The Wabash National Study - The Impact of Teaching Practices and Institutional Conditions on Student Growth

Charles Blaich and Kathleen Wise
Center of Inquiry
Wabash College
Crawfordsville, IN

Submitted for 2011 American Educational Research Association Annual Meeting

Three documents are included with this submission:

1. Blaich, C. F., & Wise, K. S. (2011, January). From gathering to using assessment results: Lessons from the Wabash National Study. (NILOA Occasional Paper No. 8). Urbana, IL: University of Illinois and Indiana University, National Institute of Learning Outcomes Assessment
2. Additional four-year data from the Wabash National Study (Figures 1-5)
3. Appendix A: Wabash Study Research Methodology

From Gathering to Using Assessment Results:

Lessons for the Accountability Movement from the Wabash National Study

Charles Blaich and Kathleen Wise

Center of Inquiry

Wabash College

accountable (ac·count·a·ble) – *adjective* (of a person, organization, or institution) required or expected to justify actions or decisions; responsible.[1]

The Accountability Movement: Common Assumptions and Practices

The fact that the word "accountability" runs through much of the national conversation about assessment in higher education says a lot about the presumed motives and responsibilities of the parties in this discussion. The implication, of course, is that unless they are held accountable by an outside authority, some colleges and universities at least would not make good on their commitment to students.

It is both reasonable and necessary for public entities, such as the federal government or regional accreditors, to hold colleges and universities accountable for educating students consistent with basic standards and institutional missions (Kuh & Ikenberry, 2009). At the same time we cannot ignore the fact that quality assurance efforts take place in the midst of a broader public discourse in which politicians and pundits wag their fingers at Wall Street bankers, their political opponents, government agencies, teachers, and other alleged evildoers—demanding they be held accountable. The not-too-subtle subtext here is that those who make these calls for accountability are acting for the public good while those expected to respond to them most definitely are not. As anyone who reads the higher education trade papers knows, the accountability movement is as much about politics as it is about student learning.

Due in part to these political realities, accountability efforts in higher education—even when applied carefully and with good intentions—shift how institutions do their work. In a recent survey by the National Institute for Learning Outcomes Assessment (NILOA), provosts and chief academic officers from over 1,500 institutions across the U.S. reported that assessment

[1] http://oxforddictionaries.com/view/entry/m_en_us1219575#m_en_us1219575

data at their institutions was most commonly used to prepare for accreditation (Kuh & Ikenberry, 2009). Institutions have to invest enormous resources to meet accountability requirements. Yet, as Peter Ewell (2009) has pointed out, institutions' engagement in assessment for the purposes of accountability—focusing on gathering evidence to prove that student learning has occurred—is different from their engagement in assessment to improve student learning. The counterargument to this critique, of course, is that without accountability efforts a significant portion of colleges and universities would not serve their students as they should—bringing us back to the politically charged suspicions of the motives of the parties in higher education.

An Alternative Approach to Accountability

One can, however, think of accountability from a different, higher standpoint: not in terms of the standards an outside authority holds us to but, rather, in terms of the responsibility we, as staff, faculty, and administrators, assume as teachers and as professionals. As Lee Shulman (2003, para. 4) has stated,

> My point is that excellent teaching, like excellent medical care, is not simply a matter of knowing the latest techniques and technologies. Excellence also entails an ethical and moral commitment—what I might call the "pedagogical imperative." Teachers with this kind of integrity…inquire into the consequences of their work with students. This is an obligation that devolves on individual faculty members, on programs, on institutions, and even on disciplinary communities. A professional actively takes responsibility; she does not wait to be held accountable.

The Association of American Colleges and Universities (AAC&U) and the Council for Higher Education Accreditation (CHEA) (2008) have also called for the higher education community to take on this pedagogical imperative:

> Finally, and perhaps most important, higher education has an obligation to our democracy as well as our economy. A college degree should ensure that graduates are well prepared to contribute to society as knowledgeable, engaged, and active citizens.
>
> In order to meet these challenges, we in the higher education community must continually seek, and find, better ways to reach our common goal of helping all the students we serve realize their full potential. We need to make clear, for ourselves and our various constituencies, what our aims are, how we seek to achieve them, and how well we do so. This requires continuing efforts in many quarters to make higher education a challenging and rigorous experience for all students—for their benefit and society's as well. To do so, we in higher education must constantly monitor the quality of student learning and development, and use the results both to improve achievement and to demonstrate the value of our work to the public. We must not settle for anything less. (p. 1)

Yet accountability for improving student learning as an enactment of the moral and professional commitments of faculty, staff, and institutions, rather than as a reaction to externally imposed obligations, is an idea that rarely surfaces in the public discussion about assessment and accountability.

The Wabash College Center of Inquiry and the Wabash National Study

The Center of Inquiry at Wabash College collaborates with institutions across the country to collect and use evidence to improve student learning. At times, our experience working with colleges and universities has shown the necessary role of external authorities in holding institutions accountable for promoting student learning. But hundreds of staff, faculty, students, and administrators we have worked with across the country—rather than joining the strident fray over assessment and accountability—have taken Shulman's admonition to heart. In working with these committed professionals, we have discovered common assumptions about and practices in assessment that inhibit their efforts and that have implications for the accountability movement. This paper describes the important lessons we have learned in the Wabash National Study about structuring and implementing assessment programs to advance the work of these "improvement agents."

The Wabash National Study, the primary means by which the Center of Inquiry collaborates with institutions for assessment, is a longitudinal research and assessment project designed to deepen our understanding of the teaching practices, student experiences, and institutional conditions that promote the development of students' critical thinking, moral reasoning, leadership towards social justice, well-being, interest in and engagement with diversity, and interest in deep intellectual work (Table 1). Since its pilot version in 2005, over 17,000 students from 49 colleges and universities have joined the Wabash Study. The study is still in progress, and 30 institutions joined a new version of the study in fall 2010. The first institutions to join the Wabash Study did so in response to the national conversation about accountability. In some cases this response was driven directly by impending accreditation review and in others by initiatives from foundations or higher education organizations.

The Wabash Study was collaboratively designed by researchers from the University of Iowa, the University of Michigan, Miami University, ACT Inc., and the Center of Inquiry. Our goal was to create a "gold-standard" longitudinal study that included measures of what students brought to college, what they experienced during college, and a wide range of learning outcomes.

In designing the Wabash Study, we made three core assumptions about what helps and what hinders effective assessment. First, we believed that a lack of high-quality data was the primary obstacle that institutions faced in using assessment evidence to promote improvements in student learning. Second, we thought that providing detailed reports describing study findings would be the key mechanism for kicking off a sequence of events on campus that would culminate in evidence-based improvements. Finally, we assumed that the intellectual approach that faculty and staff took in their scholarship would facilitate their work on assessment projects to produce improvements in student learning.

These assumptions account for the extensive range of measures we adopted, the long and detailed reports about study findings we developed for institutions, and the mechanisms we added to help institutions merge Wabash National Study data with institutional data. They also led us to focus our data analyses, so that we provided institutions with information on the practices and conditions that promoted growth on the outcomes that we measured. In creating the study, we thought that once faculty and staff had "good" data from a high-quality research project, they would use it to improve student learning. In essence, we designed the Wabash Study to solve a "lack of quality data" problem. Insofar as we, as well as our institutional partners, assumed that study data would inevitably lead to improvements in student learning, we spent most of our time and energy on building mechanisms to gather data. We did not ask our

institutional partners to consider what they would do with our detailed reports once they landed on their collective desks.

