American Educational Research Journal
June 2010, Vol. 47, No. 2, pp. 466–496
DOI: 10.3102/0002831209353595

Can 1st-Year College Students Accurately Report Their Learning and Development?

Nicholas A. Bowman
University of Notre Dame

Many higher education studies use self-reported gains as indicators of college student learning and development. However, the evidence regarding the validity of these indicators is quite mixed. It is proposed that the temporal nature of the assessment—whether students are asked to report their current attributes or how their attributes have changed over time—best accounts for students' (in)ability to make accurate judgments. Using a longitudinal sample of over 3,000 first-year college students, this study compares self-reported gains and longitudinal gains that are measured either objectively or subjectively. Across several cognitive and noncognitive outcomes, the correlations between self-reported and longitudinal gains are small or virtually zero, and regression analyses using these two forms of assessment yield divergent results.

KEYWORDS: self-report, validity, college students, student learning, student development

Many constituencies within and outside of higher education are concerned with the learning and development of college students. College administrators, staff, faculty, parents, legislators, taxpayers, and others all want to know whether and how students benefit from their (rather expensive) years spent in college. This desire was recently placed center stage in the Spellings Commission, which called for greater public accountability and transparency regarding college learning outcomes (U.S. Department of Education, 2006). To date, there are relatively few ways to measure student learning and development at the individual, institutional, state, or national level (Burke, 2005; National Center for Public Policy and Higher Education, 2004).

NICHOLAS A. BOWMAN is a postdoctoral research associate in the Center for Social Concerns at the University of Notre Dame, 164 Geddes Hall, Notre Dame, IN 46556-4633; e-mail: *nbowman@nd.edu*. His research interests include the outcomes and psychological processes associated with college diversity experiences, the validity of college student outcomes assessments, and the impact of college rankings on various constituencies.

The need for high-quality assessment and accountability begs the question of how student outcomes are effectively and accurately measured. Compared with conducting longitudinal assessments with objective measures, asking students to report how much they have gained during college is relatively inexpensive, requires minimal financial and human resources, and provides results in a short period of time. As a result, higher education researchers and administrators frequently use self-reported gains as indicators of student growth. According to a review by Gonyea (2005), self-reports constitute a useful proxy, but not a substitute, for objective measures. However, the literature in this area is quite mixed, with some studies reporting strong agreement between subjective and objective assessments (Berdie, 1971; Pike, 1995, 1996; Pohlmann & Beggs, 1974), whereas others find substantial biases in self-report measures (Pike, 1993) and strong divergence between subjective and objective assessments (Astin, 1993; Bowman & Seifert, 2009; Dumont & Troelstrup, 1980). This article explores the degree to which the most common form of assessment of learning and development—college students' self-reported gains—accurately reflects longitudinal gains.

When Do Subjective and Objective Measures Diverge?

As Gonyea (2005), Pike (1996), and others have noted, previous studies have differed greatly when assessing how closely subjective and objective measures align with each other. Two explanations that account for this variability have been proposed. Pike (1995) and Dumont and Troelstrup (1980) suggested that subjective and objective measures might differ in the content areas that they assess. That is, disparities between subjective and objective measures may stem, in part, from the fact that these measures capture largely divergent constructs. In addition, Astin (1993) and Baird (1976) proposed that low correspondence may be related to the breadth of content measured. They argue that objective measures, particularly of cognitive abilities, may be very effective at assessing particular skills (i.e., they have "high fidelity"), whereas they may cover only a small range of skills (i.e., they have "low bandwidth"). Conversely, self-report measures may have high bandwidth but low fidelity. In some ways, this second explanation is a subset of the first, because it proposes that the content of objective measures may represent only a portion (and a more accurately defined portion) of the content covered in self-report measures. Both of these explanations are intuitively plausible and may account for some of the variation in previous research. However, it is not clear whether there is a difference in content overlap between studies that find lower correlations and those that find higher correlations.

A third explanation, which has not been offered in the previous literature, focuses on whether students are estimating their current attributes or

how their attributes have changed over time. Importantly, the temporal nature of the assessment nicely explains apparent discrepancies in previous research. According to the literature, students are reasonably accurate when estimating their own abilities and attributes at a single point in time, as defined by how closely subjective assessments align with objective assessments. For example, Berdie (1971) found correlations ranging from .47 to .74 between participants' self-assessments of general knowledge and their objectively tested knowledge of famous people. In another study, the correlations between students' self-assessments of current academic subject knowledge and their scores on academic subject exams were also quite high, ranging from .52 to .67 (Pohlmann & Beggs, 1974). Moreover, when Pike (1995, 1996) compared students' self-ratings of their existing academic skills with objective measures of those skills, he found a strong relationship between these forms of assessment within a larger structural equation model.

On the other hand, studies that use self-reported gains show a much greater divergence between these estimates and objective measures. For example, Astin (1993) found that significant predictors of gains in standardized test scores bear little to no resemblance of predictors of subjective gains. Moreover, Pike (1993, 1999) found that significant "halo effects" are apparent in self-reported gains; that is, self-reported gains in various domains were moderately or strongly correlated with one another, but this correspondence may have been substantially inflated by biases in students' judgments. This tendency was most pronounced among 1st-year students, for whom halo effects accounted for 47% to 75% of the explained variance in self-reported gains (Pike, 1999).

Other researchers do not explicitly argue for a strong divergence between self-reported gains and objective gains, but their data speak to this issue directly. Anaya (1999) compared the results of regressions predicting GRE verbal and math performance (controlling for SAT scores) with regressions predicting self-reported verbal and math gains; all other predictor variables were the same. Replicating Astin's (1993) earlier observation, the results of analyses predicting self-reported gains versus objective, longitudinal gains diverged notably. In addition, Whitt and colleagues (Whitt, Edison, Pascarella, Nora, & Terenzini, 1999) examined the relationship between interactions with peers and changes in cognitive outcomes. They used two measures of peer interactions and multiple cognitive measures, including objective tests (administered upon entering college and later in college) and self-reported gains. When controlling for a host of precollege factors and college experiences, peer interactions were consistently and positively related to self-reported gains in cognitive development. Specifically, Whitt et al. (1999) observed significant positive effects of peer interactions on self-reported cognitive gains in 25 separate regression analyses, whereas not a single analysis contained a negative relationship. In contrast, when

predicting objectively measured longitudinal gains, peer interactions were associated with *decreases* in cognitive skills in the same number of analyses in which they predicted gains.

Biases and Errors in Subjective Judgments

Why do students' self-reported gains not align more closely with gains on objective, longitudinal measures? There are numerous factors that contribute to this divergence. The most mundane reason is simple measurement error. No construct, whether measured subjectively or objectively, is ever captured perfectly, and this imperfection reduces the observed relationship between these two types of outcomes. As noted previously, the degree of content overlap between the two measures also can be an important factor (Pike, 1994, 1995); the smaller the content overlap, the smaller the correspondence. Social desirability also may play an important role in shaping students' subjective responses (Cole & Gonyea, 2008; Gonyea, 2005). Regardless of their actual beliefs, students may report sizable gains on a number of attributes, such as academic skills, ability to interact with diverse groups, and interpersonal skills and relationships, even if they do not think they have gained very much. Indeed, one study found that self-reported gains are positively correlated with a well-validated social desirability scale, and this relationship persists when controlling for numerous other variables (Bowman & Hill, 2009). The tendency toward providing seemingly desirable responses could occur for at least two reasons. First, students do not want to admit their own (perceived) lack of learning; thus, they may report large gains when they do not believe these have occurred. Second, if respondents think they know what the researchers want to find, many people will change their responses in an effort to "help" the researchers achieve the desired results (Aronson, Ellsworth, Carlsmith, & Gonzales, 1989). This may constitute a substantial problem if students believe the purpose of a survey is to assess their growth and development.

Furthermore, students may respond to specific self-reported gain items in a way that reflects their overall perception of gains, which would diminish the validity of specific self-report responses. The tendency to respond to specific items based on general perceptions of a subject is known as the halo effect. This phenomenon has been well documented in a variety of domains over time, and it has been observed in people's judgments about themselves and about others (see Cooper, 1981). As noted earlier, Pike (1999) has found that halo effects can be massive, accounting for over half of the explained variance in self-reported gains among 1st-year students and at least one quarter of explained variance among seniors (also see Pike, 1993).

Perhaps most importantly, people tend to be surprisingly inaccurate when making various judgments about themselves. For instance, Gilbert (2007) has demonstrated that people are not very good at predicting what

makes them happy. Intuitively, it seems quite easy to predict which experiences will lead to happiness, but people's predictions do not align well with their own subsequent ratings of happiness in real-world situations. In general, people are confident about all kinds of self-knowledge, such as identifying factors that influence their decision making and knowledge of certain traits and characteristics, but they are often incorrect (Nisbett & Wilson, 1977; Pronin & Kugler, 2007; Wilson, 2002). In one of the classic works of social psychology, Nisbett and Wilson (1977) reviewed relevant research on people's awareness of their own mental processes. They conclude that (a) people are largely unaware of their previous attitudes when these views have shifted over time; (b) in experimental settings, people's conscious assessments of their mental states and the behavioral manifestations of those states are virtually uncorrelated; and (c) people's perceptions of attitude change and mental processes reflect a priori causal theories about these processes.

