CHAPTER ONE

Studying College Outcomes in the 2000s

Overview and Organization of the Research

The purpose and value of higher education are under fire. As national confidence in the aims of higher education and the subsequent value of degree attainment erode (see Arum & Roksa, 2011, 2014), scholars interested in college and its influence on students are faced with a series of emergent challenges, ranging from the decoupling of the once tightly held belief that participation in higher education was the primary means for learning and thus social mobility to ontological questions about learning itself: Is learning about making money? Why is learning important if it does not lead to financial gain? Indeed, some students are paid to forgo college-going for pursuing entrepreneurial start-ups. Peter Thiel, founder of the Thiel Foundation, an organization that pays up-and-coming entrepreneurs to leave formal education, noted, "University administrators are the equivalent of mortgage brokers, selling you a story that you should go into debt massively, that it's not a consumption decision, it's an investment decision. Actually, no, it's a bad consumption decision. Most colleges are four-year parties" (Jenkins, 2010, p. A.13). This comment exemplifies the emergent American learning conundrum: How utilitarian and pragmatic does learning need to be in order to hold value in and to American society? Is higher education an investment in one's future or a consumable good of questionable value?

In light of these questions and challenges, educators from across disciplines are designing and executing rigorous college impact studies that draw on the scholarly work of generations past to further develop a robust understanding of college as critical to not only the learning enterprise but to other

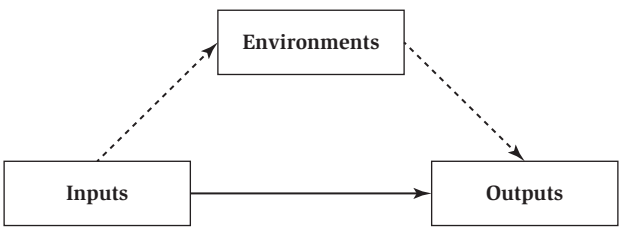
social and economic factors as well. Rather than shy away from the difficulties of studying outcomes that many think are ineffable and even irrelevant, these scholars are approaching the study of college impact with the thoroughness needed to appraise historic claims regarding the roles and purposes of higher education and the innovation needed to tackle questions once believed too challenging to address. Our aim in this volume is not to provide silver-bullet answers to these pressing and difficult questions but to review carefully the evidence for helping educators make claims about college and its impact on students.

Conceptually, this volume is based on Astin's (1984) framework for understanding how college affects students. Put simply, this framework deconstructs the college experiences into three discrete categories: inputs, environments, and outcomes. Inputs include demographic characteristics, academic preparedness, and predispositions that students bring with them to campus (e.g., race, high school grade point average, SAT scores, degree aspirations, and academic motivation, to name a few). Environments include, but are not limited to, institutional cultures and climates and specific educational experiences designed to shape students in some meaningful way. Outcomes relate to the attitudes (e.g., student satisfaction), aptitudes (e.g., critical thinking), and behaviors (e.g., departure) that students exhibit as a result of going to college.

Of critical importance to this review is how these categories work together to explain college and its effects on students. When organizing studies, we based our review on two relationships: that which we call "general" to describe the relationship between environments and outcomes (i.e., how exposure to and participation in college generally affect all college students) and that which we call "conditional" to underscore the relationship between environments and outcomes as it relates to student inputs (i.e., how exposure to and participation in college experiences affect students differentially based on students' input characteristics).

Figure 1.1 is a graphic representation of Astin's model. These relationships are represented by the dotted arrows in the figure. Note that the relationship between inputs and outcomes is displayed with a solid arrow to reflect that the review did not focus on studies that examined this relationship.

Figure 1.1 Astin's Framework (1984) for Understanding College and Its Influence on Students



With this conceptual map as our guide, we used the organizational framework developed by Ernest Pascarella and Patrick Terenzini (1991, 2005) to synthesize the many thousands of empirically based articles designed to better understand college and its relationship to student outcomes. Building on the generous work of many scholars and employing the organizational framework used in the previous two volumes of this work, we addressed each of these six issues for each set of outcomes: the development of verbal, quantitative, and subject matter competence; cognitive skills and intellectual growth; psychosocial change; attitudes and values; moral development; educational attainment and persistence; career and economic impacts of college; and quality of life after college. Specifically, we adopted Pascarella and Terenzini's six-question framework for organizing the literature within each chapter. This framework, which developed out of previous work by G. Gurin (1971), Nucci and Pascarella (1987), and Pascarella (1985), asks six basic questions that serve as the organizing feature for each chapter:

1. What evidence is there that individuals change during the time in which they are attending college?
2. What evidence is there that change or development during college is the result of college attendance?
3. What evidence is there that attending different kinds of postsecondary institutions have a differential influence on student change and development during college?
4. What evidence exists that engaging in different experiences in the same institution are associated with student change and development during college?
5. What evidence is there that the collegiate experience produces conditional, as opposed to general, effects on student change or development?
6. What are the long-term effects of college?

Question 1, which we sometimes refer to by the shorter phrasing of "change during college," refers to whether change occurred while students were exposed to postsecondary education. Question 2, regarding the net effects of college, focuses on whether the change is attributed to postsecondary exposure, as opposed to precollege characteristics, maturation, or other noncollege experiences. Question 3, between-college effects, explores the degree to which institutional conditions (e.g., size, control, geographic location) or organizational characteristics (e.g., average level of peer cognitive development, whether the school is bureaucratic or collegial, structural diversity of the faculty) have an influence on the learning and development of the student. Question 4, within-college effects, summarizes the articles that address student change as a function of exposure to or participation in specific collegiate

experiences. Question 5, conditional effects of college, gauges the extent to which the relationship between student change and any given college experience differs based on student characteristics, such as race, gender, or academic major. Question 6, long-term effects of college, addresses the duration or permanence of the college influence based on student's postcollege activities, attitudes, beliefs, and behaviors. Table 1.1 summarizes the framework used to guide this review.