Findings from the Wabash National Study

To date, four principal findings have emerged from the Wabash Study. A discussion of each follows.

First, as Arum and Roska (in press) have reported, students do not always grow as much as we hope or in the directions that we expect in college. As shown in Figures 1–3, during four years of the Wabash Study, students grew on some of the outcomes we measured, such as critical thinking and moral reasoning, and declined on others, including academic motivation and openness to diversity.

Second, students still benefit from the good practices and conditions that Chickering and Gamson (1987) highlighted over 20 years ago (Pascarella, Cruce, Wolniak, & Blaich, 2004; Pascarella, Seifert, & Blaich, 2010; Pascarella, Wolniak, Seifert, Cruce, & Blaich, 2005; Seifert, Goodman, Lindsay, Jorgensen, Wolniak, Pascarella, et al., 2008; Seifert, Pascarella, Goodman, Salisbury, & Blaich, 2010). These good practices and conditions, which are easily and well measured by surveys such as the National Survey of Student Engagement and the Higher Education Research Institute's Your First College Year and College Senior Survey, have an impact on almost every outcome we measured. Our research highlighted four dimensions of these good practices and conditions (see Table 2 for examples):

1. Good Teaching and High-Quality Interactions with Faculty
2. Academic Challenge and High Expectations
3. Diversity Experiences

4. Higher-Order, Integrative, and Reflective Learning

Third, as Kuh (2003) described, the variability within our institutions—both in terms of growth on the outcomes and the level of good practices and conditions experienced by students—dwarfs the differences between institutions on these variables (see also the NSSE 2008 Annual Report at http://nsse.iub.edu/NSSE_2008_Results/). Although many discussions about assessment focus on the importance of creating measures by which to compare institutions, the underlying reality is that any overall institutional measure belies the complex range of student learning and experiences that occurs within our institutions. As depicted in Figures 4 and 5, the variation among students within Wabash National Study institutions is vastly larger than the median differences between institutions. Even if a school has greater average growth on critical thinking or some other outcome than its peer institutions, it is likely that many of its students will not have grown or may even have declined on these outcomes. The comparatively high average growth of the institution does not matter for these students. Learning what differentiates the students who learn substantially more or substantially less than their institution's average score on an outcome is the grist for good assessment work. Faculty, staff, and administrators at almost every Wabash Study institution have been surprised and concerned that their students seem to experience small growth or even declines on outcomes in the study. Nonetheless, we were optimistic about the possible benefits of the study because we found evidence at every participating institution about the good practices and conditions that played a role in how their students were, or were not, changing in college. In other words, every Wabash Study institution could address their concerns about students' growth by responding to specific evidence about the strengths and weaknesses of their teaching and learning environments.

This brings us to the fourth primary finding from the study: It is incredibly difficult to translate assessment evidence into improvements in student learning. Unfortunately, we learned early on that *gathering data*, even with the complicated longitudinal methodology employed in the Wabash Study, is much easier than *using* the information to improve student learning. As we monitored how institutions were using the information from the Wabash Study through numerous follow-up phone calls, meetings, and site visits—and even by tracking how often institutions downloaded our reports—we learned that evidence from the study was having little impact. Although all 19 institutions from the first cohort of the Wabash National Study in 2006 worked extraordinarily hard to collect data multiple times from students, nearly 40% of the institutions have yet to communicate the findings of the study to their campus communities, and only about a quarter of the institutions have engaged in any active response to the data.

Why So Little Action?

As we worked with institutions in the first version of the study, our assumptions concerning the importance of gathering additional high-quality data; of creating long, detailed reports; and of engaging the scholarly energies of faculty and staff proved to be completely wrong. We had focused too much on gathering, analyzing, and reporting assessment evidence and not enough on helping institutions use it.

In our work with the Wabash Study, we learned that most institutions already had more than enough actionable assessment evidence—not only in terms of national surveys and standardized outcome measures but also from information in institutional databases, student interviews, reports from external reviewers, and many other sources of information about student learning. A typical and somewhat disappointing experience in working with institutions was that

many of the actionable findings we thought we had discovered in the Wabash Study—for example, evidence about changes in student attitudes about diversity, the quality of student-faculty interactions, the level of binge drinking, or the amount of time students spent working off campus—were either already well known by a couple of people on campus or were tucked away unnoticed among assessment data collected previously.

Why so much collection—but so little utilization—of data? Most institutions have routinized data collection, but they have little experience in reviewing and making sense of data. It is far easier to sign up for a survey offered by an outside entity or to have an associate dean interview exiting students than to orchestrate a series of complex conversations with different groups on campus about what the findings from these data mean and what actions might follow. The norm for many institutions is to gather data, to circulate the resulting reports among a small group of people, and then to just shelve them if nothing horrible jumps out—and sometimes even if it does! In recent years of the Wabash Study, we posted reports of institutional data on a website, allowing us to get an idea of how many people at each campus actually opened and either read or downloaded the reports. The extent to which the reports were opened varied dramatically across campuses. At several institutions only a handful of people opened the report, while faculty and staff at one small institution viewed the report over 150 times.

Even when assessment reports are disseminated widely, most of us behave as though the data in the reports will speak loudly enough to prompt action. We tend to believe that interesting findings will naturally prompt discussions and ultimately revisions in our courses and programs. But this denies the reality on most of our campuses—that the current state of affairs in our departments, curricular structures, and programs is usually a compromise carefully negotiated

among numerous parties over the course of years. Unless the findings are truly devastating, assessment data has little impact on this tightly constrained arrangement.

General reports about outcome changes or student experiences that are not embedded into an ongoing campus conversation about student learning are just quickly filed away and forgotten, sometimes without even being read. Implicitly, we are relying on people's curiosity as the mechanism to generate discussion and, ultimately, action about data. For the most part, faculty, staff, and students are curious about their institutions, but in the busy, multitasking environments in which we all work, general curiosity does not compete well against the classes we need to prepare, the papers we need to write or grade, and the programs we need to implement. The way we govern and structure our institutions means that the simple reporting of assessment data has little hope of generating the kind of "data-informed, continuous improvement" that many of us hope for. Assessment data has legs only if the evidence collected rises out of extended conversations across constituencies about (a) what people hunger to know about their teaching and learning environments and (b) how the assessment evidence speaks to those questions.

The kinds of community processes necessary for identifying assessment questions and for making sense of assessment evidence, furthermore, are different from the individual or small-team interactions that typically lead to the identification of a research question and the analysis of texts or quantitative data in our scholarly work. Scholarship is mostly a solitary endeavor, and its few communal interactions are generally among people who share similar intellectual training and backgrounds. Done correctly, using assessment to improve student learning is an entirely public process in which people with different levels of experience and different intellectual backgrounds must work together toward a common end.

As Upcraft and Schuh (2002) have pointed out, research and assessment also have different goals. Although research practice varies by discipline, most researchers focus on discovering or identifying something and communicating this finding with others within their discipline. The actions entailed by good research are communication with colleagues via presentations and publications and, ultimately, more research. The goal of assessment, on the other hand, is to create changes that improve student learning. Assessment also entails communication with colleagues, but the communication must at some point move from talking about the data to talking about, and then enacting, changes. Research and assessment are not just different processes; at some point the goals of each process are in opposition. For scholars, it is hard to imagine reaching too deep a level of understanding about one's subject. One way of inviting faculty to engage in assessment is to frame it as a form of inquiry. The challenge then is to engage faculty's interest in inquiry without engaging the other familiar scholarly skills that will lead them to gather more data and write reports rather than taking concrete actions. Not only does constantly gathering and analyzing additional data fit neatly into faculty's intellectual wheelhouse, it also allows faculty and administrators to avoid expending their political capital by advocating for change. It's far less risky and complicated to analyze data than it is to act.