In the context of self-reported gains, students probably feel that they have a great deal of insight into their own learning and development, but their judgments may be quite erroneous. According to Nisbett and Wilson (1977) and Ross (1989), students' self-reported gains reflect causal theories that make intuitive sense to them, and they will be correct to the degree that these implicit causal theories align with reality. For example, students report greater knowledge gains in their major than in other disciplines (Pace, 1984); this subjective assessment, which is consistent with general "common sense," seems exceedingly likely to be replicated through any objective assessment. However, experiences that would logically appear to influence students' skills and attributes may not always do so, which can lead to errors in judgment. For example, Conway and Ross (1984) randomly assigned college students into either a series of study-skills workshops or a waitlist for the workshops. Students who participated in the workshops expected to receive higher grades in their major, and they reported receiving higher grades when subsequently asked about the semester in which they took the workshops. However, these students overestimated their grades in both instances; that is, they expected higher grades than they later received, and they falsely recalled having received higher grades than they had actually been given. These faulty self-reports were consistent with students' causal theory, which was that study-skills workshops should lead to better study skills and higher grades. In contrast, students who were on the waitlist for the workshops had no reason to expect their grades to improve, and they did not overestimate their predicted or actual grades.

Why Estimating Gains Is Especially Difficult

Given the numerous potential sources of bias for student self-reports, it is quite impressive that Berdie (1971) and Pohlmann and Beggs (1974) found

such high correlations between subjective and objective measures at a single time point. However, there are additional reasons that gains over time may be more difficult to estimate than one's current attributes. Some of these reasons are best understood through Tourangeau and colleagues' four-stage model of the psychology of survey responses (Tourangeau, Rips, & Rasinski, 2000). The four steps involved, in order, are *comprehension* of the question, *retrieval* of memories associated with the question, *judgment* of the completeness and relevance of the memories, and mapping the judgment onto a *response* represented by one of the options provided. It seems that retrieval and judgment processes for self-reported gains may be especially difficult. To provide an accurate judgment, students would need to estimate their current attributes, remember their previous attributes on an outcome, and have some means of comparing the previous level with their current level. However, Ross (1989) shows that this is not what people do in practice; instead, people tend to estimate their current attributes or abilities and then decide whether or how these have changed. This judgment process can result in a host of errors. Overall, people tend to overestimate how much their skills and abilities have changed, yet underestimate how much their attitudes have changed (Conway & Ross, 1984; Goethals & Reckman, 1973; Markus, 1986; McFarland & Ross, 1987; Reiter, 1980). Importantly, both of these biases are consistent with students' lay theories of change and stability (Ross, 1989). Therefore, self-reported gains should be very difficult to estimate accurately, because students generally believe that college contributes to growth and development, and self-reported gains place high demands on retrieval and judgment processes.

The four-step process that Tourangeau and colleagues (2000) describe implies that respondents are making a concerted effort to respond to the best of their knowledge and ability. However, this may not be the case on every question in a college student survey. Krosnick (1991) argues that many respondents will provide a satisfactory answer instead of exerting the cognitive effort that many survey items require, which is a process known as "satisficing." Some of the conditions that promote satisficing are more likely to occur when students are asked to estimate self-reported gains than when asked to assess their current traits and abilities. First, satisficing is more likely when respondents answer more cognitively challenging items. As noted earlier, reporting gains over time is more difficult than estimating current attributes, because identifying one's current attributes serves as only part of the process of estimating gains. Second, satisficing occurs more often when people have no preconceived answer to a given question. It seems much more likely that people have preformed answers regarding their current traits than how these traits have changed over time. Third, having a succession of response categories with the same scale—regardless of whether these items are intended to measure the same construct—may lead to greater satisficing, because participants who respond to similar items

with the same available answers may lose focus. This concern seems particularly problematic for self-reported gains, because perceived gains on numerous outcomes are often assessed consecutively, the response options are typically the same for all items, and none of the choices are reverse coded.

The creation of a meaningful scale also may be more problematic for self-reported gains than for current assessments. Scales that ask students to report their own gains typically contain responses such as having gained "a lot." However, the definition of gaining "a lot" in critical thinking, for example, may differ dramatically for students who are entering college with relatively higher or lower levels of critical thinking. That is, the same response may mean something very different to different people (e.g., Pace & Friedlander, 1982), and this problem is likely to be more pronounced for self-reported gains than for current subjective assessments. In addition, as Pascarella (2001) describes, some students are much more likely to report having gained from their experiences, and the use of demographic control variables in regression equations is largely ineffective in reducing this bias. This tendency among some students likely contributes to an overall halo effect in self-reported gains. To help correct this bias in cross-sectional studies, Pascarella recommends asking students about gains in high school on the same measures, but this practice is rarely employed in higher education research.

Regarding all types of self-reports, Kuh and colleagues (Kuh et al., 2001; Umbach & Kuh, 2006) propose five conditions that should be met for participants' responses to be considered valid:

> (a) the information requested is known to the respondents, (b) the questions are phrased clearly and unambiguously, (c) the questions refer to recent activities, (d) the respondents think the questions merit a serious and thoughtful response, and (e) answering the questions does not threaten, embarrass, or violate the privacy of the respondent or encourage the respondent to respond in socially desirable ways. (Umbach & Kuh, 2006, p. 173)

For college students providing self-reported gains on well-established surveys, it seems quite possible that some of these conditions—particularly, (b), (c), and (d)—are often met. However, the first (and perhaps most important) condition is unlikely to be satisfied; that is, regardless of how much students would like to provide honest answers in response to clear, thought-provoking questions, they may not have access to the necessary information for judging their own gains.

Measuring Longitudinal Change in Student Outcomes

Although estimating change with longitudinal measures may seem like a simple task, scholars in education, psychology, and other disciplines have long discussed and debated the most effective way to do so. If several

observations over time are available for each student, then multilevel or growth curve modeling—with time as the Level 1 variable and students at Level 2—constitutes an effective technique for measuring change (Raudenbush & Bryk, 2002; Rogosa, Brandt, & Zimowski, 1982; Singer & Willett, 2003). However, the vast majority of longitudinal studies of college student development contain data from two time points (a pretest and a posttest), and growth curve modeling is not appropriate when this is the case (Raudenbush & Bryk, 2002; Rogosa et al., 1982).

Despite the difficulties associated with estimating change, researchers have concluded that change scores from two-wave data can be reliable and valid in many circumstances (Rogosa et al., 1982; Rogosa & Willett, 1983; Zimmerman, 2009). Several approaches for computing change scores from pretest-posttest data have been offered, but none of these is universally accepted. The most straightforward approach is to calculate a difference score by subtracting the pretest from the posttest. However, this approach has been critiqued for a number of reasons. Lord (1956) and others (Cronbach & Furby, 1970) have advocated for the use of "true change" or "true gain" scores that are designed to minimize or eliminate some potential problems. The true gain score is an adjusted form of the difference score that accounts for the unreliability of the pretest and posttest scores and the correlation between pretest and posttest scores. Alternatively, other scholars (DuBois, 1957; Tucker, Damarin, & Messick, 1966) have argued for the use of residual scores, which are derived from the residuals when posttest scores are regressed on pretest scores. This technique is designed to yield gain scores that are uncorrelated with pretest scores. To complicate matters further, Rogosa et al. (1982) argue that difference scores actually constitute the most appropriate measure of change between two points in time, because (a) difference scores are unbiased estimators of true change scores, (b) residual scores and true change scores have problematic characteristics that difference scores do not, and (c) most critiques of difference scores are fallacious or apply only when certain assumptions are true.

In higher education contexts, researchers are generally most interested in identifying precollege characteristics or college experiences that are related to changes in student learning and development. Among studies that use two-wave data, the most common approach is to use multiple regression or related analyses to predict posttest scores while controlling for pretest scores. This technique for examining predictors of change is considered preferable to other approaches, such as predicting difference scores or predicting posttest scores without controlling for pretest scores (Nunnally, 1982; Werts & Linn, 1970).

Present Study

In examining the validity of self-reported data, previous studies have compared the results of objective and subjective measures, but it is

sometimes unclear what a divergence between these two types of measures actually means. Does a large disparity between objective and subjective measures suggest that the subjective measures are less valid, or does it suggest that subjective measures reflect a different (yet valid) set of attributes? The current study addresses this issue directly by comparing self-reported gains and longitudinal gains on a variety of outcomes with a large, multi-institutional sample of 1st-year college students. Some of the longitudinal assessments are measured objectively, and others are measured subjectively. If subjective and objective measures simply gauge divergent constructs, then the correlation between self-reported gains and *objective* longitudinal gains should be low, whereas the correlation between self-reported gains and *subjective* longitudinal gains should be reasonably high. On the other hand, if students are unable to make accurate judgments of self-reported gains, then the correlations between self-reported gains and *all* longitudinal gains should be low. A strong divergence between longitudinal and self-report results across all outcomes would have profound implications, because this would mean that the form of assessment would substantially affect the main findings of any study of college student outcomes. Four research questions were addressed:

1. How closely do college students' self-reported gains correspond with longitudinal gains that purportedly measure the same construct?
2. How closely do longitudinal gains correspond with one another, and how closely do self-reported gains correspond with one another?
3. How similar are the results of regression analyses that predict self-reported and longitudinal gains?
4. To what degree does the correspondence between self-reported gains and longitudinal gains depend upon whether the longitudinal outcome is measured subjectively or objectively?