Table 1.1 Overview of Review Framework

	Conceptual Orientation	*Shorthand*	*Description*	*Example Research Question*
Question 1	General	Change during college	Whether change occurred while in college	Do college students demonstrate gains in moral development during college?
Question 2	General	Net effects of college	Whether the change can be attributed to college-going, as opposed to maturation, for example	Does moral development occur as a result of college-going, accounting for a host of potential confounding influences?
Question 3	General	Between-college effects	Whether the change can be explained by institutional conditions, organizational characteristics, and/or peer socialization	What role does institutional type and public (versus private) control play in shaping students' moral development?
Question 4	General	Within-college effects	Whether the change can be explained by exposure to and participation in specific educational experiences	How does participation in a service-learning experience influence moral development?

(continued)

Table 1.1 Overview of Review Framework (continued)

	Conceptual Orientation	*Shorthand*	*Description*	*Example Research Question*
Question 5	Conditional	Conditional effects of college	Whether the change that occurs as a result of participation in any given college experience differs based on student inputs such as race, gender, living status	Does the relationship between participating in a service-learning experience and moral development differ between residential and commuter students?
Question 6	General	Long-term college effects	If the changes due to college are sustained after graduation	Are the moral development gains made during college sustained beyond graduation?

Building on these six questions used to frame the literature, we organized studies within each question based on themes emerging from the articles reviewed for each chapter. This decision came from our collective value to review articles in the spirit in which they were written. We wanted to stay as close to the authors' intentions as possible. Of course, this decision produced a distinctive set of challenges regarding structural continuity across chapters. For example, for the within-college effects section of each chapter, some authors studied honors colleges while others did not; some articles discussed interactional diversity while others examined quality of diversity interaction or nonclassroom-based diversity peer interaction; some studies investigated work on campus while others reflected interest in part-time employment. Given these and the many more examples of themes that emerged from the studies themselves, we chose not to try to force articles into categories for the sake of consistency across chapters; rather, we let the literature base specific to the chapter's focus inform the organization of that chapter, at least to some degree. Similarly, a number of outcomes examined in the literature do not fit neatly and discretely into one chapter or another. For example, one could argue that a self-reported gain in general education is a measure of the general skills, like

verbal and quantitative competence, that students develop in college; a parallel argument could advance that this is a reflection of students' academic and intellectual self-concept. We shaped our review with the authors' intentions in mind while recognizing the potential overlap between researchers' definitions of outcomes and the conceptual outcome framework used in the book.

HOW THE LITERATURE HAS CHANGED

Since the first volume of this book was published (Pascarella & Terenzini, 1991), terms and definitions continue to change at a remarkable pace. Words like *how*, *college*, *affect*, and *students* have taken different meanings in the higher education research context since the beginning of the century. For example, with the advent and momentum of computer-mediated distance education, for-profit institutions, and massive open online courses (MOOCs), "college," as we know it today, has moved beyond chartered boundaries to be more inclusive than ever before. As the college experience extends its reach, its "effects" are more difficult than ever to ascertain; indeed, new methods are continuously being offered and refined to manage issues with studying students in their natural, albeit nonrandom, learning environments. Finally, there are many definitions of *student*: Is a student someone enrolled in one MOOC? A degree program? A certification program? A GRE course offered at an institution? Since these words—*how*, *college*, *affects*, *students*—underlie the syntheses provided in this volume, we consider the meaning of each to discuss trends in the literature since the previously published volume (Pascarella & Terenzini, 2005) and to note where this volume departs from those previously written.

How: Changes in the Ways College May Have Influenced Students

Based on the rich 30-plus years of research linking college-going to development and change across a variety of domains, scholars have moved from empiricism to assumption: rather than question if college-going has an influence on students, scholars assume that the relationship exists and subsequently focus on investigating the specific practices and psychological mechanisms responsible for student change. In other words, since the previous volume, scholars are asking more questions about *why* college affects students than *if* college affects students. Such a trend presented a particular set of challenges for this review, including how to speak to change over time with very few longitudinal designs that tracked students over multiple time points, address the net effects of college-going as so few studies compared students to their peers who did not attend college, and evaluate and summarize the theoretical claims across the empirical studies.

Another disruption in our understanding of the "how" comes in the form of the many competing approaches designed to interrogate college and its effects on students. Clearly, the frameworks researchers use to position their inquiries

into college and its effects on students play a role in the questions researchers ask and their subsequent choices regarding data and methods of analysis. Like previous volumes, we overrepresented studies that adopted a positivist or post-positivist paradigm for asking questions about college and its effects on students. Perhaps this overrepresentation is an artifact of the types of studies that are published in most peer-reviewed journals. Alternatively, the overrepresentation may result from our decision to review only the studies that measured the relationship between college and its effects on students. Either way, we own that our collective perspective also informed our approach to this review from its conceptualization to its organization.