For assessment to be successful, it is necessary to put aside the question, "What's the best possible knowledge?" and instead to ask, "Do we have good enough knowledge to try something different that might benefit our students?" Ultimately, the most fruitful way to learn if the conclusions that we have drawn from assessment data are correct is to try to change something and see what happens.

Designing Assessment for Improvement

Fortunately, we observed the trends described above early enough in the study that we could adjust course. We have continually revised the way we work with institutions to use evidence from the Wabash Study and other forms of evidence to promote improvements in student learning. The revisions are based on our new understanding that "closing the loop" and using evidence to promote improvements is as much a political and sociocultural process as it is an evidence-driven process (Blaich & Wise, in press). We now encourage institutions that enter the Wabash Study to pay as much attention to creating and sustaining processes by which faculty, staff, and students reflect on and consider responses to the evidence as they do to developing processes for collecting assessment evidence. Specifically, we work with institutions to develop and implement detailed, three-year plans that include the following components:

Data audits. To ensure that institutions entering the Wabash Study are aware of and use the evidence they already have, we ask that they complete a data audit survey—listing all of the assessment data they already possess along with data they plan to collect on what students bring to college, what they experience in college, and what they learn in college (see http://www.liberalarts.wabash.edu/storage/Institutional-Assessment-Portfolio-Data-Survey.pdf). This prompts institutions to consider not only standardized surveys and tests but also ways they could use student work and data from their student information systems.

A clear focus. We strongly encourage institutions to focus their plans on no more than two or, at most, three outcomes, but our preference is that they focus on only one. In our experience, institutions that try to engage in too many initiatives wind up accomplishing none of them. With assessment projects prioritized on one, two, or maybe three specific

outcomes, institutions can then sift through their typically vast piles of assessment evidence to focus on specific elements that relate to their chosen outcomes.

Communication. Creating sustained conversations about assessment data and engaging in sense-making activities is akin to a campaign—not a series of reports posted on a website. We ask institutions to list the individuals, constituencies, and governance structures that need to be engaged in their discussions of assessment evidence and then to develop plans for how they will engage these constituencies in conversations for both making sense of as well as developing responses to the data. Even before these groups get data, it is important to consider engaging them in planning for what different findings might imply. Creating effective plans for structuring conversations and activities under various "good news" and "bad news" scenarios may happen more productively through a series of "what if" exercises than with the actual evidence, and all that it implies, sitting in front of you.

Resources. We ask institutions to set aside $10,000 to devote to activities designed to encourage institutional conversations about and responses to the data. We also ask them to develop plans for obtaining any additional resources and support they will require from institutional leaders to promote these conversations and to consider when visits from outside consultants might help their work on campus.

Student involvement. Finally, we encourage institutions to engage students to help them make sense of assessment data. One of the most important lessons we learned from the first version of the Wabash Study is how much you can learn by sitting informally with groups of students and asking them to reflect on some of your institutional data or to respond to simple questions like, "What things have faculty and staff done that have

made a difference in what you have learned this year?", "What classes are hardest for you and what makes them hard?", or "What surprised you when you first came to college?" We have found that these conversations are a critical way of linking assessment data with specific qualities of students' experience to get a richer sense of the data. (For more information about the plans institutions develop in the new version of the Wabash Study, see the plan template at http://www.liberalarts.wabash.edu/storage/Wabash-Study-Plan-template-final.docx and a description of the institutional assessment portfolio at http://www.liberalarts.wabash.edu/storage/Institutional-Assessment-Portfolio-description.pdf.)

Of course, even the most carefully thought-out plan will shift and require revision as soon as it is implemented. The point of planning is not to create a rigid procedure that institutions will follow regardless of what happens on campus but to help campuses frame assessment from the start as a many-step process that culminates in improvements. To be successful, institutions must stop thinking about assessment as a process that begins with data-gathering and ends with a report.

Enhancing Institutions' Engagement in Assessment and Improvement Efforts

To help institutions keep on track as they revise their plans in response to unforeseen events as well as to ensure that the Center's support actually benefits institutions, we continually assess both the progress of institutions and our work with them. Like any other complex social project, it is important to "iterate," enhancing an institution's engagement with assessment evidence on the fly. Banta and Blaich (in press) have described a number of questions to

consider when evaluating the progress of assessment projects and determining whether the projects need to be revised:

Resources. Are institutions devoting as many resources—in terms of time, money, personnel, and effort—to creating mechanisms for making sense of and developing responses to assessment evidence (workshops, meetings, faculty and staff development opportunities, small grant programs, conversations with students, etc.) as they are to gathering the evidence? If not, the process needs to be revised. If all of an assessment program's resources are devoted to gathering evidence and none toward making sense of and using evidence, no change is likely to occur.

Communication of assessment results. If asked, can faculty, staff, and students readily identify the outcomes, measures, and recent findings of their institution's assessment program? If asked, would all faculty members in a department be able to cite the same two or three things that their department is doing well and the same two or three areas for improvement along with evidence that supports their assertions? If the answer to these questions is "no," then it is time for the institution, department, or program to revise how it communicates about assessment.

Getting evidence to potential users. Assessment evidence will have no impact if it is not widely shared and discussed on campus. Hiding data because they are too controversial, sending out a report via email, or posting information on a website without creating opportunities for people to come together to reflect on and make sense of the findings will ensure that assessment evidence has little long-term impact. Institutions also need to create structures and resources to take advantage of faculty and staff interest that emerges in response to assessment evidence. For example, do faculty, staff, and students know

where to go to find assessment evidence to address questions about their programs, departments, or majors? Is there someone they can contact if they have questions about the information? If they do contact that individual, will they get a timely response? The key is getting evidence into the hands of people who are able and interested in using it to improve student learning and student experience and then supporting their efforts to understand and use the data.

Conclusion

Patience is an important virtue for those engaged in assessing student learning and using the findings to improve student and institutional performance. In our current work, we plan that institutions will take at least three to four years to make sense of and act on assessment evidence on one or two learning outcomes. Yet we suspect this plan is overly ambitious.

Where does the time go? The vast majority of our work with institutions focuses on the politics and procedures of using evidence, not on collecting it. For all of the challenges we face trying to gauge student growth on our institutional outcomes, it is far easier to collect data measuring student learning and experiences than it is to use these data. One reason for this difference is that there are many nationally known standardized tests, surveys, predesigned rubrics, or e-portfolio systems that institutions can adopt to collect assessment data, and in some cases, to deliver detailed reports. We have sometimes heard these assessment options referred to as “assessment in a box” or “plug in and play assessment.” This way of gathering assessment evidence is still not easy, but it cuts down on the things that institutions have to design from scratch.

Unfortunately, there is no "plug in and play" system for using assessment data to change our institutions. The messy processes that inevitably follow once the data has been collected cannot be outsourced in the same way that we can outsource components of evidence-gathering for assessment.