Method

Data Source and Participants

Data from the Wabash National Study of Liberal Arts Education were used for this study. Nineteen colleges and universities (11 liberal arts colleges, 2 community colleges, 3 research universities, and 3 regional universities) were included in the sample on the basis of their strong commitment to liberal arts education. The study sample contained both private and public institutions, along with religiously affiliated, single-gender, and minority-serving schools. Institutions exhibited a wide range of selectivity, tuition costs, and geographic diversity.

Students who were beginning their freshman year in fall 2006 were invited to participate in a longitudinal study. Before classes began or during their

first 2 to 3 weeks on campus (Time 1), students completed a registration form that included demographic information; a questionnaire of various high school experiences, interests, attitudes, and values; and a battery of five assessments. All students completed all assessments, except half of the students completed a critical thinking measure (the Critical Thinking module of the College Assessment of Academic Proficiency [CAAP]; ACT, 1991), whereas the other half completed a measure of moral reasoning (the Defining Issues Test 2 [DIT2]; Bebeau & Thoma, 2003; Rest, Narvaez, Thoma, & Bebeau, 1999). Students received $50 for their participation, and a total of 4,501 students completed this first wave. At the end of their freshman year (Time 2), students who took part in the initial assessment were invited to participate in a second round of data collection. They completed the same battery of assessments, along with questionnaires that asked about their college experiences, interests, attitudes, values, and self-reported gains during their 1st year. Once again, students who completed all measures received $50 as compensation. A total of 3,081 students participated in the second wave, yielding a retest response rate of 68%. Of these students, 3,072 had valid data on the assessments; 1,569 completed the CAAP Critical Thinking module, and 1,503 completed the DIT2. Among students who responded to both waves of the survey, 54.9% were female, 81.5% were White non-Hispanic, 7.4% were Asian or Pacific Islander, 5.1% were Hispanic, 4.0% were Black non-Hispanic, 0.4% were American Indian or Alaska Native, and 1.5% did not report their race or ethnicity.

Measures

Dependent variables. Four items that asked students to estimate self-reported gains were included. These items were "thinking critically and analytically," "developing a personal code of values and ethics," "understanding people of other racial and ethnic backgrounds," and "understanding yourself." All items used a 4-point scale (1 = *very little* to 4 = *very much*).

Four corresponding longitudinal measures were used, all of which have been validated in previous research (for descriptive statistics, see Table 1). The CAAP Critical Thinking module was used to gauge critical and analytical thinking. Developed by ACT, this 40-minute, 32-item instrument measures a student's ability to clarify, analyze, evaluate, and extend arguments. Students receive an overall score for the number of questions that they answer correctly. The internal consistency reliabilities for the CAAP Critical Thinking test range between .81 and .82 (ACT, 1991), and CAAP Critical Thinking scores correlate very highly (r = .75) with the Watson-Glaser Critical Thinking Appraisal (Pascarella, Bohr, Nora, & Terenzini, 1995).

Developing a personal code of values and ethics was measured longitudinally with the DIT2. The DIT2 and its predecessor, the Defining Issues Test, are designed to measure students' moral judgments. The application

Table 1
Means and Standard Deviations for Dependent Variables and Difference Scores

	Time 1		Time 2		Difference Score		
Dependent Variable	*M*	*SD*	*M*	*SD*	*M*	*SD*	Effect Size
CAAP Critical Thinking	51.99	28.21	53.30	30.36	1.31	19.35	0.04
DIT2 N2 score	31.73	15.72	36.61	15.95	4.88	11.94	0.31
MGUDS Relativistic Appreciation	4.78	0.66	4.72	0.74	−0.06	0.68	−0.08
SRLS Consciousness of Self	3.92	0.56	3.97	0.55	0.05	0.49	0.09

Note. Cohen's *d* (Cohen, 1988) was used to compute effect sizes for the difference between average pretest and posttest scores. All dependent variables were subsequently standardized with a mean of zero and a standard deviation of one for inclusion in the regression analyses. CAAP = College Assessment of Academic Proficiency (ACT, 1991); DIT2 = Defining Issues Test 2 (Bebeau & Thoma, 2003; Rest, Narvaez, Thoma, & Bebeau, 1999); N2 = new index (University of Minnesota, Center for the Study of Ethical Development, n.d.); MGUDS = Miville-Guzman Universality-Diversity scale (Fuertes, Miville, Mohr, Sedlacek, & Gretchen, 2000; Miville et al., 1999); SRLS = Socially Responsible Leadership Scale (Tyree, 1998).

of one's own complex framework for considering moral problems is associated with high levels of reasoning; in other words, students who have high N2 (new index) scores on the DIT2 have developed and are employing their own code of values and ethics. Previous evidence has shown that the N2 score is reliable (α = .77 to .81; Rest et al., 1999; University of Minnesota, Center for the Study of Ethical Development, n.d.), and the N2 and its predecessor (the P-score) predict a variety of forms of moral thinking and behavior (for a synthesis of this literature, see Pascarella & Terenzini, 2005).

The two longitudinal measures described above—the CAAP Critical Thinking module and DIT2—were clearly objective in nature. The other two longitudinal measures are primarily subjective; that is, they are based on participants' self-perceptions. Understanding people from other racial and ethnic backgrounds was measured longitudinally with the relativistic appreciation subscale of the Miville-Guzman Universality-Diversity scale (MGUDS; Fuertes, Miville, Mohr, Sedlacek, & Gretchen, 2000; Miville et al., 1999). Basing their conception on the work of Vontress (1986, 1988) and Yalom (1985), Miville and colleagues (1999) describe a relativistic appreciation of diversity as a recognition and valuation of the similarities and differences among diverse people. Clearly, recognizing commonalities and similarities among diverse groups of people is a necessary component of understanding people from diverse racial backgrounds. The overall score for the MGUDS scale is positively associated with empathy and a more complex racial identity, whereas it is negatively associated with dogmatism and

homophobia (Miville et al., 1999). The current study used the short form of the MGUDS (Fuertes et al., 2000); the Relativistic Appreciation subscale contains five items that are each rated on a 5-point scale (1 = *strongly disagree* to 5 = *strongly agree*). In the present sample, the internal reliability of the MGUDS Relativistic Appreciation scale is .77 at Time 1 and .79 at Time 2, and this scale is highly correlated with Pascarella and colleagues' (Pascarella, Edison, Nora, Hagedorn, & Terenzini, 1996) Openness To Diversity and Challenge Scale (r = .50 at Time 1 and r = .56 at Time 2).

Understanding oneself was measured by the Consciousness of Self subscale of the Socially Responsible Leadership Scale (SRLS; Tyree, 1998), which is based on Astin and colleagues' (1996) social change model of leadership development. In this model, knowing who you are is seen as a critical component for exhibiting effective leadership skills. This subscale very directly and straightforwardly asks participants to rate their level of self-knowledge with items such as "I know myself pretty well." The scale included nine items, and the internal reliabilities were .78 at Time 1 and .80 at Time 2. In the present sample, consciousness of self is highly correlated with Ryff's (1989) Self-Acceptance Scale (r = .64 at Time 1, and r = .68 at Time 2).

To examine the direct correspondence between self-reported and longitudinal gains, three separate measures of longitudinal gains were computed: the difference score, the residual change score, and the true change score. McNemar's (1958) equation for the true change score was used, because—unlike other formulas—this formula does not rely upon the questionable assumption that the reliabilities and standard deviations are identical for the pretest and posttest.

Independent variables. Consistent with many studies of higher education, numerous independent variables—including demographic and precollege characteristics, institutional type, and curricular and cocurricular college experiences—were used. The variables within each of these categories are also common in higher education research (e.g., race-ethnicity, gender, age, and family income as demographic variables). Moreover, the independent variables in this study were almost identical to those used in previous studies that examined the same dataset (Bowman, in press-a, in press-b).

Several demographic variables were included. First-generation students were defined as students whose parents had not attended any postsecondary education (1 = *first generation*, 0 = *other*). In addition, a series of dummy-coded variables was used to indicate race-ethnicity, which included Black non-Hispanic, American Indian or Alaska Native, Asian or Pacific Islander, Hispanic, and students who did not report their race or ethnicity. White non-Hispanic served as the referent group. Other demographic variables were gender (0 = *female*, 1 = *male*) and whether students were traditional college age (0 = *19 years old and younger*, 1 = *20 years old and older at the beginning of freshman year*). Parental income was recoded into several

categories, because over 10% of participants did not report their parents' income. Dummy variables were computed for low-income students (parents' combined income less than $35,000 per year), high-income students (at least $100,000 per year), and students who did not report their income (many of whom reported that they were economically self-sufficient). Middle-income students ($35,000-$99,999) served as the referent group.