Like its predecessors, this review theoretically draws from many disciplines for studies and explanations of the relationship between college and students. Each chapter tended to rely on certain disciplinary perspectives based on the material published on the chapter's subject; for example, chapters focused on outcomes with developmental dimensions often drew from psychology, while those that emphasized earnings were based largely in economic studies. Due to the distinctiveness that each theoretical perspective offered for making meaning of empirical findings, we decided to discuss the theoretical underpinnings of each outcome within its related chapter and to provide a review of only the theories that this volume's researchers most often used to frame their inquiries. To be clear, *this volume is not intended to cover, or even mention, all theories, conceptualizations, and frameworks that have informed higher education scholarship since its inception*. Instead, we provided brief overviews of these elements as contextual support for conclusions offered by the authors of the articles reviewed in this volume. Placing the theoretical overview section within each chapter marks a departure from previous efforts where an overview of guiding theory for all chapters was offered in Chapter 2.

Turning to our approach to the literature review, we gave greater weight to issues of design over analysis when making decisions about article inclusion and subsequent exposition (see Rubin, 2008). When compared to articles that used cross-sectional designs, articles that included research designs that were longitudinal and included a pretest and a comparison group, or that were quasi-experimental (e.g., propensity score or regression discontinuity) were relied on more heavily as evidence of particular empirical trends. Due to our collective commitment to help readers understand the criteria we used for reviewing and ultimately including articles in this volume and marking a departure from previous volumes, we included a detailed methodological overview as a methodological appendix in this review.

Of course, the issue of survey fatigue also has made making claims about college and its effect on students more problematic. Technological advances in data collection and control have equipped scholars and institutional researchers with the infrastructure needed to support more institution-specific data collection efforts. Although we encourage these practices as they lead to data-driven decisions administrators can use to ameliorate institutional practices, we also

recognize that the proliferation of these data collection efforts makes multi-institutional research efforts more challenging. Survey fatigue presented another issue that complicated the "how" with regard to understanding college and its effects on students.

College

What is college? Since its inception, higher education in the United States has been in constant evolution. The particular sociohistoric and political location in which this volume was drafted marks no exception to this trend. However, in the 10 years since the previous volume was published, a number of developments have changed the way that many understand and relate to college and student experiences therein.

The term *college* is complicated. For example, in the United States, *college* could refer to higher education in general, a single institution within the higher education system (e.g., Pomona College), or a subunit within a larger university system (e.g., College of Business within the University of Iowa). In other countries, the term *college* carries different meanings, often reflecting each nation's interest in, values concerning, and organization of higher/postsecondary/tertiary education (Jones, 2012). Despite the challenges that accompany different interpretations of college, especially across national borders, we broadened the scope of this review to include relevant college impact research executed in Australia, New Zealand, the United Kingdom, and Canada. Marking a departure from previous efforts, the inclusion of studies from these countries as part of an expanded scope of the review reflects an acknowledgment that higher education has become much more internationalized since the previous volume's publication (Altbach & McGill Peterson, 2007; Guruz, 2008; Knight, 2008) and that much could be learned from understanding student experiences outside the United States. Acknowledging and appreciating the differences in these countries' respective approaches to higher education, we selected these nations based on their use of English as the primary language for instruction and research dissemination, as well as their historic grounding in the Oxford-Cambridge residential colleges model.

The technological movement has advanced the notion of college from being a context bound by geographic borders to one that is essentially borderless, with many individuals claiming student status without having set foot on a college campus (Selino, 2013). Indeed, even President Obama has enacted policies that challenge the notion of equating college with a degree, as involvement in at least one year has taken federal priority over four-year degree completion (e.g., Complete College America, 2011, 2012, 2013). With the increasingly widespread and mobile nature of Internet technologies and social media shifting the landscape for educational delivery, technology has complicated research on college and its effects on students by challenging assumptions that any scholar could ever isolate the effects of any measured experience on any student outcome.

The movement toward integration of the college experience has changed the research landscape, as evidence-based best practices (e.g., service-learning,

living-learning communities; see Kuh, 2008) often reflect integrated educational delivery models designed in an effort to educate the whole student. What is service-learning? What is a living-learning community? Are these academic, social, or functional experiences (see Milem & Berger, 1997)? To date, despite a robust research base on these topics, few practices have attained definitional consensus. As a result, the college experience itself has become harder to define, making the study of a presumed best practice for its influence on college student learning more challenging.

The changing nature of the peer environment has rendered historic higher education vernacular increasingly difficult to understand. For example, what do we mean by *college major*? A series of courses tightly threaded together by a common academic interest? A means for generating a pseudo-academic cohort effect by engaging students with common interests around a set of ideas presented sequentially in the curriculum? Another way of grouping students, similar to identity patterns based on social identity group organization or residence hall participation? Again, these questions provided some conceptual challenges to researchers interested in unpacking college experiences as a set of embedded peer networks and to us as we confronted some organizational obstacles in deciding where to discuss peer effects in each chapter.

Similar challenges emerged from studies that linked faculty practice to student outcomes. Who are the faculty who have the greatest impact on students? Are these adjunct faculty? Faculty who teach more courses? Faculty who engage students in undergraduate research opportunity programs? To complicate matters further, faculty practice sometimes is mediated fully through a particular delivery mechanism: authors may study an educational context (e.g., diversity course) for its association with a particular outcome without specifically examining the practice within that context. Given these and the many other issues that remain unmentioned, it is often difficult to draw conclusions about the potential impact of faculty behaviors on students.