There are many wonderful books and papers about how to use assessment evidence, including works by Banta (1999), Woolvard (2004), Patton (2008), and Suskie (2009). But the leap from reading good advice about working formatively with assessment data to applying that advice in the academic polis is far greater than that between reading about and implementing concrete suggestions for gathering a better research sample of students. We believe the next step in developing the necessary scholarship and expertise for assessment is to create mechanisms to systematically train campus assessment leaders in the political skills and organizational knowledge they need to more fully utilize their assessment data. To effectively promote improvements in student learning, it is just as important for assessment leaders to be able to draw on the work of, for example, Kezar (2001) and Kezar and Lester (2009) on facilitating institutional change as it is for them to know the reliability of assessment measures or how to create an e-portfolio.

We began this paper with references to the national discourse on accountability in higher education. Our collaborative work with colleges and universities over the last five years on the Wabash National Study has led us to wonder whether the advocates for accountability and improvement have a realistic sense—both in terms of student learning as well as in terms of institutional change—of what kind of change is possible over a four- or five-year period. The research on institutional change suggests that "institutional transformation" is rare and that, if anything, incremental change is what is best and what is possible (Kezar, 2001). The question

we are left with, then, is whether any institution—even the colleges and universities most committed to being accountable and to improving student learning—can meet the standards set by the discourse.

References

Arum, R., & Roska, J. (in press). *Academically adrift: Limited learning on college campuses.* Chicago, IL: University of Chicago Press.

Association of American Colleges and Universities (AAC&U) and the Council for Higher Education Accreditation (CHEA). (2008). *New leadership for student learning and accountability: A statement of principles, commitments to action.* Washington, DC: Authors.

Banta, T. W. (1999). *Assessment essentials: Planning, implementing, and improving assessment in higher education.* San Francisco, CA: John Wiley & Sons.

Banta, T. W., & Blaich, C. (in press). Why is closing the loop so hard? And what can be done to change that? *Change*.

Blaich, C., & Wise, K. (in press). Moving from assessment to institutional improvement. In T. Seifert (Ed.), Using longitudinal assessment for institutional improvement (Assessment Supplement 2010). *New Directions for Institutional Research.*

Chickering, A. W., & Gamson, Z. (1987). Seven principles for good practice in undergraduate education. *American Association for Higher Education Bulletin, 39,* 3–7.

Ewell, P. (2009, November). *Assessment, accountability, and improvement: Revisiting the tension* (NILOA Occasional Paper No.1). Urbana, IL: University of Illinois and Indiana University, National Institute of Learning Outcomes Assessment.

Kezar, A. J. (2001). *Understanding and facilitating organizational change in the 21st century: Recent research and conceptualizations* (ASHE-ERIC Higher Education Report, Vol. 28, No. 4). San Francisco, CA: Jossey-Bass.

Kezar, A. J., & Lester, J. (2009). *Organizing higher education for collaboration: A guide for campus leaders.* San Francisco, CA: John Wiley & Sons.

Kuh, G. (2003). What we're learning about student engagement from NSSE. *Change, 35*(2), 24–32.

Kuh, G., & Ikenberry, S. (2009, October). *More than you think, less than we need: Learning outcomes assessment in American higher education.* Urbana, IL: University of Illinois and Indiana University, National Institute of Learning Outcomes Assessment. Retrieved from http://www.learningoutcomesassessment.org/NILOAsurveyresults09.htm

Pascarella, E. T., Cruce, T. M., Wolniak, G. C., & Blaich, C. F. (2004). Do liberal arts colleges really foster good practices in undergraduate education? *Journal of College Student Development, 45,* 57–74.

Pascarella, E. T, Seifert, T. A., & Blaich, C. (2010). How effective are the NSSE benchmarks in predicting important educational outcomes. *Change, 42,* 16–22.

Pascarella, E. T., Wolniak, G. C., Seifert, T. A., Cruce, T. M., & Blaich, C. F. (2005). *Liberal arts colleges and liberal arts education: New evidence on impacts* (ASHE Higher Education Report, Vol. 31, No. 3). San Francisco: Jossey-Bass.

Patton, M. Q. (2008). *Utilization-focused evaluation* (4th ed). Los Angeles, CA: Sage.

Seifert, T. A., Goodman, K. M., Lindsay, N. K., Jorgensen, J. D., Wolniak, G. C., Pascarella, E. T., et al. (2008). The effects of liberal arts experiences on liberal arts outcomes. *Research in Higher Education, 49,* 107–125.

Seifert, T. A., Pascarella, E. T., Goodman, K. M., Salisbury, M. H., & Blaich, C. F. (2010, January/February). Liberal arts colleges and good practices in undergraduate education: Additional evidence. *Journal of College Student Development, 51*, 1–22.

Shulman, L. S. (2003). No drive-by teachers. *Carnegie Perspectives*. Retrieved from http://www.carnegiefoundation.org/perspectives/no-drive-teachers

Suskie, L. (2009). *Assessing student learning: A common sense guide.* San Francisco, CA: Jossey-Bass.

Upcraft, M. L., & Schuh, J. H. (2002). Assessment versus research: Why we should care about the difference. *About Campus, 7(1)*, 16–20

Woolvard, B. E. (2004). *Assessment clear and simple: A practical guide for institutions, departments, and general education.* San Francisco, CA: Jossey-Bass.

Table 1. Wabash National Study Outcome Measures

Academic Motivation Scale
ACT Collegiate Assessment of Academic Proficiency Critical Thinking Test
Contribution to the Arts and Humanities Scale
Contribution to the Sciences Scale
Defining Issues Test of Moral Reasoning (Version 2)
Miville-Guzman Universality-Diversity Scale (Short Form)
Need for Cognition Scale
Openness to Diversity and Challenge Scale
Political and Social Involvement Scale
Positive Attitude toward Literacy Scale
Ryff Scales of Psychological Well-Being
Socially Responsible Leadership Scale (Revision 2)

Note. More information about these scales can be found at http://www.liberalarts.wabash.edu/study-instruments/

Table 2. Examples of Good Practices and Conditions from the Wabash National Study

Having faculty and staff who
Have a genuine interest in teaching and are interested in helping students grow in more than just academic areas
Provide timely feedback
Check to see if students learned the material before moving on to new material
Design clear explanations of their course or program goals and requirements
Develop organized classes and presentations
Provide clear explanations of course goals and requirements
Engage in high-quality nonclassroom interactions that influence students' growth, values, career aspirations, and interest in ideas
Ensure that students work hard to prepare for their classes and are required to read and write a substantial amount of material
Challenge students to analyze and synthesize information and make judgments about ideas, experiences, and theories
Ask students to integrate ideas and information from different sources and to include diverse perspectives in their work
Ask students to examine the strengths and weakness of their ideas and to understand someone else's view by imagining how an issue looks from his or her perspective

Note. These examples are based on questions from the Wabash National Study Student Experiences Survey and the National Survey of Student Engagement. For a complete list of the effective practices and conditions see http://www.liberalarts.wabash.edu/study-research/

Figure 1. Four-Year Change in Moral Reasoning, Critical Thinking, Socially Responsible Leadership, and Need for Cognition Among Four-Year Institutions in the 2006 Cohort (in Standard Deviations)