A number of additional precollege variables were also used. Degree aspirations were coded on a 6-point scale (1 = *vocational or technical certificate or diploma* to 6 = *doctorate degree*). High school grade point average (GPA) was also included. Because high school GPA was strongly skewed, dummy codes were created for students who had a B average (B+ to B–) and for those with a C or D average (C+ or lower); students who had an A average (A+ to A–) served as the referent group. Academic motivation upon entering college was measured with an eight-item scale (α = .69). The racial-ethnic composition of one's high school was also included. For all students, this variable was coded such that higher values reflect a high school student body that is similar to oneself. That is, for White non-Hispanic students, 1 = *almost all students of color*, whereas 5 = *almost all White students*. On the other hand, for students of color, 1 = *almost all White students*, whereas 5 = *almost all students of color*.

A number of college experience variables were also included. Because the number of hours spent working on campus and the number of hours spent working off campus were weakly and negatively correlated, these were treated as separate variables. Moreover, given the strong skew of both variables, these were both recoded with 0 hours per week as the referent group. Dummy-coded variables were created to indicate working 1 to 15 hours per week, and 16 or more hours per week. In addition, dummy-coded variables indicated whether a student had an athletic scholarship and whether he or she was a member of a social fraternity or sorority. To indicate one's living situation, dummy-coded variables also denoted whether a student lived in a fraternity or sorority house and whether he or she lived in non-Greek on-campus housing (e.g., residence halls), with living off campus as the referent group. Continuous single-item measures gauged the number of hours spent participating in cocurricular activities and the number of hours spent relaxing and socializing (for both measures, 1 = *0 hours* and 8 = *more than 30 hours*). The frequency of drinking alcoholic beverages also was included. Because a majority of students reported not drinking during their 1st year, dummy-coded variables were created to gauge the number of times per week that a student drank alcohol: one to two times per week and three or more times per week, with zero times as the referent group.

Several items were included to reflect interactions with diverse peers. An eight-item scale gauged the frequency of positive diversity experiences (α = .89); each item was measured with a 5-point scale (1 = *never* to 5 = *very often*). A five-item index also was created to measure the frequency of

negative interactions with diversity (α = .83) using the same 5-point scale. Because most students reported that these interactions occurred hardly ever or never, this mean-scaled index was recoded into dummy-coded variables representing rare negative interactions (at least 1.5 and less than 2.5 on the 5-point scale) and somewhat common negative interactions (at least 2.5), with hardly ever or never as the referent group (i.e., less than 1.5).

Additional measures reflected students' course work and classroom experiences. The number of courses taken that focused on issues of diversity, gender, and social justice was included. Because the continuous variable for the number of courses was strongly skewed, several dummy-coded variables were used in the analyses; zero courses served as the referent group, and dummy-coded variables were computed to reflect one course, two courses, and three or more courses. In addition, an index of total number of courses taken was computed by adding the number of courses taken in nine subject areas. To avoid overlap among the independent variables in the regression models, the number of diversity courses taken was subtracted from this total. Several indices were created to gauge the impact of experiences with faculty and course work. A four-item index assessed the frequency of faculty contact (α = .70); each item was gauged with a 4-point scale (1 = *never* to 4 = *very often*). Also, a six-item index measured the level of challenge that occurred in the classroom (α = .82) using a 5-point scale (1 = *never* to 5 = *very often*). *In-class challenge* refers to how often students and faculty challenged each other's ideas and how often students argued for their point of view. A three-item index reflected the promptness of instructor feedback (α = .68), with all items using this same 5-point scale. The clarity and organization of teaching was gauged with 10 items (α = .89), which also used the same 5-point scale (1 = *never* to 5 = *very often*).

Finally, institutional type was included to gauge institutional differences that were not measured by other college experience variables. Dummy-coded variables were created for research universities, regional universities, and community colleges, with liberal arts colleges as the referent group.

Analyses

To determine the simple correspondence between self-reported gains and longitudinal gains, Pearson correlations between self-reported gains and the three measures of longitudinal gains (i.e., difference scores, residual change scores, and true change scores) were computed. Moreover, correlations among the self-reported gains and among the longitudinal gains were calculated.

A series of ordinary least squares (OLS) multiple regression analyses was conducted with the four longitudinal measures at Time 2 as outcomes and the four self-reported gain measures as outcomes. To ensure comparability between the results, regression analyses that predicted self-reported gains in critical thinking and developing a personal code of ethics examined

only students who had valid scores on the CAAP and DIT2, respectively. This sample selection was performed so that the same students were included in the corresponding self-report and longitudinal analyses. The full sample was used to analyze understanding people from other racial-ethnic backgrounds and understanding oneself. Moreover, to permit the comparisons of regression coefficients discussed below, all dependent variables were subsequently standardized with a mean of zero and a standard deviation of one.

The same independent variables were used for all regression equations, except the longitudinal measures contained Time 1 levels of the relevant construct as an additional control variable. The independent variables in all analyses were first-generation status, race-ethnicity, age, gender, parental income, high school GPA, high school demographics, degree aspirations, academic motivation, institutional type, time spent working on campus, time spent working off campus, time spent relaxing and socializing, time spent participating in cocurricular activities, living situation, Greek membership, athlete status, total courses taken, the number of diversity courses taken, positive experiences with diverse peers, negative or hostile experiences with diverse peers, frequency of faculty contact, in-class challenge, promptness of feedback from instructors, and teaching clarity and organization. Because the self-reported gain measures used a 4-point scale, ordered logit regression analyses are actually preferable to OLS regression for these outcomes (Long, 1997). However, OLS analyses were conducted so that the coefficients for regression analyses predicting longitudinal gains and self-reported gains could be compared systematically. (Preliminary analyses showed that OLS and ordered logit regressions provided very similar patterns of significant results.) Variance inflation factor (VIF) statistics were computed for all regression analyses. Across every analysis, no independent variable had a VIF greater than 5, and only one predictor had a VIF greater than 3 (dummy-coded variable for community college students).

Two separate methods were used to compare the results of the regression analyses predicting self-reported and longitudinal gains. First, the regression coefficients for each independent variable were examined using a technique described by Cohen and colleagues (Cohen, Cohen, West, & Aiken, 2003, pp. 46–47). For example, the coefficient for involvement in a fraternity or sorority predicting longitudinal gains in critical thinking was compared to the coefficient for same independent variable predicting self-reported gains in critical thinking. A confidence interval method was used to determine whether the unstandardized coefficients were significantly different from each other ($p < .05$).

Although this technique from Cohen et al. (2003) provides evidence of measurement differences, many consumers of higher education studies are largely interested in whether a significant relationship exists between an independent variable and the dependent variable. As a result, the difference between two positive regression coefficients in which $p = .04$ and $p = .14$ is

substantial in terms of interpretation, even though the corresponding regression coefficients would almost certainly not differ from each other. Therefore, a second technique was used to compare variables that were significant predictors of the relevant outcomes ($p < .05$). After each regression was performed (e.g., predicting self-reported gains in critical thinking), the degree to which the regression analyses on the corresponding measure (e.g., CAAP Critical Thinking module) replicated these findings was examined. For instance, the regression analyses showed that 12 independent variables significantly predicted self-reported gains in critical thinking. The number of these 12 independent variables that also significantly predicted CAAP scores (in the same direction) was determined and then divided by 12 to yield the proportion of effects that were replicated in the alternative form of analysis. The same procedure also was performed in the opposite direction; that is, the proportion of the significant effects from the longitudinal measure analyses that were replicated in the self-report analyses also was determined. The pretest measures for the longitudinal analyses were not included in any of these computations.

Limitations

Some limitations should be noted. First, the sample includes only 1st-year college students. Because halo effects in self-reported gains tend to be greater for 1st-year students than for seniors (Pike, 1999), the correspondence between longitudinal measures and self-reported gains for this sample may be lower than for more advanced students. Moreover, the 1st year of college often is a time of substantial change and transition (e.g., Upcraft, Gardner, & Barefoot, 2004), so 1st-year students might experience different developmental trajectories than more advanced undergraduates. Second, the objectively measured longitudinal variables in this study are cognitive in nature, whereas the subjectively measured variables are intra- or interpersonal. Thus, one cannot conclusively determine whether the nature of the outcome or the form of assessment is responsible for any differences in the results (this issue will be discussed later in more detail). Third, the intended content of the longitudinal measures may not perfectly match that of self-reported gains, particularly in the case of developing a personal code of ethics and understanding people from other racial backgrounds. Although the constructs measured by DIT2 and the MGUDS Relativistic Appreciation scale are closely related to these concepts, they are not one and the same. Fortunately, the intended constructs for the self-report and longitudinal measures are virtually identical for critical thinking and understanding oneself. This does not imply that the *actual* content of these measures is the same; indeed, one would expect well-validated, multi-item scales to have different psychometric properties than single items. Also, as shown in the Results section, the observed correspondence between self-report and

Table 2
Correlations Between Self-Reported Gains and Longitudinal Gains of the Same Construct

Construct	Measure of Longitudinal Gains		
	Difference Score	Residual Change Score	True Change Score
Thinking critically and analytically	.01	.02	.03
Developing a personal code of values and ethics	−.01	.01	−.01
Understanding people of other racial and ethnic backgrounds	.13***	.22***	.15***
Understanding yourself	.12***	.18***	.12***

*$p < .05$. **$p < .01$. ***$p < .001$.

longitudinal gains for critical thinking and understanding oneself is quite similar to that for personal code of values and understanding people from other racial backgrounds.