Affects

Given the explosion of research on college and its effects on students over the past decade, the use of causal language has been increasingly scrutinized in making claims about college and its relationship to college student learning and development. In tandem with criticism about causal language, questioning such verbs as *affects* and, to some degree, *influences* (Swanson, 2010, 2012), many scientists have also questioned the use of the term *quasi-experimental*, even for research that uses longitudinal designs with control or comparison groups. Unless researchers can randomly assign students into a certain educational experience (i.e., experimental) or methodologically make adjustments to samples through the use of propensity scores or regression discontinuity (i.e., quasi-experimental), causal claims about college and its relationship to students must be made cautiously or, in some cases, not at all. The disruption concerning what constitutes a quasi-experimental design marks a point of departure in our synthesis of the literature when juxtaposed against

previous reviews. In addition to being more thoughtful in our use of terms like *affects* and *influences*, we were equally careful to use the term *quasi-experimental* only in studies with adjusted sample designs.

As with previous volumes, *affects* is a term reserved for studies that measure the relationship between college experiences and outcomes, not necessarily for studies that use college students as samples of convenience for examining relationships between certain phenomena or for scholars only interested in how outcomes differed among certain student characteristics, like race or high school achievement. In short, all of the studies reviewed for this volume involved researchers' empirical attempt to link educational experiences to student outcomes.

What do we mean by a college experience *affecting* students? Most of the studies reviewed for this volume used developmental language for making meaning of college and its impact on students: we use phrases like "helping students make cognitive gains," "more likely to demonstrate gains in pluralism orientations," and "make moral gains" as communicative proxies for college's impact on students. When the studies depart from developmental frames, we aimed to use the authors' voices to describe the kind of learning or achievement, if any, that occurs and its relationship to college-going. Examples include "helping students achieve outcomes related to critical thinking" to "outcomes with moral dimensions."

Students

Who a college student was, is, and is becoming plays a central role in framing this review. As stated in previous volumes, the demographic characteristics of college-going students continue to rapidly change, forcing us to reconsider the ways we have traditionally defined the college student. According to the U.S. National Center for Educational Statistics, the percentage of undergraduates of color has risen from 29.2% in 2000 to 39.7% in 2012 (U.S. Department of Education, National Center for Education Statistics, 2015). Given these shifting characteristics, especially as they relate to students' racial identities, comparisons between studies conducted in the 1990s with those reviewed in this volume must be interpreted cautiously.

Similarly, more international students are enrolling in U.S. higher education institutions than ever before (Institute of International Education, 2012). Institutions continue to expand their reach into international markets through strategic partnerships with global partners and increase revenue streams through recruiting more international students to campus in order to remain globally relevant and economically viable (Altbach & McGill Peterson, 2007; American University Office of Institutional Research and Assessment, 2014; Guruz, 2008; University of California Office of the President, 2015; University of Notre Dame, 2013). Given this increase of international students, college impact researchers are beginning to be more attentive to other variables (e.g., English

as a second language) that may exert influence on either the college experience or the student outcome.

Related to these complications are notions of multiple, intersecting identities for college students. With a greater number of students coming to college more cognizant of their multiple identities and/or more familiar with the lexicon used to describe intersecting identities, it is important to understand that the effects of "race" or "worldview" or "sexual orientation" may involve intersections across these characteristics. Moreover, we were cautious in our use of terms that sidestep these intersecting realities and tried to shy away from using terms like "controlling for race" because they, although technically correct, probably do not provide the most accurate picture of student experience. How can anyone really control for race? In addition, because the research reviewed was broadened to include studies outside the United States, it is necessary to be cognizant of how perceptions of identity are deeply rooted in the unique history, culture, and systemic social structure of the various international postsecondary contexts reviewed.

Another complication arises when trying to capture the experience of the traditional college student. Traditional-aged college students are now a minority of undergraduates in U.S. postsecondary education, as Pascarella and Terenzini (2005) predicted in the previous volume. In short, we attempted to include studies that spoke to the undergraduate experience of all students, regardless of age, college choice, degree aspiration, or preferred mode of educational delivery (e.g., online). In doing so, we hope to extend the reach of this volume to any person interested in undergraduate postsecondary education.

VOLUME 3: RESEARCH FROM THE 21ST CENTURY

This volume adheres closely to the guidelines provided in previous iterations of *How College Affects Students* (1991 and 2005). As such, we echo the sentiments expressed in 1991 and 2005, respectively. This book is an attempt to synthesize the college impact research evidence that has accumulated since the review period of the 2005 publication. At times, we relied on articles from previous decades to frame arguments made by the authors whose work is reviewed in this volume. This review covered articles written between 2002 and 2013. In addition, we included some articles published from 2014, depending on the time that the chapter was written. This approach was consistent with the previous volumes' presentation of the evidence.

In terms of focus, this book collected information from over 10,000 sources of literature. Of those pieces, 1,848 peer-reviewed articles served as the foundation for this synthesis. Unlike previous volumes, we chose not to include conference papers or dissertations due to the overwhelming number of quality-controlled research published over the past decade. Articles were

located in journals representing an array of audiences. Every article identified as relevant (i.e., it addressed some aspect of "college effects" on students) was initially reviewed and flagged for potential use for this review. In addition to this approach, we located articles through the use of search engines such as Google Scholar, ERIC, and PsycInfo, among others. In addition, we conducted a hand-search of general higher education journals (e.g., *Journal of Higher Education*, *Research in Higher Education*, *Review of Higher Education*, along with some other journals (e.g., *Journal of College Student Development*, *Review of Educational Research*, *Journal of College Student Retention*). Also, we conducted forward searches in Google Scholar to see who cited eligible articles. After articles were identified as relevant, we scoured that article's references as a means for tracking down other cited works germane for this review. Once the articles were compiled, they were then organized based on chapter focus. On completion of this step, articles were then systematically coded based on their fit within (and often across) the six-question framework offered by Pascarella and Terenzini (1991, 2005) as part of their syntheses and methodological quality.

