixed effects, as described above, and the third expands the sample to include Realistic and Conventional majors. Because Pike et al. assume that critical thinking is an outcome for both Investigative and Artistic majors, these majors are combined and used as the reference category for the Holland major grouping dummy variables.

Four patterns are evident. First, the statistically significant coefficients are all negative, consistent with the idea that students in each of these majors are scoring lower on critical thinking tests than Investigative and Artistic majors. Second, as with the SRLG dependent variables, the differences between Holland major groupings decrease when student self-selection is taken into account. For first-year Social majors, the differences decrease from six points on the CAAP to no statistically significant difference for first-year students. The differences decrease 1 point in two analyses and increase four points in the gains score analysis for seniors. For Enterprising majors, the differences decrease from 2 to five points to no statistically significant difference for seniors, depending on the type of analysis. Third, there are no statistically significant differences between major groupings for first-year students, while some small effects remain for seniors. Given what we know about student learning, academic majors, and the effect of college, these results are more plausible than the results in Table 4 showing large differences for first-year students. Most students have not been in a major for a significant amount of time by the end of their first year, so we would expect to see null findings across major groupings for these students. Fourth, and most importantly, the effect sizes in this table are much smaller than the effect sizes for the Investigative outcome in Tables 4 and 5, consistent with Prediction 4. Given the standard deviation of the CAAP for this sample, the effect sizes for the statistically significant regression coefficients in the last column of the table range from .25 to .35, much smaller than the effect sizes in Tables 4 and 5 and those reported by Pike et al.

In sum, the results presented here provide evidence in favor of the belief-sampling model of student response. First, first-year students demonstrate a very similar factor structure for SRLG as seniors. However, first-year students have had much less time to be influenced by the academic environment compared to seniors, which implies that a similar process (other than the effect of college) must be driving their responses. Second, first-year students and seniors show similar differences in SRLG between Holland major groupings; this is also consistent with both groups of students using a belief-sampling approach to generate a response, rather than generating a response based on how much they have learned since entering college. In addition, inclusion of additional Holland major categories yields predictions opposite to what we should see if SRLG were valid measures of learning. Third, the differences in SRLG between Holland major groupings decrease once student self-selection into schools and majors is taken into account; we also observe differences between Holland major groupings opposite of what we should see if these were valid measures of learning. Fourth, there are almost no differences between Holland major groupings once self-selection is taken into account, and the models are estimated with a measure of learning that cannot be answered using a belief-sampling response process.

Table 6 Effects of Holland disciplinary groupings on CAAP critical thinking

	First-years	First-years	First-years	Seniors	Seniors	Seniors
Pre-test as control						
Social major	−5.84*	−4.70	−4.57	−8.64**	−7.70*	−7.84*
Enterprising major	−4.39	−1.48	−1.92	−4.68*	−2.27	−2.35
Realistic major			−0.90			−0.98
Conventional major			−4.06			−2.61
Gain score analysis						
Social major	−3.03	−4.14	−4.09	−2.04**	−5.43*	−5.64*
Enterprising major	−2.37	−2.62	−3.11	−1.89*	−4.80	−4.94
Realistic major			3.29			6.68
Conventional major			−4.92			−2.15
Residual analysis						
Social major	−5.50*	−4.37	−4.30	−8.11**	−6.85*	−7.00*
Enterprising major	−4.15	−2.15	−2.61	−4.46*	−3.22	−3.33
Realistic major			1.51			1.93
Conventional major			−4.55			−2.44
N	624	624	668	431	431	465
Pike et al. controls?	Yes	Yes	Yes	Yes	Yes	Yes
Student and school controls?	No	Yes	Yes	No	Yes	Yes
Includes R and C majors?	No	No	Yes	No	No	Yes

Note The reference category for the Holland major dummy variables is Investigative and Artistic majors combined. Pre-test as control models use Y at end of year as the dependent variable, with Y at entry as a control variable. Gain score analysis models use Y at end of year minus Y at entry as the dependent variable. Residual analysis models regress end of year Y on Y at entry, and use the difference between the actual end of year Y and predicted end of year Y as the dependent variable. * $p < 0.05$, ** $p < 0.01$

Conclusion

Colleges and universities increasingly face pressure to demonstrate that they are actually doing what they are supposed to be doing. While there is some debate as to what the primary metric for postsecondary institutions should be, student learning is generally considered a primary outcome. Given these pressures, it would be wonderful if we could simply ask students how much they have learned in college, and then use these data for institutional assessment, as well as for academic research. The results of this study suggest that student responses to these questions are largely unrelated to actual gains in learning.