Results

The Pearson correlations between self-reported gains and longitudinal gains are generally small or virtually zero. As shown in Table 2, the correlations between self-reported gains and all three measures of longitudinal gains for critical thinking (rs = .01 to .03) and personal code of ethics (rs = –.01 to .01) do not differ significantly from zero. The correlations between self-reported gains and longitudinal gains for understanding people from other racial backgrounds (rs = .13 to .22) and understanding oneself (rs = .12 to .18) are all significantly greater than zero (ps < .001), but these are still small according to Cohen's (1988) guidelines (.10 to .30 is considered small; .30 to .50, medium; and .50 and greater, large).

The correlations among the longitudinal gains are also quite small, ranging from –.15 to .24 (median r = .05). As shown in Tables 3 and 4, regardless of the way in which longitudinal gains are measured, there is a significant negative correlation between gains in CAAP Critical Thinking scores and SRLS Consciousness of Self (rs = –.15 to –.05, ps < .05). There are also significant positive correlations between gains in SRLS Consciousness of Self and DIT2 scores (rs = .06 to .11, ps < .05) and between gains in SRLS Consciousness of Self and MGUDS Relativistic Appreciation (rs = .18 to .24, ps < .001). With the exception of the correlation between residual change scores for DIT2 scores and MGUDS Relativistic Appreciation (r = .09, p < .01), all other correlations were nonsignificant. In contrast, the correlations among self-reported gains are

Table 3
Correlations Among Difference Scores and Among Residual Change Scores

	CAAP	DIT2	MGUDS	SRLS
CAAP Critical Thinking	—	N/A	−.01	−.05*
DIT2 N2 score	N/A	—	.09**	.11***
MGUDS Relativistic Appreciation	−.03	.05	—	.24***
SRLS Consciousness of Self	−.15***	.06*	.18***	—

Note. CAAP = College Assessment of Academic Proficiency (ACT, 1991); DIT2 = Defining Issues Test 2 (Bebeau & Thoma, 2003; Rest, Narvaez, Thoma, & Bebeau, 1999); N2 = new index (University of Minnesota, Center for the Study of Ethical Development, n.d.); MGUDS = Miville-Guzman Universality-Diversity scale (Fuertes, Miville, Mohr, Sedlacek, & Gretchen, 2000; Miville et al., 1999); SRLS = Socially Responsible Leadership Scale (Tyree, 1998); N/A = not applicable. Correlations between CAAP Critical Thinking module and the DIT2 are not observed, because respondents completed only one of the two measures. Numbers below the diagonal represent correlations among difference scores, and those above the diagonal represent correlations for residual change scores.
*$p < .05$. **$p < .01$. ***$p < .001$.

Table 4
Correlations Among True Change Scores

	CAAP	DIT2	MGUDS	SRLS
CAAP Critical Thinking	—			
DIT2 N2 score	N/A	—		
MGUDS Relativistic Appreciation	−.03	.05	—	
SRLS Consciousness of Self	−.13***	.07*	.18***	—

Note. CAAP = College Assessment of Academic Proficiency (ACT, 1991); DIT2 = Defining Issues Test 2 (Bebeau & Thoma, 2003; Rest, Narvaez, Thoma, & Bebeau, 1999); N2 = new index (University of Minnesota, Center for the Study of Ethical Development, n.d.); MGUDS = Miville-Guzman Universality-Diversity scale (Fuertes, Miville, Mohr, Sedlacek, & Gretchen, 2000; Miville et al., 1999); SRLS = Socially Responsible Leadership Scale (Tyree, 1998); N/A = not applicable. Correlations between CAAP Critical Thinking module and the DIT2 are not observed, because respondents completed only one of the two measures.
*$p < .05$. **$p < .01$. ***$p < .001$.

medium to large and uniformly positive; these range from .32 to .55, with a median of .43 (see Table 5).

The parallel regression analyses also showed a divergence between the results using self-reported gains versus longitudinal gains (see appendix). Across the four outcomes, 32% (53 out of 168) of the pairs of regression coefficients for self-reported and longitudinal gains differed significantly. For

Table 5
Correlations Among Self-Reported Gains

	Critical Thinking	Personal Code of Ethics	Understanding People From Other Races	Understanding Oneself
Critical thinking	—			
Personal code of ethics	.40***	—		
Understanding people from other races	.32***	.45***	—	
Understanding oneself	.34***	.55***	.51***	—

*$p < .05$. **$p < .01$. ***$p < .001$.

example, as shown toward the top of the appendix, the regression coefficient for Asian or Pacific Islander students predicting self-reported gains in critical thinking ($\beta = -.078$) is significantly more negative than the respective coefficient for predicting longitudinal gains ($\beta = -.004$); the significant difference between these values is denoted by the asterisk to the right of the two coefficients. The proportion of significant differences in coefficients is virtually identical for analyses predicting critical thinking (31%), personal code of ethics (33%), understanding people from other racial-ethnic backgrounds (29%), and understanding oneself (33%). Clearly, these figures are substantially higher than the 5% of significant differences that would be expected by random chance (this expected figure is 5% because the cutoff for statistical significance was set at $p < .05$).

Moreover, the correspondence between significant predictors of these two forms of assessment is also quite low. As summarized in Table 6, only 32% of significant effects that occurred in the analyses of self-reported gains were replicated in the longitudinal analyses. Similarly, 31% of significant effects from the longitudinal analyses were replicated in the self-reported gain analyses. Moreover, the level of correspondence diverged greatly across the various outcomes. For the subjectively measured longitudinal outcomes, 52% of significant effects from self-reported gain analyses and 52% of significant effects from longitudinal analyses were replicated with the other type of measure. Conversely, for the objectively measured longitudinal outcomes, only 4% of significant effects were replicated in either direction. In fact, three independent variables that significantly predict self-reported gains are actually significant in the *opposite* direction when predicting longitudinal gains. For instance, frequency of faculty contact is positively associated with self-reported gains in critical thinking, whereas it is negatively related to objective longitudinal gains in critical thinking (see appendix).

Table 6
Summary of Correspondence Between Significant Predictors of Self-Reported Gains and Longitudinal Measures From Multiple Regression Analyses

	Proportion of Significant Predictors of Self-Reported Gains Replicated by Longitudinal Measures	Replication by Type of Longitudinal Outcome	Overall Replication
Critical thinking	1/12 (8%)	Objective:	18/56 (32%)
Personal code of ethics	0/11 (0%)	1/23 (4%)	
Understanding people from other races	10/19 (53%)	Subjective: 17/33 (52%)	
Understanding oneself	7/14 (50%)		
	Proportion of Significant Predictors of Longitudinal Measures Replicated by Self-Reported Gains	**Replication by Type of Longitudinal Outcome**	**Overall Replication**
Critical thinking	1/13 (8%)	Objective:	18/58 (31%)
Personal code of ethics	0/12 (0%)	1/25 (4%)	
Understanding people from other races	10/17 (59%)	Subjective: 17/33 (52%)	
Understanding oneself	7/16 (44%)		

Discussion

The most noteworthy finding is the overall low correspondence between longitudinal and self-reported gains. The correlations between longitudinal and self-reported gains were certainly greater when the longitudinal measures were subjective rather than objective, but the correlations between self-reported gains and subjective longitudinal gains were still small. Thus, it seems that the disparity between self-reported and longitudinal gains primarily stems from the temporal nature of these assessments, not from whether the outcomes are measured objectively or subjectively. Because people are much better at making judgments about their current attributes than their past attributes (Goethals & Reckman, 1973; Markus, 1986; Reiter, 1980), the lack of correspondence between longitudinal and self-reported gains calls into question the validity of self-reported gains as an accurate indicator of college student learning and development. Given the prevalence of self-report measures in higher education research (Gonyea, 2005), researchers and administrators at many institutions should seriously consider the efficacy and validity of their current assessment practices. Although college administrators are unlikely to base their decisions on a single self-report item, it is unclear whether combining or simultaneously

considering a number of self-reported gains would actually yield more valid conclusions.

Self-reported gains also seem to contain a substantial amount of halo error; in other words, the correlations among self-reported gains are artificially inflated by errors in judgment. Although it can be difficult to discern whether correlations among self-reported gains constitute errors or true relationships among constructs (Pike, 1999), a comparison of the intercorrelations among longitudinal gains and self-reported gains is instructive. The correlations among longitudinal gains are all quite low, and the median correlation ($r = .05$) is not significantly different from zero. Perhaps not coincidentally, the highest correlation for longitudinal measures occurred between the two subjective assessments, yet this correlation is still fairly small (rs = .18 to .24 across the three measures of longitudinal gains). Conversely, the correlations among self-reported gains are all greater than .30, with a median of .43. For the sake of argument, one might assume that Astin (1993) and Baird (1976) are correct when suggesting that objective measures more accurately gauge a smaller set of skills or attributes than do self-report measures or that Dumont and Troelstrup (1980) and Pike (1995) are correct in asserting that divergence between objective and subjective measures often stems from differences in content. Even if these assertions are true, one would still expect to find substantial positive correlations between objective and/or longitudinal measures if positive associations actually exist, but these correlations are virtually zero. As a result, the moderate to large correlations among self-reported gains seem to be largely driven by errors in students' judgments.