Exemplary studies of college impact received greater weight in our review of the literature. Specifically, we placed an emphasis on studies that used research designs that permitted stronger causal conclusions (i.e., experimental, quasi-experimental, and nonexperimental with rigorous analytical controls), obtained multi-institutional samples, conducted multilevel analyses (when appropriate), explored direct and indirect effects (when appropriate), employed a longitudinal design, and used well-validated measures of outcomes and experiences. In subsequent chapters, we cite studies that contain a variety of methodological characteristics, but we generally describe the findings of stronger research in greater detail, and we use considerable caution when evaluating the results of studies that meet few of these criteria.

To provide readers context for understanding our approach to weighing the evidence provided in this volume, we offer some points about measuring and modeling the student outcomes represented and reviewed in Chapters 2 through 9 of this book. Although technical in some regards, this strategy enables readers to make meaning of the research designs and numbers derived for this volume. We begin with a brief discussion of the complexities involved with measuring student change as a result of exposure to and participation in postsecondary education. We then discuss issues of whether and when effects are practically meaningful, and we provide guidelines for making these decisions.

Measuring and Modeling Student Outcomes

The measurement of changes in student outcomes is more complicated than one might expect. Direct measures of change necessarily involve collecting data on the same students (or institutions) on two or more occasions in time and then comparing the outcomes at these different time points. However, longitudinal data collection (with or without random assignment) presents some logistical difficulties: (1) students' data from the pretest must be linked to their responses

on the posttest(s), which requires keeping track of students' personal information; (2) many students who completed the pretest may drop out, transfer, or simply not respond to the posttest; (3) collecting data multiple times requires more human and financial resources than conducting a single data collection; (4) the time between the pretest and posttest may be too short for the expected effect to occur; and (5) the primary results from longitudinal analyses cannot be determined until two or more waves of data collection have occurred. To alleviate these challenges, some college impact studies conduct a single cross-sectional assessment. This may be less problematic for outcomes that do not have a true pretest (e.g., college satisfaction, perceptions of campus climate), but this is certainly a concern for determining changes in cognitive, attitudinal, and psychosocial outcomes. Researchers who administer a single questionnaire often ask students for an estimate how much they have changed on a variety of outcomes, which serves as a proxy for longitudinal measures of growth.

Although college student self-reported gains are often interpreted as if they reflect changes in student outcomes over time (Gonyea & Miller, 2011), considerable evidence suggests that this is not the case. If these self-reports were accurate, there should be a strong correlation between students' self-reported gains on a particular outcome and longitudinal changes on a well-validated measure of that same outcome. Across a variety of outcomes, the correlations between longitudinal and self-reported gains on the same construct are consistently weak and are often not significantly different from zero (Bowman, 2010a, 2011b; Bowman & Brandenberger, 2010; Gosen & Washbush, 1999; Hess & Smythe, 2001). In addition, the variables that significantly predict longitudinal growth (e.g., college experiences, student demographics, institutional attributes) are often nonsignificant—and sometimes even significant in the opposite direction—when predicting self-reported gains for the same construct (Anaya, 1999; Bowman, 2010a; Bowman & Brandenberger, 2010; Porter, 2013). Earlier research has established consistent biases in self-reported growth among college students and older adults, such that 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; M. Ross, 1989). A meta-analytic review further suggests that people may be somewhat accurate in reporting their current knowledge, whereas they are highly inaccurate at reporting changes in knowledge over time (Sitzmann, Ely, Brown, & Bauer, 2010). In short, using student self-reported gains as a proxy for college impact may yield substantially flawed results.

That said, longitudinal studies that use objective assessments also face some difficulties for measuring changes in student outcomes. Perhaps the most important concern is students' effort on assessments that have substantial cognitive demands, such as critical thinking instruments. If students do not exert considerable effort, then the results of these assessments may be questionable. Indeed, providing monetary incentives for student performance results in

higher test scores than providing no incentive (for a meta-analysis, see Duckworth, Quinn, Lynam, Loeber, & Stouthamer-Loeber, 2011). In addition, both telling students that their test scores will be used as a means of assessing the quality of their institution or assessing their own skills result in better performance than telling students that their responses are simply part of a research project (Liu, Bridgeman, & Adler, 2012; also see Wise & DeMars, 2005). Although the impact of different motivational conditions is concerning, there are at least two ways in which these problems can be at least partially remedied. First, researchers or administrators can frame the purpose of the study carefully to increase student motivation (i.e., the results will be used not only for research purposes, but also for assessing individual students or the institution as a whole). Second, a variety of techniques are available for identifying unmotivated test takers, especially when the exams are administered via computer (e.g., Swerdzewski, Harmes, & Finney, 2011; Wise & Kong, 2005). Finding and removing unmotivated students from the sample can lead to more accurate conclusions about student performance.