The paper proposes a simple theory of college student survey response based on the belief sampling model. Hypotheses derived from this theory suggest that students' pre-college characteristics, and experiences in their academic major, drive the considerations that pop into students' minds when answering a self-reported learning gains question. These considerations are then used to create a response. This in turn suggests that much of the survey response is caused by factors unrelated to actual learning, and explains why Bowman's work shows no relationship between SRLG and actual measures of learning gains. The theory also explains why SRLG vary across majors as we might expect: students with pre-college interests in a content area self-select into a major, and their experiences within their major further color their responses.

The theory yields four predictions confirmed by the data: (1) first-year students and seniors should have similar factor structures for SRLG; (2) first-year students should show large differences across majors in SRLG, even though they have not spent enough time in their major to be affected by it; (3) SRLG differences across majors will decrease once we take into account student self-selection into majors; and (4) when students are given an objective test of learning that cannot be answered by reflecting upon their experiences, the large differences across majors observed for SRLG will also decrease in size.

Besides demonstrating the problems with recent validation studies supporting the validity of SRLG questions, this paper also demonstrates the great need for college student researchers to develop theoretical models of how college students can correctly answer autobiographical questions on the many surveys that we use. The field has been conspicuously silent on this front, instead constantly reciting the following mantra:

Research suggests that self-report data are likely to be valid under five conditions:

1. The information is known to respondents;
2. The questions are phrased clearly and unambiguously;
3. The questions refer to recent activities;
4. The respondents think the questions merit a serious and thoughtful response; and
5. Answering the question does not threaten, embarrass, or violate the privacy of the respondent or encourage the respondent to respond in socially desirable ways (Kuh (2001, p. 4), Pike et al. (2012, p. 559), and many others).

This is not a theoretical model of student survey response, but simply a list of conditions that rarely hold for the typical college student survey. It is difficult to defend the use of student survey responses in both institutional and academic research when we have no idea how students are generating the responses that we use in our analyses.

Looking beyond college student surveys, this paper also raises two issues that Holland researchers should consider when using Holland's theory to study college students. First, it is not clear why researchers in this area have consistently excluded students in Realistic and Conventional majors. Although claims are made about too few students, none of the studies found in the literature listed these numbers, nor provided an explicit statistical explanation, such as a power analysis, for why these students could not be included in the analyses. As the results here demonstrate, inclusion of these majors leads to findings contrary to Holland's theoretical predictions, and demonstrates the importance of their inclusion whenever possible.

Second, the larger coefficients for the Wabash data compared with the Pike et al. results in Table 4 raises the question of how college majors are coded into Holland categories. Using program CIP codes from colleges and mapping them to an existing Holland crosswalk like the *Dictionary of Occupational Codes* is fairly straightforward, but mapping by major name likely leads to errors. For example, Smart et al. (2000) and Pike (2006) classify Electrical Engineering as a Realistic major, while Pike (2011), Pike et al. (2011b) and Gottfredson and Holland (1996) classify Electrical Engineering as Investigative. An examination of the *Dictionary* shows that similar-sounding majors may be classified as different Holland types. It lists, for example, 13 different nursing majors, some of which are classified as Social, others as Investigative, while Smart et al. (2000) state that Nursing in general should be classified as Social. The correct coding of majors is undoubtedly essential for any analysis, and it is likely that the academic major categories collected on student surveys, such as the CIRP and NSSE, are too broadly worded to be correctly coded into Holland environments.

The vast majority of quantitative studies of college students relies on survey data collected from students. While there is a rich literature studying nonresponse in student surveys, and whether nonresponse leads to bias in results, very few studies have investigated the validity of college student survey questions. Yet error introduced by poorly constructed questions could be worse than that introduced by nonresponse behavior. Recent research, for example, has suggested that nonresponse bias is unrelated to survey response rates (Groves 2006; Groves and Peytcheva 2008; Keeter et al. 2006). However, researchers are more focused on obtaining a good response rate for their survey, instead of using questions on their survey that have evidence for their validity. Given our reliance on student survey data, it is essential that we learn more about how students comprehend and construct a response to our survey questions. We know little about whether students can accurately report autobiographical information, and whether common attitudinal questions, such as satisfaction with college, exhibit response stability over time.