However, we cannot entirely attribute the self-reported gain correlations to halo error, because the self-reported gain items appeared consecutively in the survey, which may artificially inflate the correlation among these items to some degree. That is, the observed correlations may be the product not only of students' difficulties with assessing their learning and development in various domains, but also of survey characteristics that contribute to students' responding similarly to a set of adjacent items. The greater the role that survey design plays in contributing to these errors, the greater the possibility that the validity of self-reported gains can be improved through fairly minor changes in the questionnaires.

Predicting Factors Associated With Student Growth and Development

Higher education researchers are largely concerned about the degree to which the factors that predict student outcomes are contingent upon the form of assessment. The current findings suggest that regressions predicting student growth yield substantially different results depending upon how that growth is measured. In several pairs of regressions that contained numerous precollege characteristics and college experiences as predictors, about one

third of parallel regression coefficients (e.g., fraternity or sorority participation predicting critical thinking) differ significantly if the outcome is assessed via longitudinal gains versus self-reported gains. Moreover, fewer than one third of predictor variables that are significant with one form of assessment (e.g., self-reported gains) are also significant when using the other form of assessment. Stated differently, more than two thirds of the significant effects found in analyses predicting self-reported gains are *not* replicated when using the same independent variables to predict longitudinal gains. This incongruity between the two forms of assessment is particularly startling, because college administrators and practitioners often make decisions on the basis of these statistical analyses. Specifically, programs may be established to promote certain outcomes, because previous research has identified the efficacy of such programs. In other cases, certain groups of students who appear to exhibit less growth and development may be targeted for participation in (seemingly) beneficial experiences.

There are two types of error that might occur when using self-reported gains to gauge student outcomes. First, researchers might not identify certain experiences or characteristics that are associated with longitudinal gains. This constitutes a sizable problem, because more than two thirds of the significant predictors of longitudinal gains are not significant predictors of self-reported gains. This error might be seen as a missed opportunity to identify promising practices and behaviors. A more important error, though, occurs when certain experiences are positively related to self-reported gains, but the same experiences are unrelated or inversely related to longitudinal gains in student learning and development. In the present analyses, more than two thirds of significant predictors of self-reported gains fall into this category. This error is particularly pernicious, because it may result in institutions' spending time and money on programs that are not beneficial (at least in the manner that they are intended).

The divergence between self-reported and longitudinal gains (as assessed by comparing whether significant effects are apparent in both forms of assessment) varies greatly depending on whether the longitudinal outcome is assessed objectively or subjectively. For the objectively measured outcomes, virtually no correspondence exists between predictors of longitudinal and self-reported gains. For these outcomes, only one predictor is significantly related to longitudinal and self-reported gains in the same direction, whereas three predictors are significantly related to both forms of gains, but in the *opposite* direction. As a result, practitioners who want to know what experiences their students should engage in or avoid might take opposite actions, depending solely on the form of assessment that is used to inform their decisions. Conversely, the correspondence between predictors of longitudinal and self-reported gains for subjectively measured longitudinal outcomes is much greater. These two methods of assessment do not yield remarkably similar results, though: Only about half of the

predictors for self-reported gains are replicated for longitudinal gains, and vice versa.

At first glance, this pattern seems to contradict the pattern for the formal interaction analyses (Cohen et al., 2003), in which significant differences between regression coefficients were equally prevalent for objective and subjective longitudinal outcomes. This apparent discrepancy can be explained, at least in part, by the smaller sample size for the objective outcomes. As noted earlier, about half of all participants completed the CAAP Critical Thinking module, whereas the other half completed the DIT2. To ensure that the regressions compared the same groups of students, the regressions for self-reported gains included only students who had also completed the relevant objective assessment. The reduced statistical power from this smaller sample likely resulted in relatively fewer significant differences between regression coefficients and in fewer regression coefficients that differed significantly from zero. Therefore, when combining the results of the correlational analyses with the two comparisons of the regression analyses, the most plausible integration of these findings is the following: (a) The actual correspondence between self-reported and longitudinal gains is lower when the longitudinal measures are objective (as opposed to subjective), and (b) differences in sample size across the analyses contribute to a masking of this disparity in the formal interaction analyses and an overestimate of this disparity in the comparison of significant regression results.

It is unclear whether the greater correspondence for understanding oneself and diverse others relates to the form of the assessment (subjective versus objective) or to the type of construct that is being measured. Both of the objective longitudinal measures in this study captured skills and tendencies related to thinking and reasoning, whereas both subjective longitudinal measures gauged understanding of self and others. Therefore, one cannot conclusively disentangle the form of assessment and the type of outcome. It would be fruitful in future research to compare self-reported gains to objective, longitudinal assessments of noncognitive attributes (e.g., intercultural development or leadership skills) and to subjective, longitudinal measures of cognitive attributes (e.g., critical thinking, knowledge of academic subject areas).

Finally, from a measurement perspective, some scholars have suggested that virtually all of the variance in pretest-posttest measures of change can be attributed to error, which would mean that two-wave data may be so unreliable as to provide very little useful information (e.g., Baird, 1988). However, others argue that these concerns are misplaced or are relevant only in certain circumstances (Rogosa et al., 1982; Rogosa & Willett, 1983; Zimmerman, 2009). Specifically, if residual change scores for student learning consisted almost entirely of error, then very few independent variables within a multiple regression analysis would be significantly associated with the dependent variable. However, the longitudinal results in the appendix strongly suggest

that this is not the case in the current study. In fact, the number of significant regression coefficients is virtually identical for analyses predicting self-reported gains and for those predicting longitudinal gains. For the four regression analyses predicting longitudinal outcomes, over one third (36%) of the independent variables are significantly associated with the relevant dependent variable. Clearly, having a substantial number of significant predictors is possible only if the measures of change are reasonably reliable.

Conclusion

In sum, students' estimates of self-reported gains may not accurately reflect longitudinal gains, regardless of whether the longitudinal measures are objective or subjective. Although it is faster, easier, and less expensive to assess self-reported gains than longitudinal gains, the use of these cross-sectional assessments may lead researchers to draw erroneous conclusions about student learning and development; in other words, the potential problems with validity may outweigh the benefits of expediency. Self-reported gains can lead administrators and practitioners not only to mistakenly endorse practices and programs that have no true longitudinal impact, but also to mistakenly reduce or eliminate funding for programs that actually yield longitudinal improvements.

However, this validity issue does not imply that self-reported gains do not provide any information of interest. Perceived gains in learning and development are positively associated with students' satisfaction with college (Pike, 1993; Terenzini, Pascarella, & Lorang, 1982), because students' perceptions of having gained something useful arguably constitute an important domain of college satisfaction. It would certainly be problematic for any college or university if a large proportion of its students felt that they had not gained anything from attending college, regardless of any actual gains they had achieved. Thus, it seems quite reasonable to ask students about self-reported gains to provide useful information about students' college experience. Ironically, though, these self-report measures may be more effective at gauging student satisfaction or other student perceptions than the aspects of learning and development that they are purported to measure.

Future studies should explore the generalizability of the current findings. Because Pike (1999) showed that halo effects are greater among freshman than seniors, the correspondence between self-reported gains and longitudinal gains may be greater among more advanced undergraduates. Moreover, certain groups of students may be more adept at estimating their own gains than are other students. Some evidence suggests that students with high cognitive ability are more adept at estimating their test performance than low-ability students (Truxello, Seitz, & Bauer, 2008); this self-awareness may extend to judgments of self-reported gains in college. Furthermore, analyzing a greater number of outcomes would provide

additional support for the utility (or disutility) of self-reported gains as proxies for longitudinal gains. Although the findings for the current sample seem quite clear, additional evidence is needed to understand the conditions in which students are able (or unable) to judge their own development.