Additional difficulties may occur when attempting to estimate the overall impact or net effect of college, since a decrease or increase in some outcome measure during the college years does not necessarily suggest that college contributed to that increase or decrease. To address this college impact question, one would need to explore whether changes in college students' outcomes differ from those of people who are the same age but not in college (which is why this book distinguishes between "change during college" and "net effects of college"). One example is particularly illustrative. A number of studies found a decline in religious behaviors during the college years, such as attending religious services, frequency of prayer, discussing religion, and perceiving oneself as religious (e.g., Bryant, Choi, & Yasuno, 2003; also see Pascarella & Terenzini, 1991). This result, which merely reflects change during college, can be interpreted as demonstrating that college attendance has a secularizing effect. However, Uecker, Regnerus, and Vaaler (2007) examined a large sample of young adults who did and did not attend college so that they could accurately assess the net effects of college. Both college students and noncollege adults declined on several measures of religiosity, but these decreases were actually greater among young adults who were not attending college. Thus, simply exploring change during college as a proxy for net effects can yield conclusions that are exactly the opposite of those obtained when using appropriate noncollege comparison groups.

A final issue with measuring and modeling college outcomes is that some dependent variables are not continuous, whereas many statistical analyses make assumptions that are violated when the distribution of the outcome is not at least approximately normal. These nonnormal outcomes occur because variables can be dichotomous (a student graduates or does not), categorical (a student could remain at the same institution, transfer to another institution, or drop out of college entirely), ordinal (a student might respond to the perceived importance of a life goal on a four-point scale from "not at all important" to

"essential"), or a count of the number of times an event occurs (a student could take no diversity courses, one course, two courses, and so on). These types of outcomes can be modeled successfully through the use of logistic, multinomial, ordinal logit, and Poisson regression, respectively (for more information, see Agresti, 2013; Long, 1997; Smithson & Merkle, 2013; Xie & Powers, 2008). These treatments of categorical and limited dependent variables can then be incorporated within some of the statistical techniques discussed in the appendix, such as multilevel modeling, structural equation modeling, and quasi-experimental analyses.

Practical and Statistical Significance

When considering research or assessment results, various stakeholders seek to answer a fundamental question: Are these effects meaningful? The vast majority of research studies emphasize one definition of meaningful, which is whether the results are statistically significant at some specific threshold of confidence (most commonly, $p<.05$). Statistical significance is arguably necessary to determine whether an effect is meaningful, since it suggests whether a particular finding is unlikely to have occurred by chance. However, it is also crucial to decide whether a finding is not only "real" (nonrandom), but also whether it is practically meaningful. Some national studies of higher education collect data on tens of thousands of students; because statistical tests are sensitive to the sample size, a result could be statistically significant while also being very small and therefore having little practical importance. For example, within a sample of 10,000 students, a seemingly trivial correlation of .02 would be statistically significant at $p<.05$. Many people would likely agree that this correlation is not meaningful in practice. Thus, given that higher education researchers, practitioners, and policymakers all want to create change that improves student outcomes, statistical and practical significance are both necessary for determining the importance of a finding.

This point brings up the difficult issue of how to determine whether an effect is substantively or practically meaningful. Any answer to this question depends on a variety of circumstances. First, obtaining a reasonable return on the investment of human and financial resources when making a change in institutional practices is a valid and important consideration. For example, a reasonably small effect might be considered worthwhile if it were obtained through minor and virtually cost-free adjustments in teaching or academic advising practices, but not if a proposed change was to open a comprehensive student success center that required numerous new employees and expensive facilities. Second, how much attention a relationship deserves is shaped in part by the rigor of the study that produced it. For instance, an effect size from a randomized experiment is probably more worthy of attention than one of similar size found in a study with fewer controls for potentially confounding variables, because more rigorous designs will often provide a more accurate estimate of a causal relationship. Third, in a somewhat related point, some outcomes have no true pretest (e.g., college satisfaction, perceptions of campus

climate). Because the pretest is often strongly related to the posttest, there is more variance to be explained by within- and between-college attributes for outcomes that have no true pretest. Fourth, some outcomes are more stable over time than others, so the same effect size is more impressive when it occurs for a less malleable outcome than for a variable known to fluctuate greatly. Fifth, the length of time between an experience and the outcome is also a relevant consideration. It certainly seems reasonable to expect a larger impact of service-learning (or any other experience) when civic engagement is measured in the next semester rather than several years after college graduation. Sixth, predictor variables can be dichotomous or continuous. It is therefore difficult, for example, to directly compare the size of effects for a dichotomous and a continuous independent variable predicting the same outcome. In short, the meaningfulness of the magnitude of an effect should be considered contextually to some extent.

Despite these complexities, college impact researchers need to have some basis for determining what constitutes a practically meaningful effect. We offer a discussion of this issue and provide specific recommendations. We necessarily discuss some statistical detail, because these guidelines are provided for specific statistical results. We hope that these recommendations will be widely used by researchers, who can then interpret the magnitude of these effects to better inform practice.

Effect Size Guidelines

Many social science studies use Cohen's (1988) general guidelines for effect sizes. Cohen frequently notes that researchers should not rely too heavily on these guidelines since they are general and provided for all behavioral sciences, which clearly includes a large and diverse array of disciplines and fields of study. These guidelines are also frequently misused not only through overreliance, but also through incorrect interpretation of the actual text. Most notably, Cohen suggests specific values for "small," "medium," and "large" effects, but he is often miscited as providing ranges of values for these criteria. He also provides guidelines for various types of effect sizes; we will focus on those that are most relevant to college impact research. A Cohen's *d* is the standardized difference between the means of two groups; this statistic is calculated as the difference of the means divided by the pooled standard deviation (i.e., the standard deviation for both groups combined). These guidelines state that a difference of .2 standard deviations is small, .5 is medium, and .8 is large. For correlation coefficients (which are ideal for indicating the simple relationship between two continuous variables), Cohen asserts that a small effect is a correlation of .1, medium is .3, and large is .5. He also provides guidelines for the variance explained (or *R*-squared) of a multiple regression analysis that includes multiple independent variables predicting an outcome variable; these are approximately .02 for a small effect, .13 for medium, and .26 for large.