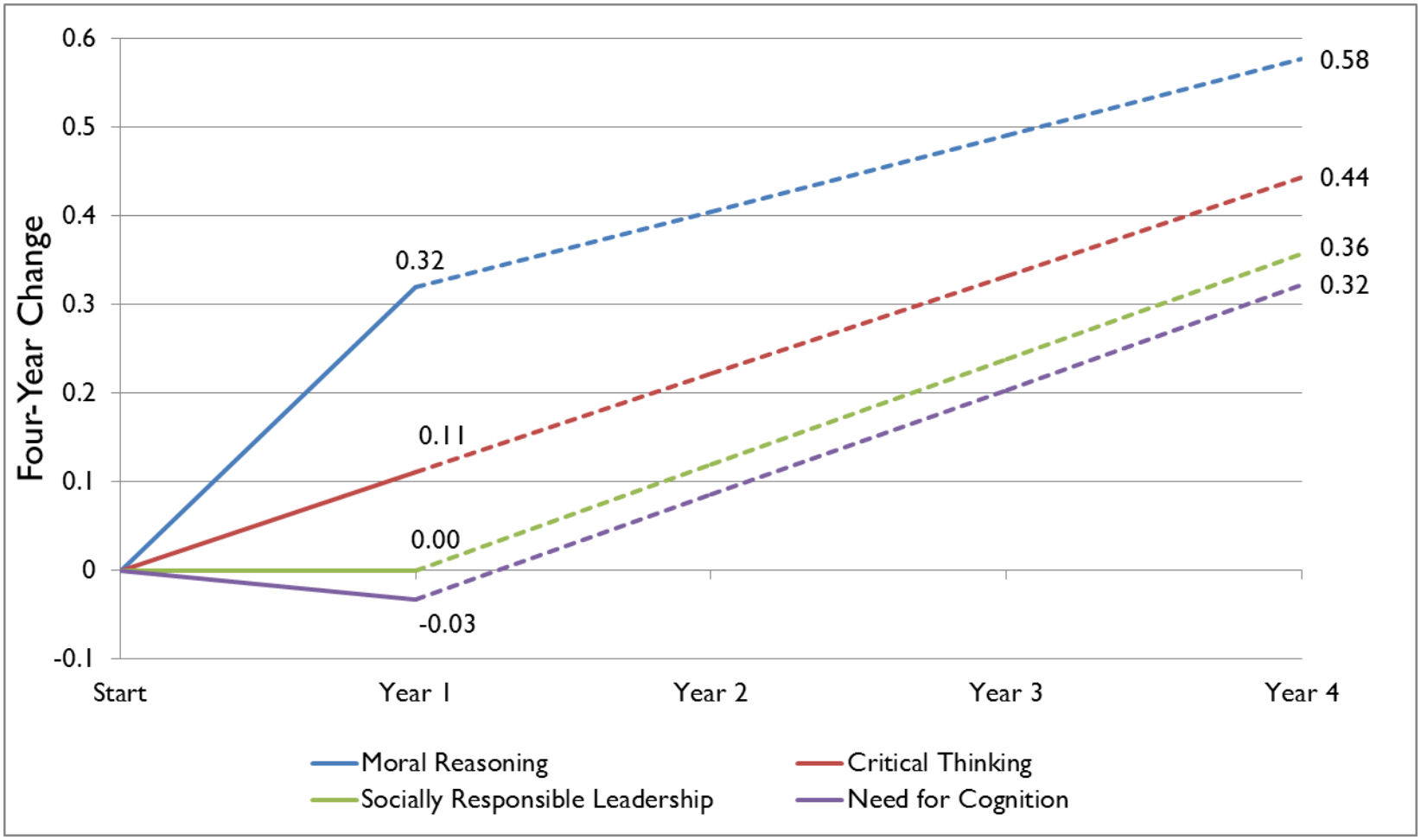


Figure 2. Four-Year Change in Psychological Well-Being, Universality-Diversity Awareness, Political and Social Involvement, and Openness to Diversity and Challenge Among Four-Year Institutions in the 2006 Cohort (in Standard Deviations)

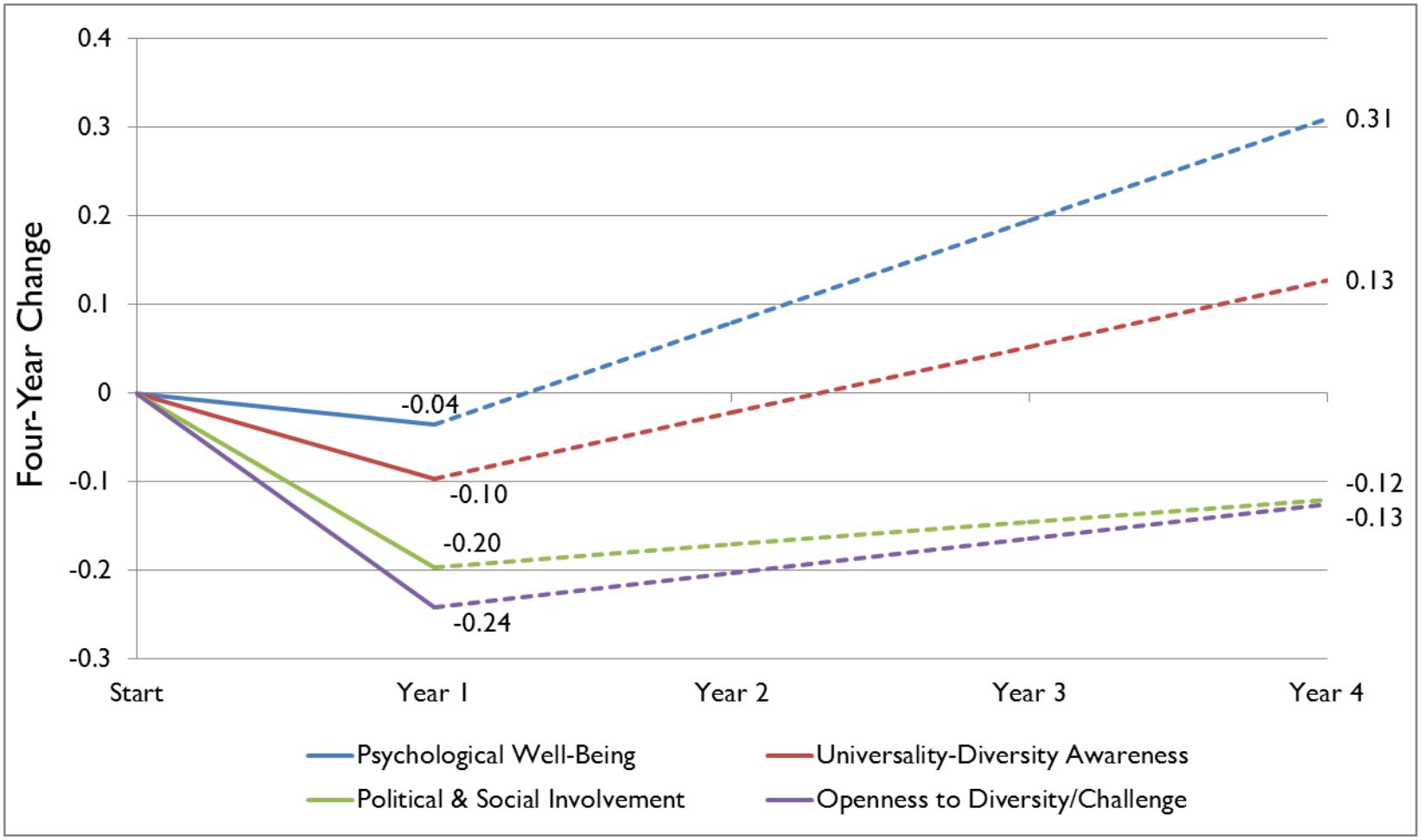


Figure 3. Four-Year Change in Positive Attitude Toward Literacy, Contribution to the Arts, Contribution to the Sciences, and Academic Motivation Among Four-Year Institutions in the 2006 Cohort (in Standard Deviations)