Appendix

Standardized Regression Coefficients Predicting Self-Reported and Longitudinal Gains

	Critical Thinking			Personal Code of Ethics		
Independent Variable	Self-Report Gains	Longitudinal Measures	Coef Diff	Self-Report Gains	Longitudinal Measures	Coef Diff
First generation	−.038	−.013		−.041	**−.041**	
Male	.004	−.026		**−.076**	−.026	
Black/African American	.025	−.015		−.017	−.032	
American Indian/ Native American	−.038	−.012		−.003	−.025	
Asian/Pacific Islander	**−.078**	−.004	*	.010	−.028	
Latino/Hispanic	**.099**	**−.043**	*	**.051**	−.001	
Race not reported	−.015	.001		−.013	−.013	
Low income	.003	−.029		.038	**−.046**	*
High income	−.038	.009		.052	**−.054**	*
Income not reported	−.045	.027	*	.041	**−.042**	*
Age 20+ in freshman year	**.230**	**−.064**	*	−.028	**.080**	*
HS GPA B average	−.001	−.015		.006	**−.101**	*
HS GPA C average	.001	**−.043**		**−.062**	−.017	
HS racial composition	.012	−.024		.018	.015	
Degree aspirations	**.054**	**.071**		−.009	−.006	
Academic motivation at T1	.052	.009		.022	.002	
Research university	.042	.000		.042	.024	
Regional university	−.042	−.030		−.026	−.005	
Community college	−.066	−.012		−.039	**−.131**	
Fraternity or sorority member	.057	**−.103**	*	**−.061**	.030	*
Varsity athlete	.024	**−.036**		.010	−.009	
Living in residence halls	.038	−.020		−.067	.015	
Living in Greek housing	.004	**.036**		.055	−.027	*
Work on campus 1–15 hr/wk	−.011	.026		.041	−.006	
Work on campus 16+ hr/wk	**−.056**	.003	*	**−.059**	−.030	
Work off campus 1–15 hr/wk	.017	−.028		−.003	**−.061**	
Work off campus 16+ hr/wk	.030	**−.104**	*	**.089**	−.018	*
Cocurricular activities	−.021	.023		**.103**	−.014	*
Relaxing and socializing	.001	.011		.040	−.007	
Drink alcohol 1–2 times/wk	**−.078**	.000	*	−.020	**−.058**	
Drink alcohol 3+ times/wk	.028	**−.036**	*	−.013	**−.072**	
Positive diversity experiences	.038	**.046**		**.069**	.017	
Anxious diversity experiences: rare	**−.052**	−.014		−.020	−.031	

(continued)

Appendix **(continued)**

Independent Variable	Critical Thinking			Personal Code of Ethics		
	Self-Report Gains	Longitudinal Measures	Coef Diff	Self-Report Gains	Longitudinal Measures	Coef Diff
Anxious diversity experiences: common	.000	**−.053**		.050	**−.098**	*
Diversity courses: 1	.032	.007		.003	.037	
Diversity courses: 2	**.071**	.007	*	.028	.027	
Diversity courses: 3 or more	.009	−.020		.053	.026	
Total courses	**.056**	.005		.012	**−.057**	*
Frequency of faculty contact	**.092**	**−.076**	*	**.171**	.030	*
In class challenge	**.224**	.015	*	**.125**	.025	*
Teaching clarity and organization	**.163**	.013	*	**.150**	.037	*
Prompt faculty feedback	.048	**.046**		.041	.035	
Longitudinal measure at T1		**.702**			**.603**	
Adjusted R^2	.248	.694		.192	.567	

Note. Regression coefficients that are significantly different from zero ($p < .05$) are in bold. An asterisk in the "Coef Diff" column means that the relevant pair of regression coefficients (e.g., fraternity or sorority member predicting longitudinal critical thinking gains and self–reported critical thinking gains) are significantly different from each other ($p < .05$). Coef Diff = coefficient difference; HS = high school; GPA = grade point average; T1 = Time 1; wk = week.

Independent Variable	Understanding People From Different Racial and Ethnic Backgrounds			Understanding Oneself		
	Self-Report Gains	Longitudinal Measures	Coef Diff	Self-Report Gains	Longitudinal Measures	Coef Diff
First generation	.010	**−.038**		**−.042**	**−.057**	
Male	−.009	**−.034**		**−.075**	−.029	
Black/African American	−.010	−.006		**−.057**	−.004	*
American Indian/ Native American	.003	**.056**	*	−.031	.020	*
Asian/Pacific Islander	.003	.004		−.014	−.030	
Latino/Hispanic	**.059**	−.004	*	.034	.028	
Race not reported	**.039**	−.012	*	−.008	**.030**	
Low income	**.143**	**.049**	*	**.126**	.029	*
High income	.010	.005		−.008	.010	
Income not reported	.005	.017		−.014	−.013	
Age 20+ in freshman year	**.084**	.033		**.062**	**.085**	
HS GPA B average	−.035	−.008		**−.065**	−.006	*

(continued)

Appendix **(continued)**

Independent Variable	Understanding People From Different Racial and Ethnic Backgrounds			Understanding Oneself		
	Self-Report Gains	Longitudinal Measures	Coef Diff	Self-Report Gains	Longitudinal Measures	Coef Diff
HS GPA C average	−.031	**−.054**		**−.065**	−.026	
HS racial composition	.034	−.018	*	−.021	.015	
Degree aspirations	−.024	−.010		**−.094**	−.005	*
Academic motivation at T1	−.024	−.016		.010	−.026	
Research university	**.065**	.037		.003	**.077**	*
Regional university	.017	.019		.002	−.015	
Community college	**.064**	**.062**		.033	.006	
Fraternity or sorority member	**.104**	**.093**		**.048**	**.071**	
Varsity athlete	**.059**	−.011	*	.010	.007	
Living in residence halls	−.041	−.039		.015	**.052**	
Living in Greek housing	**−.072**	**−.055**		−.023	.017	
Work on campus 1–15 hr/wk	−.018	.002		.002	−.029	
Work on campus 16+ hr/wk	−.024	−.030		**−.048**	.013	*
Work off campus 1–15 hr/wk	.027	−.019		.014	**.066**	*
Work off campus 16+ hr/wk	**.047**	**.079**		.016	**.091**	*
Cocurricular activities	**−.046**	−.026		.017	**.038**	
Relaxing and socializing	.006	−.024		.016	−.016	
Drink alcohol 1–2 times/wk	.008	**−.036**		.027	**−.036**	*
Drink alcohol 3+ times/wk	.004	**−.060**	*	−.032	−.005	
Positive diversity experiences	**.308**	**.180**	*	**.093**	**.079**	
Anxious diversity experiences: rare	−.018	.018		−.025	**−.060**	
Anxious div experiences: common	.020	**−.051**	*	.024	**−.122**	*
Diversity courses: 1	−.034	.004		−.031	.004	
Diversity courses: 2	**.046**	**.042**		.007	.024	
Diversity courses: 3 or more	**.084**	**.037**		.009	−.012	
Total courses	**−.046**	−.004		−.024	.017	
Frequency of faculty contact	**.075**	.005	*	**.136**	**.040**	*
In-class challenge	**.130**	**.053**	*	**.162**	**.077**	*
Teaching clarity and organization	**.156**	**.126**		**.146**	**.086**	*
Prompt faculty feedback	**−.062**	.013	*	−.029	−.004	
Longitudinal measure at T1		**.454**			**.556**	
Adjusted R^2	.244	.384		.169	.443	

Note. Regression coefficients that are significantly different from zero ($p < .05$) are in bold. An asterisk in the "Coef Diff" column means that the relevant pair of regression coefficients (e.g., fraternity or sorority member predicting longitudinal critical thinking gains and self–reported critical thinking gains) are significantly different from each other ($p < .05$). Coef Diff = coefficient difference; HS = high school; GPA = grade point average; T1 = Time 1; wk = week.

Note

I would like to thank the Center of Inquiry in the Liberal Arts at Wabash College for the use of its data.

References

ACT. (1991). *CAAP technical handbook*. Iowa City, IA: Author.

Anaya, G. (1999). College impact on student learning: Comparing the use of self–reported gains, standardized test scores, and college grades. *Research in Higher Education*, *40*, 499–526.

Aronson, E., Ellsworth, P. C., Carlsmith, J. M., & Gonzales, M. H. (1989). *Methods of research in social psychology* (2nd ed.). New York: McGraw-Hill.

Astin, A. W. (1993). *What matters in college?* San Francisco: Jossey-Bass.

Astin, A., Astin, H., Boatsman, K., Bonous-Hammarth, M., Chambers, T., Goldberg, S., et al. (1996). *A social change model of leadership development: Guidebook (Version III)*. Los Angeles: University of California at Los Angeles, Higher Education Research Institute.

Baird, L. L. (1976). *Using self-reports to predict student performance* (Research Monograph No. 7). Princeton, NJ: College Board.

Baird, L. L. (1988). Value added: Using student gains as yardsticks of learning. In C. Adelman (Ed.), *Performance and judgment: Essays on principles and practice in the assessment of student learning* (pp. 205–216). Washington, DC: U.S. Department of Education, Office of Educational Research and Improvement.

Bebeau, M., & Thoma, S. (2003). *Guide for the DIT-2*. Minneapolis: University of Minnesota, Center for the Study of Ethical Development.

Berdie, R. F. (1971). Self-claimed and tested knowledge. *Educational and Psychological Measurement*, *31*, 629–636.

Bowman, N. A. (in press-a). Dissonance and resolution: The non-linear effects of diversity courses on well-being and orientations toward diversity. *Review of Higher Education*.

Bowman, N. A. (in press-b). The development of psychological well-being among first-year college students. *Journal of College Student Development*.

Bowman, N. A., & Hill, P. L. (2009). *Social desirability and self-reported gains*. Unpublished data, University of Notre Dame.

Bowman, N. A., & Seifert, T. (2009). *Can students accurately assess what affects their learning and development?* Manuscript submitted for publication.

Burke, J. C. (Ed.). (2005). *Achieving accountability in higher education: Balancing public, academic, and market demands*. San Francisco: Jossey-Bass.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression correlation analysis for the behavioral sciences* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum.