Because these guidelines are not specific to higher education and there is little other basis for determining the magnitude of effects in college impact

research, we propose revised guidelines. We do so while recommending substantial caution about the use of these figures. We have already noted some reasons that one would consider the same effect size to be more practically meaningful in one context than another, including the cost of the intervention, the methodological rigor of the study, the stability of the outcomes examined, and the length of time between the experience and outcome. We provide guidelines for measures of effect size that we believe are particularly relevant to college impact research, including Cohen's *d*, multiple regression coefficients, and delta-*p* (which we define later). We also describe two conditions that *must* exist for the appropriate use of these effect sizes. First, these conditions are designed to describe the findings *only* from studies that used well-conducted experimental designs, quasi-experimental designs, and other multivariate analyses (e.g., regression, hierarchical linear modeling) that contain a set of rigorous control variables (i.e., which must include pretests when examining outcomes that could have a pretest). Stated differently, these guidelines are *not* appropriate for nonexperimental analyses that omit key predictors that are necessary to isolate the relationship between the experience of interest and the outcome.

Second, to compare the magnitude of effects for multivariate analyses within and across studies, all continuous variables (both dependent and independent) should be standardized with a mean of zero and a standard deviation of one, whereas dichotomous variables should not be recoded. This transformation ensures that unstandardized coefficients (i.e., for regression, multilevel modeling, structural equation modeling) for dichotomous independent variables predicting continuous outcomes are analogous to Cohen's *d*s (adjusting for all other variables in the model), whereas unstandardized coefficients for continuous independent variables predicting continuous outcomes are analogous to standardized regression coefficients or beta weights (Cohen, Cohen, West, & Aiken, 2003). This recoding of continuous predictors may be especially important for analyses predicting noncontinuous outcomes, because the coefficients from these analyses (e.g., odds ratios, delta-*p*s) depend on the coding values of the independent variable, and there are no coefficients from these analyses that are analogous to standardized regression coefficients. The unstandardized coefficients for all predictors should then be reported, and other coefficients should be provided when appropriate or helpful (e.g., delta-*p* for logistic regression). We feel that these coding and reporting choices are important regardless of whether the authors of a study choose to use these effect-size guidelines, since these will allow other researchers to easily compare the magnitude of effects across studies.

In a couple of circumstances, results of studies that do not use these coding practices can be subsequently converted to appropriate effect sizes. For studies that use unstandardized continuous variables, researchers can use the standard deviations to calculate what the results would have been if the variables were standardized before being included in the analyses. To accomplish this task, researchers must use the standard deviations provided in a descriptive table (or elsewhere) for this transformation to occur. In addition, standardized regression

coefficients when both the predictor and outcome are continuous will not be affected by our recommended coding, since these coefficients convey the relationships in terms of standard deviations by definition.

We also point out that our recommendations are not based on an exhaustive quantitative synthesis of the literature. Theoretically, one could record and analyze the tens of thousands of results that we summarize in this volume and create guidelines that are solely based on that empirical analysis. Such a synthesis would not only prove extraordinarily time-consuming, but it would also require that a sufficient number of studies have met the methodological standards and provided effect sizes that fit the specifications described here. That empirical approach would also need to determine how to define these guidelines within the sea of results. For instance, would a "large" effect size be defined as the cutoff for the top 5% of effect sizes for all eligible studies? The top 20%? How would one justify this decision? As an alternative approach, we rely on our own experience in conducting this research as well as existing literature. For instance, What Works Clearinghouse (2014) claims that "effect sizes of .25 standard deviations or larger are considered to be substantively important" (p. 23). Valentine and Cooper (2003) assert that effect sizes are generally smaller in education research than in other fields, which is consistent with Cohen's (1988) view:

> Thurstone once said that in psychology we measure [people] by their shadows. As the behavioral scientist moves from [her/his/their] theoretical constructs, to their operational realization in measurement and subject manipulation, very much "noise" (measurement unreliability, lack of fidelity to the construct) is likely to accompany the variables (p. 79).

That is, we can generally expect to find smaller effect sizes when examining real-world aspects of the college experience—in which curricula, cocurricular programs, and institutional missions are also implemented with varying degrees of effectiveness and measured with some degree of error—as predictors of real-world outcomes. Perhaps most important of all, adopting such a set of guidelines, while highly mathematical and "scientific," would also require overlooking or ignoring a vast number of relevant studies that fail to meet those criteria. In our view, as in Pascarella and Terenzini's (1991, 2005) two volumes, there is simply no substitute for judgment in developing the kinds of research syntheses that we provide in subsequent chapters. Thus, we provide estimates of effect sizes where appropriate, and we also rely heavily on the canons of good research and our professional training and experiences in making judgments about "the weight of evidence" in our summaries and conclusions.

With these caveats in mind, our guidelines for "small," "medium," and "large" effect sizes are presented in Table 1.2. Consistent with Valentine and Cooper's (2003) observation, our recommendations for Cohen's d or the standardized mean difference (small = .15, medium = .30, and large = .50) are smaller than

those that Cohen (1988) provided. However, we believe that a ½ standard deviation causal effect of a college experience on a meaningful college outcome certainly qualifies as "large." Similar to Cohen's guidelines, these figures are not minimum thresholds or specific ranges (which imply a false precision); for instance, we feel that effects of .28 and .34 standard deviations are both approximately "medium" (since they are both close to .30), while .40 standard deviations could be described as "between medium and large" in magnitude. These values for Cohen's *d* could come from a simple mean comparison (for experimental studies *only*) or from an unstandardized regression coefficient when the predictor is dichotomous and the outcome is continuous and standardized (for quasi-experimental or rigorous multivariate analyses *only*).