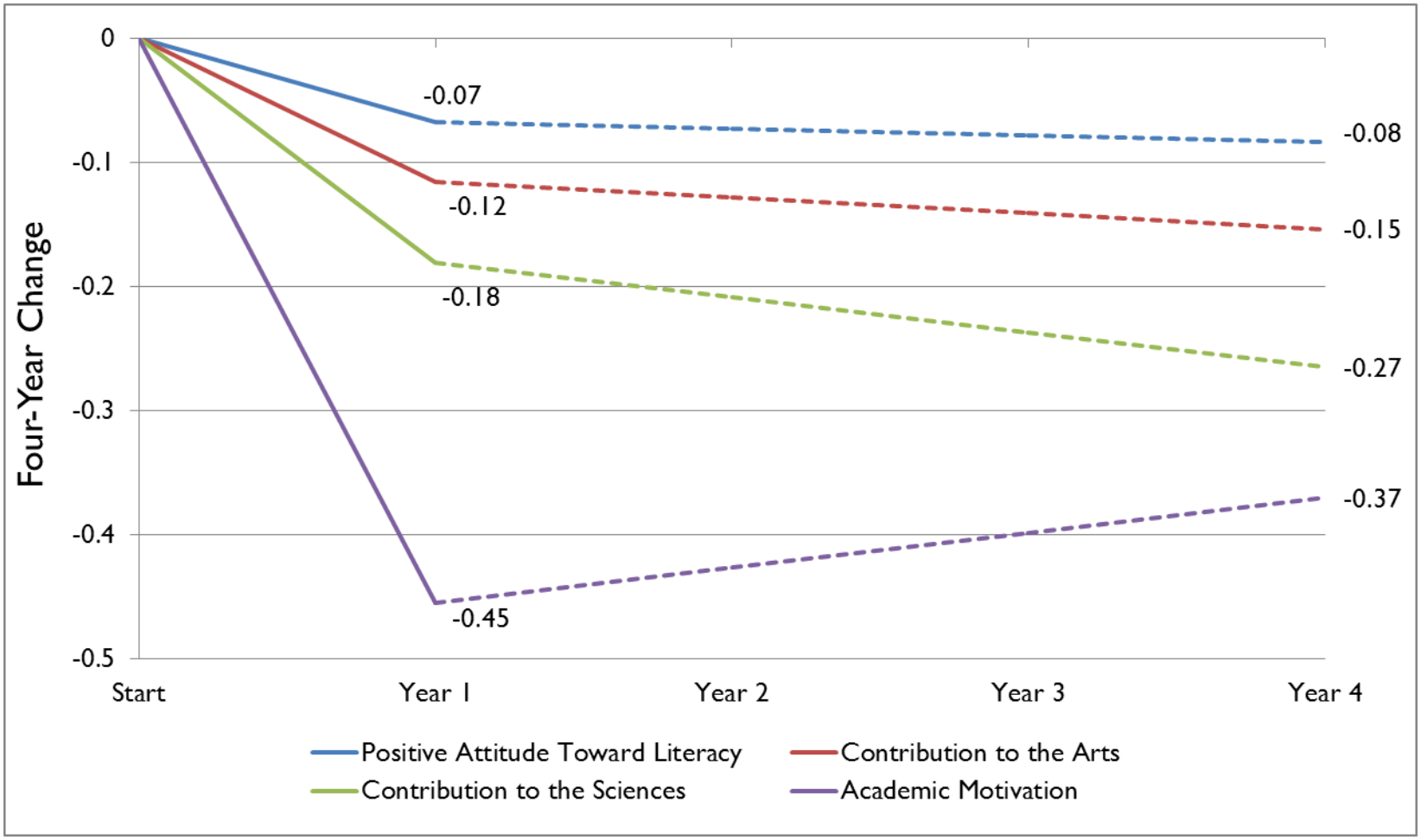


Figure 4. Within-Institution Variation at Small Colleges (S) and Large Universities (L) in Four-Year Change in Academic Motivation