Cole, J., & Gonyea, R. (2008, November). *Accuracy of self-reported SAT and ACT scores: Implications for research*. Paper presented at the annual meeting of the Association for the Study of Higher Education, Jacksonville, FL.

Conway, M., & Ross, M. (1984). Getting what you want by revising what you had. *Journal of Personality and Social Psychology*, *47*, 738–748.

Cooper, W. (1981). Ubiquitous halo. *Psychological Bulletin*, *90*, 218–244.

Cronbach, L. J., & Furby, L. (1970). How should we measure "change"—or should we? *Psychological Bulletin*, *74*, 68–80.

DuBois, P. H. (1957). *Multivariate correlational analysis*. New York: Harper.

Dumont, R. G., & Troelstrup, R. L. (1980). Exploring relationships between objective and subjective measures of instructional outcomes. *Research in Higher Education*, *12*, 37–51.

Fuertes, J., Miville, M., Mohr, J., Sedlacek, W., & Gretchen, D. (2000). Factor structure and short form of the Miville-Guzman Universality-Diversity Scale. *Measurement and Evaluation in Counseling and Development*, *33*, 157–169.

Gilbert, D. (2007). *Stumbling on happiness*. New York: Vintage.

Goethals, G. R., & Reckman, R. F. (1973). The perception of consistency in attitudes. *Journal of Experimental Social Psychology*, *9*, 491–501.

Gonyea, R. M. (2005). Self–reported data in institutional research: Review and recommendations. In P. D. Umbach (Ed.), *New directions for institutional research* (*Vol. 127*, pp. 73–89). San Francisco: Jossey-Bass.

Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology*, *5*, 213–236.

Kuh, G. D., Hayek, J. C., Carini, R. M., Ouimet, J. A., Gonyea, R. M., & Kennedy, J. (2001). *NSSE technical and norms report*. Bloomington: Indiana University, Center for Postsecondary Research and Planning.

Long, J. S. (1997). *Regression models for categorical and limited dependent variables*. Thousand Oaks, CA: Sage.

Lord, F. M. (1956). The measurement of growth. *Educational and Psychological Measurement*, *16*, 421–437.

Markus, G. B. (1986). Stability and change in political attitudes: Observed, recalled and explained. *Political Behavior*, *8*, 21–44.

McFarland, C., & Ross, M. (1987). The relation between current impressions and memories of self and dating partners. *Personality and Social Psychology Bulletin*, *13*, 228–238.

McNemar, Q. (1958). On growth measurement. *Educational and Psychological Measurement*, *18*, 47–55.

National Center for Public Policy and Higher Education. (2004). *Measuring up 2004: The state-by-state report card for higher education*. San Jose, CA: Author.

Miville, M., Gelso, C., Pannu, R., Liu, W., Touradji, P., Holloway, P., et al. (1999). Appreciating similarities and valuing differences: The Miville-Guzman Universality-Diversity Scale. *Journal of Counseling Psychology*, *46*, 291–307.

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we know: Verbal reports on mental processes. *Psychological Review*, *84*, 231–259.

Nunnally, J. C. (1982). The study of human change: Measurement, research strategies, and methods of analysis. In B. B. Wolman (Ed.), *Handbook of developmental psychology* (pp. 133–148). Englewood Cliffs, NJ: Prentice Hall.

Pace, C. (1984). *Measuring the quality of college student experiences: An account of the development and use of the College Student Experiences Questionnaire*. Los Angeles: Higher Education Research Institute.

Pace, C., & Friedlander, J. (1982). The meaning of response categories: How often is occasionally, often, and very often? *Research in Higher Education*, *17*, 267–281.

Pascarella, E. T. (2001). Using student self-reported gains to estimate college impact: A cautionary tale. *Journal of College Student Development*, *42*, 488–492.

Pascarella, E., Bohr, L., Nora, A., & Terenzini, P. (1995). Cognitive effects of 2-year and 4-year colleges: New evidence. *Educational Evaluation and Policy Analysis*, *17*, 83–96.

Pascarella, E., Edison, M., Nora, A., Hagedorn, L., & Terenzini, P. (1996). Influences on students' openness to diversity and challenge in the first year of college. *Journal of Higher Education*, *67*, 174–195.

Pascarella, E. T., & Terenzini, P. T. (2005). *How college affects students: Vol. 2. A third decade of research*. San Francisco: Jossey-Bass.

Pike, G. R. (1993). The relationship between perceived learning and satisfaction with college: An alternative view. *Research in Higher Education*, *34*, 23–40.

Pike, G. R. (1994, November). *The relationship between self-report and objective measures of student achievement*. Paper presented at the annual meeting of the Association for the Study of Higher Education, Tucson, AZ.

Pike, G. R. (1995). The relationship between self-reports of college experiences and achievement test scores. *Research in Higher Education*, *36*, 1–21.

Pike, G. R. (1996). Limitations of using students' self-reports of academic development as proxies for traditional achievement measures. *Research in Higher Education*, *37*, 89–114.

Pike, G. R. (1999). The constant error of the halo in educational outcomes research. *Research in Higher Education*, *40*, 61–86.

Pohlmann, J., & Beggs, D. (1974). A study of the validity of self-reported measures of academic growth. *Journal of Educational Measurement*, *11*, 115–119.

Pronin, E., & Kugler, M. B. (2007). Valuing thoughts, ignoring behavior: The introspection illusion as a source of the blind spot bias. *Journal of Experimental Social Psychology*, *43*, 565–578.

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

Reiter, H. L. (1980). The perils of partisan recall. *Public Opinion Quarterly*, *44*, 385–388.

Rest, J., Narvaez, D., Thoma, S., & Bebeau, M. (1999). DIT2: Devising and testing a revised instrument of moral judgment. *Journal of Educational Psychology*, *91*, 644–659.

Rogosa, D. R., Brandt, D., & Zimowski, M. (1982). A growth curve approach to the measurement of change. *Psychological Bulletin*, *90*, 726–748.

Rogosa, D. R., & Willett, J. B. (1983). Demonstrating the reliability of the difference score in the measurement of change. *Journal of Educational Measurement*, *20*, 335–343.

Ross, M. (1989). Relation of implicit theories to the construction of personal histories. *Psychological Review*, *96*, 341–357.

Ryff, C. D. (1989). Happiness is everything, or is it? Explorations on the meaning of psychological well–being. *Journal of Personality and Social Psychology*, *57*, 1069–1081.

Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence*. New York: Oxford University Press.

Terenzini, P. T., Pascarella, E. T., & Lorang, W. G. (1982). An assessment of the academic and social influences on freshman year educational outcomes. *Review of Higher Education*, *5*, 86–109.

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

Truxello, D. M., Seitz, R., & Bauer, T. N. (2008). The role of cognitive ability in self-efficacy and self-assessed test performance. *Journal of Applied Social Psychology*, *38*, 903–918.

Tucker, L. R., Damarin, F., & Messick, S. A. (1966). A base-free measure of change. *Psychometrika*, *31*, 457–473.

Tyree, T. (1998). *Designing an instrument to measure socially responsible leadership using the social change model of leadership development*. Unpublished doctoral dissertation, University of Maryland–College Park.

Umbach, P. D., & Kuh, G. D. (2006). Student experiences with diversity at liberal arts colleges: Another claim for distinctiveness. *Journal of Higher Education*, *77*, 169–192.

University of Minnesota, Center for the Study of Ethical Development. (n.d.). *New index (N2)*. Retrieved January 14, 2008, from http://www.centerforthestudyofethicaldevelopment.net/New%20Index.htm

Upcraft, M. L., Gardner, J. N., & Barefoot, B. O. (Eds.). (2004). *Challenging and supporting the first-year student*. San Francisco: Jossey-Bass.

U.S. Department of Education. (2006). *A test of leadership: Charting the future of U.S. higher education*. Washington, DC: Author.

Vontress, C. E. (1986). Social and cultural foundations. In M. D. Lewis, R. Hayes, & J. A. Lewis (Eds.), *An introduction to the counseling profession* (pp. 215–250). Itasca, IL: Peacock.

Vontress, C. E. (1988). An existential approach to cross-cultural counseling. *Journal of Multicultural Counseling and Development*, *16*, 78–83.

Werts, C. E., & Linn, R. L. (1970). A general linear model for studying growth. *Psychological Bulletin*, *73*, 17–22.

Whitt, E. J., Edison, M., Pascarella, E. T., Nora, A., & Terenzini, P. T. (1999). Interactions with peers and objective and self-reported cognitive outcomes across 3 years of college. *Journal of College Student Development*, *40*, 61–78.

Wilson, T. D. (2002). *Strangers to ourselves*. Cambridge, MA: Belknap Press of Harvard University Press.

Yalom, I. D. (1985). *The theory and practice of group psychotherapy* (2nd ed.). New York: Basic Books.

Zimmerman, D. W. (2009). The reliability of difference scores in populations and samples. *Journal of Educational Measurement*, *46*, 19–42.

Manuscript received November 18, 2008
Final revision received September 8, 2009
Accepted September 10, 2009