Cohen (1988) and others do not provide any guidance about the magnitude of standardized regression coefficients (although there is some tentative advice about translating between these multivariate statistics and raw correlations; see Peterson & Brown, 2005). Therefore, we are not able to draw on previous guidelines to supplement our thinking about the magnitude of the link between a continuous predictor and a continuous outcome. If we use a formula to convert a standardized mean difference to a point-biserial correlation (see Lipsey & Wilson, 2001), then a "large" Cohen's *d* of .5 corresponds to a correlation of .24 (if the sample sizes in the two groups are equal), .22 (if 70% of participants are in one group), or .20 (if 80% of participants are in one group). Moreover, as Lipsey and Wilson note, dichotomizing what is actually a continuous construct will result in a smaller effect size; this concern seems applicable in many cases in which students participate in a particular experience to a varying extent. Therefore, .2 seems reasonable as a "large" standardized regression coefficient. Using the same ratios as for Cohen's *d*, .06 and .12 seem reasonable as "small" and "medium" effect sizes for standardized regression coefficients, respectively.

For determining the impact of college on a dichotomous outcome, we prefer the use of the delta-*p* statistic to the odds ratio because this value can be interpreted more easily than the odds ratio (for more information, see Cruce, 2009; Petersen, 1985). Delta-*p* is the change in the probability of having a "1" on the dependent variable (rather than "0") that corresponds to a one-unit change in the independent variable. This probability change also depends on the values on the independent variables; therefore, to provide this estimate for the "average" participant, delta-*p* is often calculated for a participant who has the mean value on all predictors. It is more difficult to provide a delta-*p* value that is informed by the other effect size recommendations, since there is no way to determine the "variance explained" for a dichotomous outcome. Informed by Cohen's (1988) discussion of the *h* statistic and our own experiences, we propose that a "large" delta-*p* is .15. That is, an effect of a college experience is large if it corresponds to a 15 percentage point change in the probability of a dichotomous outcome occurring (e.g., college graduation). Using the same proportions as for the other effect size metrics, delta-*p*s of .05 and .09 would be

Table 1.2 Overview of Guidelines for Effect Size Metrics in College Impact Research When Key Conditions Are Met

Metric	*Explanation and Use*	*Small*	*Medium*	*Large*
Cohen's *d* (standardized mean difference)	Difference between two groups when predicting a continuous outcome variable (this metric should also be used for dichotomous predictors and a continuous outcome in multivariate analyses)	.15	.30	.50
Standardized regression coefficient	Relationship between a continuous predictor and a continuous outcome in a multivariate model (unstandardized coefficients and Cohen's *d* guidelines should be used if the predictor is dichotomous)	.06	.12	.20
Delta-*p*	Change in probability when predicting a dichotomous outcome in a multivariate model (for both dichotomous and continuous predictors)	.05	.09	.15

Note. These guidelines should be used only when a study meets the following conditions. First, the study employs an experimental design, quasi-experimental design, or rigorous multivariate analyses (with appropriate control variables, including a pretest). Second, for Cohen's *d* and delta-*p*, continuous dependent and independent variables are standardized with a mean of zero and a standard deviation of one, whereas dichotomous variables are not transformed. The values for delta-*p* should also be used for average marginal effects. These effect size guidelines should be considered in the context of relevant study features, such as the overall rigor of the study, the financial return on investment, the malleability of the outcome, and the length of time between the experience and outcome.

then considered "small" and "medium," respectively. These same values for delta-*p* should also be used for average marginal effects, which similarly provide the effect of a one-unit change in the independent variable on the dependent variable.

Summary

This section provided an overview of some of the challenges regarding the valid measurement of student outcomes and changes in these outcomes and determining whether effects have practical significance for higher education professionals and policymakers. We also offered effect size guidelines for college impact studies so that people who produce and use this research can have a common understanding of what constitutes a small, medium, and large effect of college. We hope that this discussion provides the needed context for helping readers understand the weight of the evidence considered in this volume.

CHAPTER CONCLUSION

This chapter provided an overview of the organization of this volume, the changing higher education landscape as a pretext for understanding some of the choices authors made in their lines of inquiry, some of the challenges with measuring and modeling student outcomes, and some methodological innovations with regard to making meaning of the numbers used as evidence for the claims in the book. Indeed, writing a book of this scope requires attention to detail without losing sight of some of the larger questions facing higher education stakeholders.

To accomplish the former, we offer Chapters 2 through 9, which are organized by student outcome areas: the development of verbal, quantitative, and subject matter competence (Chapter 2); cognitive and intellectual development (Chapter 3); psychosocial change (Chapter 4); attitudes and values (Chapter 5); moral development (Chapter 6), educational attainment and persistence (Chapter 7); career and economic impacts of college (Chapter 8); and quality of life (Chapter 9). The methodological appendix at the end of the book further illustrates some of the details important for making meaning of chapter content.

In order to address some of the larger issues facing higher education stakeholders, we have provided two summary chapters. Chapter 10 summarizes points consistently raised across all chapters. In Chapter 11, we discuss our work's implications for policymakers, researchers, and practitioners. It is our hope that a variety of stakeholders interested in higher education will use this volume to create and optimize contexts for student success.