In terms of future research on SRLG, it would be useful, for example, to administer SRLG several times during a semester. If these reflect actual learning during college, we would expect to see strong response stability during a period of only a few months. Attitudinal questions are also notoriously susceptible to context effects, because considerations from previous questions remain in the respondent's mind as they answer a question. Altering the order of SRLG questions, and including questions prior to SRLG that might change responses, would be another possible way to provide validation evidence for SRLG as measures of actual learning. Only through the use of theory and carefully constructed validation studies can we begin to understand whether the vast amount of survey data that we collect can be considered valid and reliable measures of student background, behavior, and attitudes.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for Educational and psychological testing*. Washington, DC: American Educational Research Association.

Astin, A. W., & Lee, J. J. (2003). How risky are one-shot cross-sectional assessments of undergraduate students? *Research in Higher Education, 44*(6), 657–672.

Biglan, A. (1973a). The characteristics of subject matter in different academic areas. *Journal of Applied Psychology, 57*(3), 195–203.

Biglan, A. (1973b). Relationships between subject matter characteristics and the structure and output of university departments. *Journal of Applied Psychology, 57*(3), 204–213.

Bloom, H. S., Hill, C. J. J., Black, A. R., & Lipsey, M. (2008). Performance trajectories and performance gaps as achievement effect-size benchmarks for educational interventions. *Journal of Research on Educational Effectiveness, 1*, 289–328.

Bowman, N. A. (2010a). Assessing learning and development among diverse college students. *New Directions for Instutional Research, 145*, 53–71.

Bowman, N. A. (2010b). Can 1st-year students accurately report their learning and development? *American Educational Research Journal, 47*, 466–496.

Bowman, N. A. (2011a). Examining systematic errors in predictors of college student self-reported gains. *New Directions for Instutional Research, 150*, 7–19.

Bowman, N. A. (2011b). Validity of college self-reported gains at diverse institutions. *Educational Researcher, 40*, 22–24.

Bradburn, N., Rips, L. J., & Shevell, S. (1987). Answering autobiographical questions: The impact of memory and inference on surveys. *Science, 123*, 157–161.

Campbell, C. M., & Cabrera, A. F. (2011). How sound is NSSE?: Investigating the psychometric properties of NSSE at a public, research-extensive institution. *Review of Higher Education, 35*, 77–103.

Ewell, P., McClenney, K., & McCormick, A. C. (2011). *Measuring engagement*. Inside Higher Education. Crawfordsville: Wabash College.

Gottfredson, G. D., & Holland, J. L. (1996). *Dictionary of Holland occupational codes*. Lutz, FL: Psychological Assessment Resources Inc.

Groves, R. M. (2006). Nonresponse rates and nonresponse bias in household surveys. *Public Opinion Quarterly, 70*, 646–675.

Groves, R. M., Cialdini, R. B., & Couper, M. P. (1992). Understanding the decision to participate in a survey. *Public Opinion Quarterly, 56*(4), 475–495.

Groves, R. M., & Peytcheva, E. (2008). The impact of nonresponse rates on nonresponse bias: A meta-analysis. *Public Opinion Quarterly, 72*, 167–189.

Higher Education Data Sharing Consortium. (2012a). HEDS senior survey. Downloaded June 10, 2012, from http://www.hedsconsortium.org/storage/HEDS_Senior_Survey_Sample_02-23-2012.pdf.

Higher Education Research Institute. (2012b). College senior survey. Downloaded June 10, 2012, from http://www.heri.ucla.edu/researchers/instruments/FUS_CSS/2012CSS.PDF.

Holland, J. L. (1973). *Making vocational choices: A theory of vocational personalities and work environment*. Upper Saddle River, NJ: Prentice-Hall.

Huang, Y., & Healy, C. (1997). The relations of Holland-typed majors to students' freshman and senior work values. *Research in Higher Education, 38*, 455–477.

Keeter, S., Kennedy, C., Dimock, M., Best, J., & Craighill, P. (2006). Gauging the impact of growing nonresponse on estimates from a national RDD telephone survey. *Public Opinion Quarterly, 70*(5), 759–779.

Kuh, G. D. (2001). *The national survey of student engagement: Conceptual framework and overview of psychometric properties*. Bloomington, IN:Indiana University Center for Postsecondary Research.