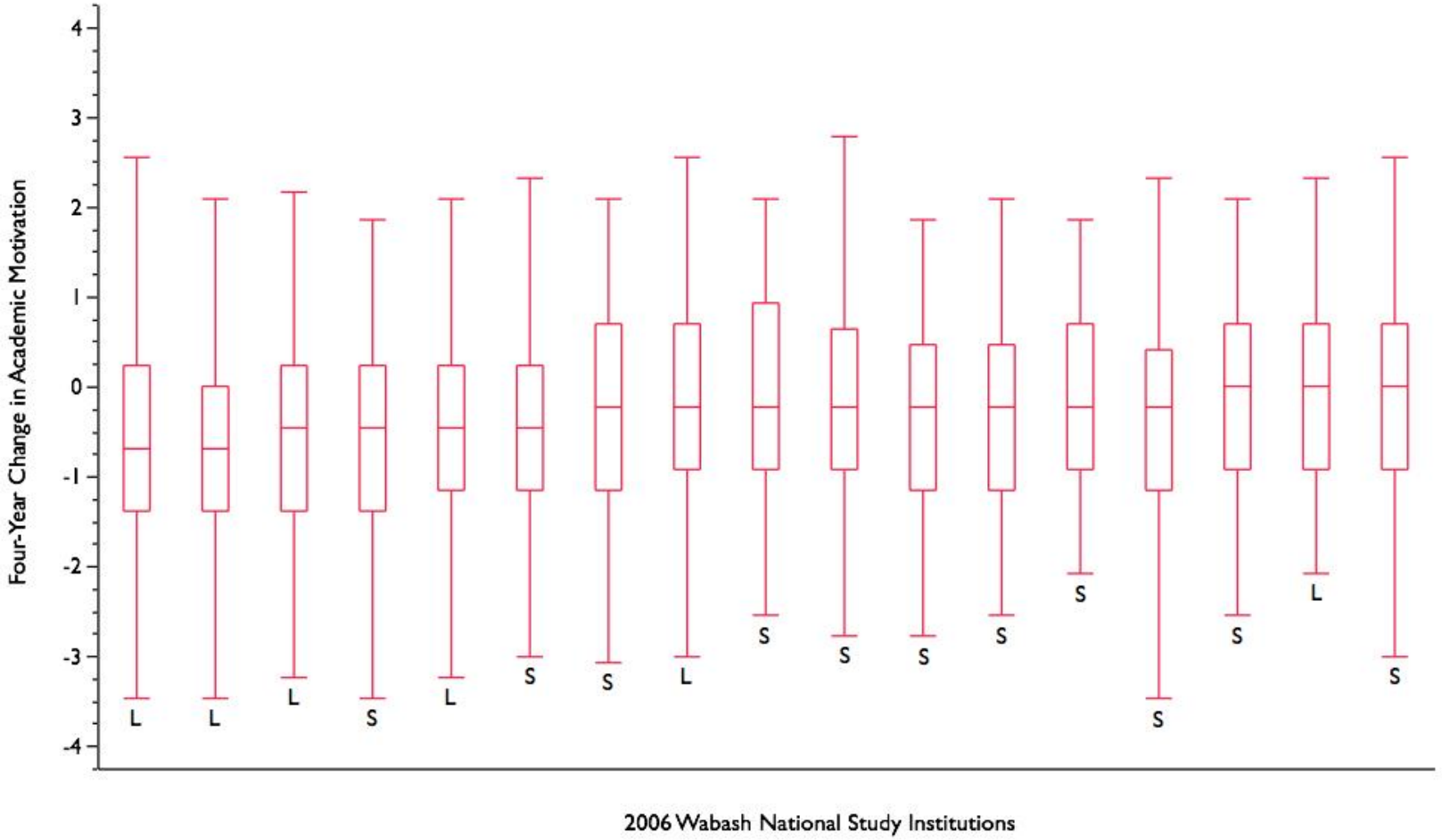
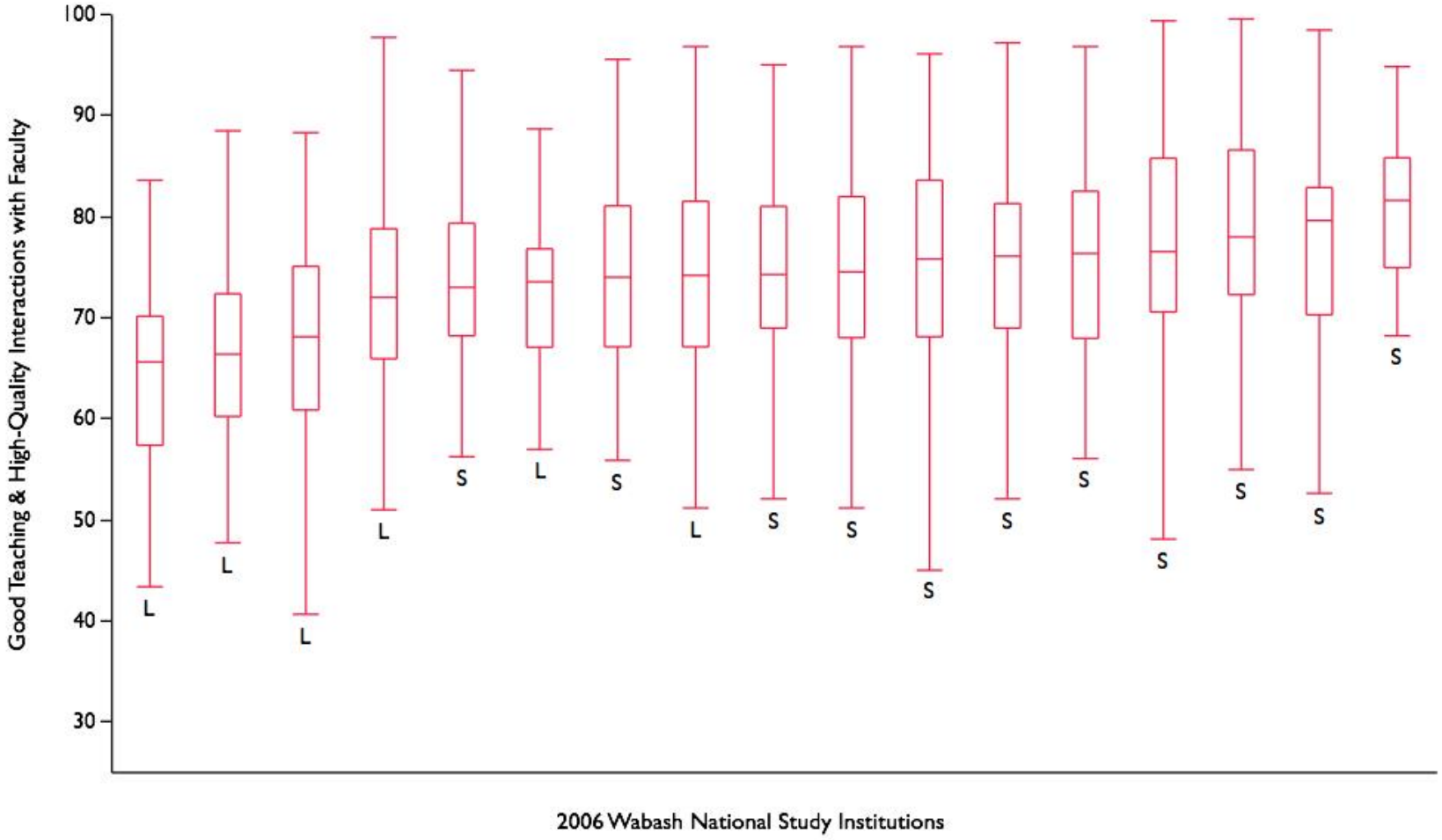
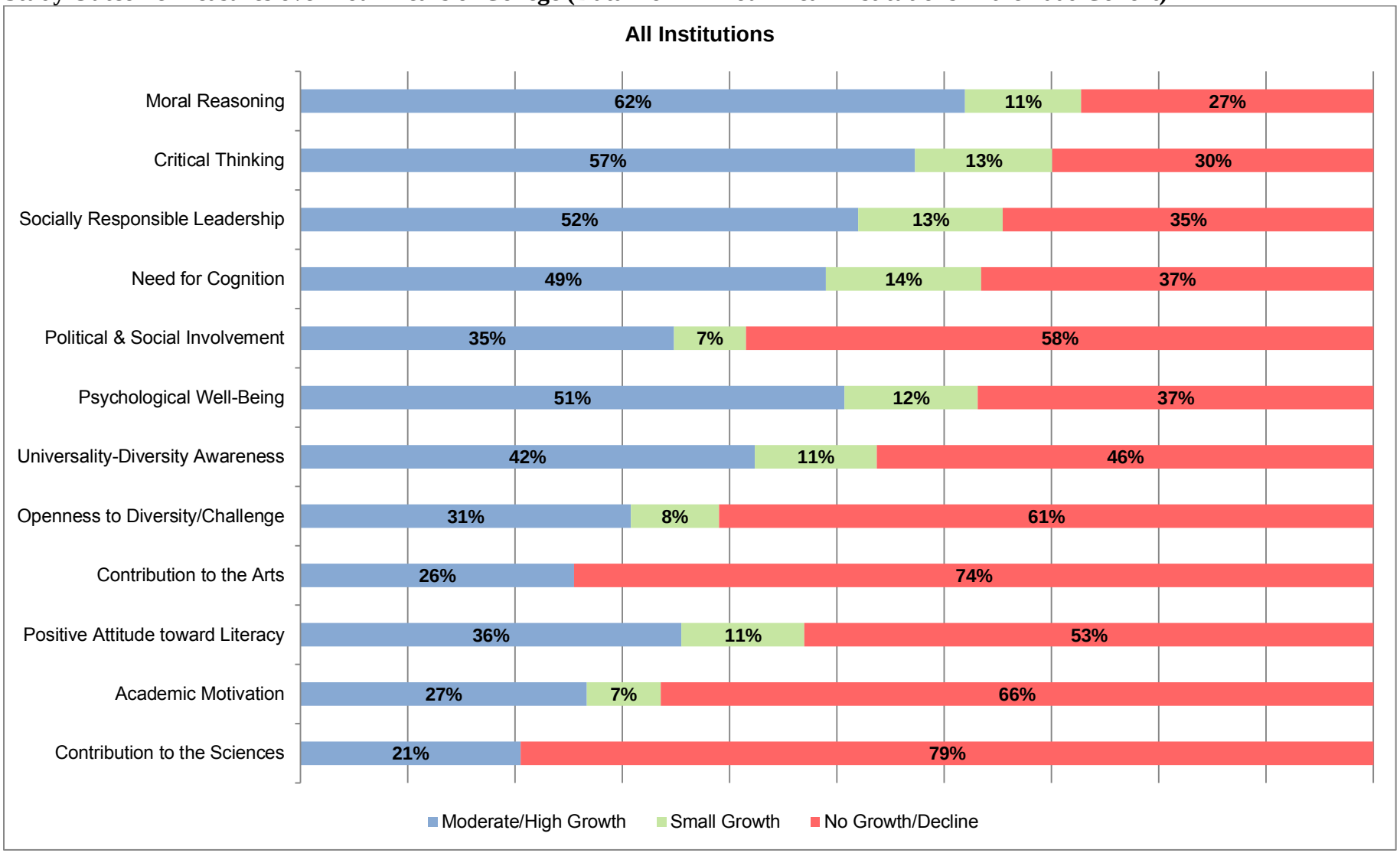