Kuh, G. D., & Vesper, N. (2001). Do computers enhance or detract from student learning? *Research in Higher Education, 42*(1), 87–102.

Laird, T. F. N., Shoup, R., Kuh, G. D., & Schwarz, M. J. (2008). The effects of discipline on deep approaches to student learning and college outcomes. *Research in Higher Education, 49*(6), 469–494.

Lambert, A. D., Terenzini, P. T., & Lattuca, L. R. (2007). More than meets the eye: Curricular and programmatic effects on student learning. *Research in Higher Education, 48*(2), 141–168.

McCormick, A., Pike, G., Kuh, G., & Chen, P.-S. (2009). Comparing the utility of the 2000 and 2005 carnegie classification systems in research on students's college experiences and outcomes. *Research in Higher Education, 50*, 144–167. doi:10.1007/s11162-008-9112-9.

McCormick, A. C., & McClenney, K. (2012). Will these trees ever bear fruit? A response to the special issue on student engagement. *Review of Higher Education, 35*(2), 307–333.

National Survey of Student Engagement. (2012). *Sample institutional report*. Downloaded June 9, 2012, from http://nsse.iub.edu/_/?cid=402.

Pace, R. C. (1985). The credibility of student self-reports. Technical report, Center for the Study of Evaluation, University of California, Los Angeles.

Pascarella, E. T., Blaich, C., Martin, G. L., & Hanson, J. M. (2011). How robust are the findings of academically adrift? *Change, 43*(3), 20–24.

Pascarella, E. T., & Padgett, R. (2011). *Using institution-level NSSE benchmarks to assess engagement in good practices: A cautionary note*. Manuscript, University of Iowa.

Pascarella, E. T., & Wolniak, G. C. (2004). Change or not to change—Is there a question? A response to pike. *Journal of College Student Development, 45*(3), 353–355.

Pike, G. R. (2000). The influence of fraternity or sorority membership on students' college experiences and cognitive development. *Research in Higher Education, 41*(1), 117–139.

Pike, G. R. (2006). Students' personality types, intended majors, and college expectations: Further evidence concerning psychological and sociological interpretations of Holland's theory. *Research in Higher Education, 47*(7), 801–822.

Pike, G. R. (2011). Using college students' self-reported learning outcomes in scholarly research. *New Directions for Instutional Research, 150*, 41–58.

Pike, G. R. (2012, June 28). Personal communication.

Pike, G. R., & Killian, T. S. (2001). Reported gains in student learning : Do academic disciplines make a difference? *Research in Higher Education, 42*(4), 429–454.

Pike, G. R., Kuh, G., McCormick, A., Ethington, C., & Smart, J. (2011a). If and when money matters: The relationships among educational expenditures, student engagement and students' learning outcomes. *Research in Higher Education, 51*(1), 81–106.

Pike, G. R., Smart, J. C., & Ethington, C. A. (2011b). *Differences in learning outcomes across academic environments: Further evidence concerning the construct validity of students' self-reports*. Paper presented at the annual conference of the Association for the Study of Higher Education, Raleigh, NC.

Pike, G. R., Smart, J. C., & Ethington, C. A. (2012). The mediating effects of student engagement on the relationships between academic disciplines and learning outcomes: An extension of Holland's theory. *Research in Higher Education, 53*(5), 550–575.

Porter, S., & Umbach, P. D. (2006). Student survey response rates across institutions: Why do they vary? *Research in Higher Education, 47*, 229–247.

Porter, S. R. (2011a). Do college student surveys have any validity? *Review of Higher Education, 35*, 45–76.

Porter, S. R. (2011b). *Student learning as a measure of quality in higher education*. Context for Success Project. Seattle: Gates Foundation.

Rothstein, J. (2009). Student sorting and bias in value-added estimation: Selection on observables and unobservables. *Education Finance and Policy, 4*(4), 537–571.

Smart, J. C. (2010). Differential patterns of change and stability in student learning outcomes in Holland's academic environments: The role of environmental consistency. *Research in Higher Education, 51*, 468–482.

Smart, J. C., Feldman, K. A., & Ethington, C. A. (2000). *Academic disciplines: Holland's theory and the study of college students and faculty*. Nashville, TX: Vanderbilt University Press

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

Zhao, C.-M., & Kuh, G. D. (2004). Adding value: Learning communities and student engagement. *Research in Higher Education, 45*(2), 115–138.