Figure 5. Within-Institution Variation at Small Colleges (S) and Large Universities (L) in Levels of Good Teaching and High-Quality Interactions with Faculty



Additional Four-Year Data from the Wabash National Study

Figure 1. Proportion of Students who had Moderate to High Growth, Small Growth, or No Growth/Decline on Wabash National Study Outcome Measures over Four Years of College (Data from 17 Four-Year Institutions in the 2006 Cohort)



Moderate/High Growth = outcomes change of 0.3 standard deviations or more
Small Growth = outcomes change between 0.05 and 0.3 standard deviations
No Growth/Decline = outcomes change of 0.05 standard deviations or less

The four line graphs below (Figures 2-5) show how students' reported level of engagement in good practices changed over four years, from the end of their first year of college to the end of their fourth year of college. Each graph has two lines, one for students at small institutions and the other for students at large institutions. Data is from the 17 four-year institutions in the 2006 cohort of the Wabash National Study.

Figure 2. Change in Students' Reported Level of Good Teaching and High-Quality Interactions with Faculty over Four Years of College

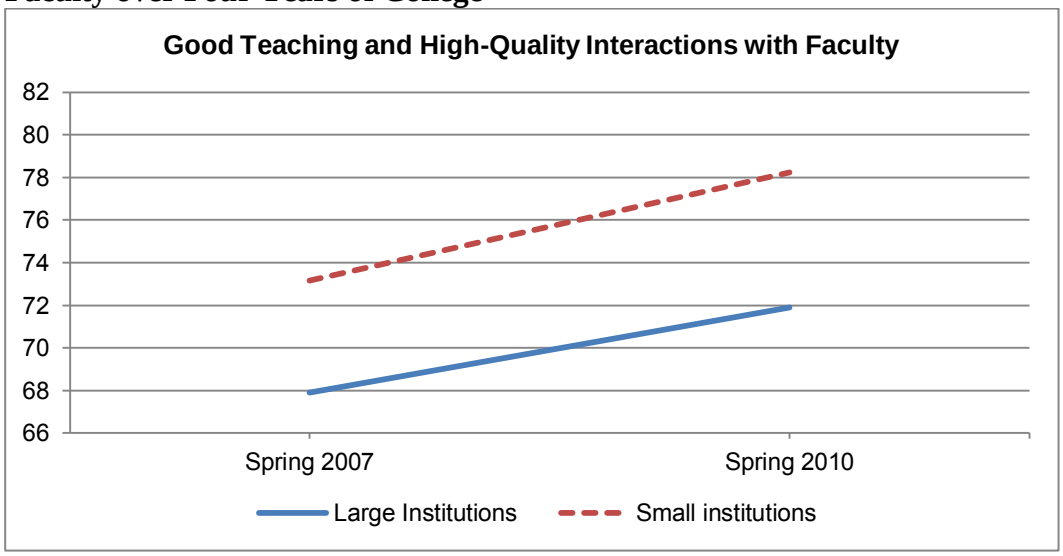


Figure 3. Change in Students' Reported Level of Academic Challenge and High Expectations over Four Years of College

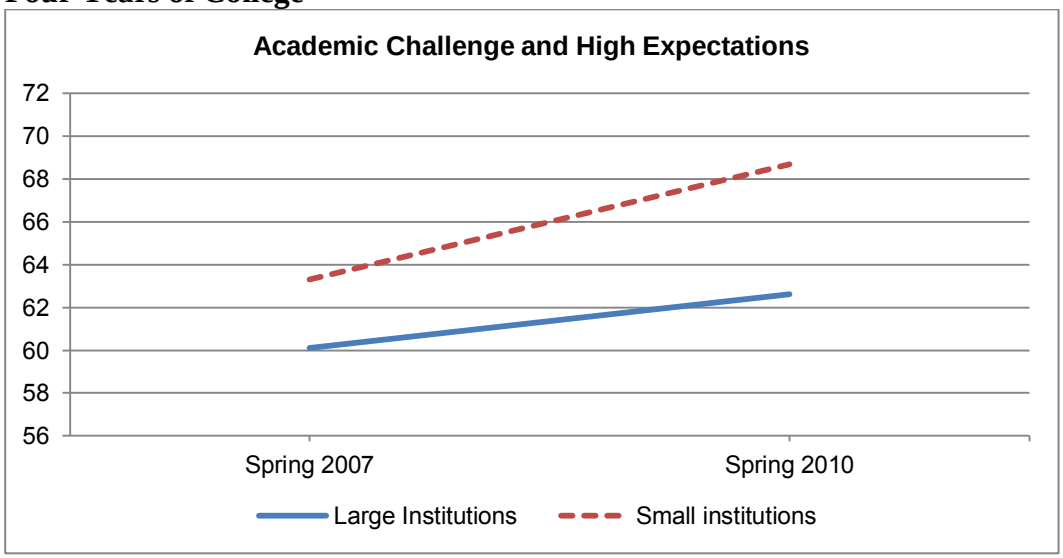


Figure 4. Change in Students' Reported Level of Diversity Experiences over Four Years of College

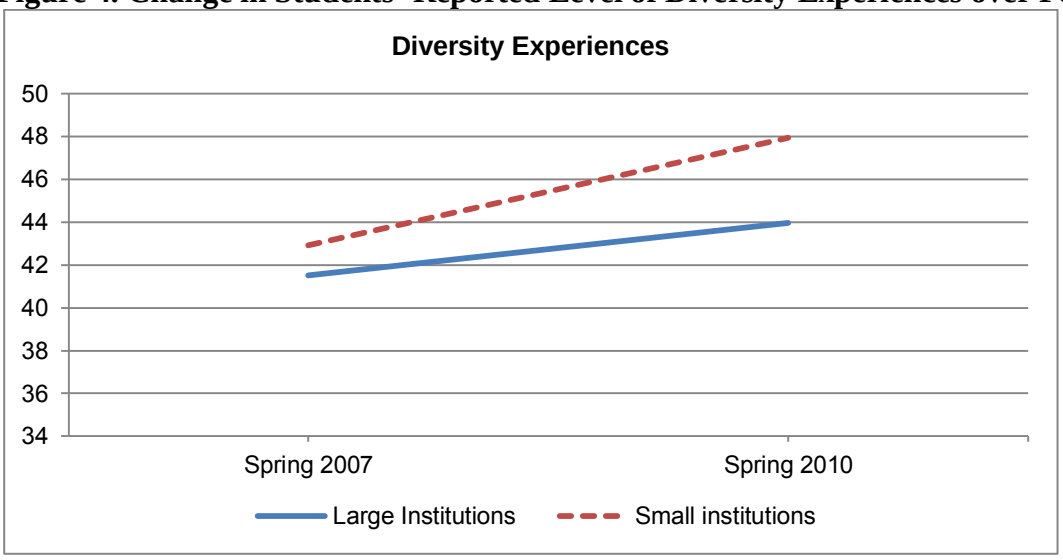
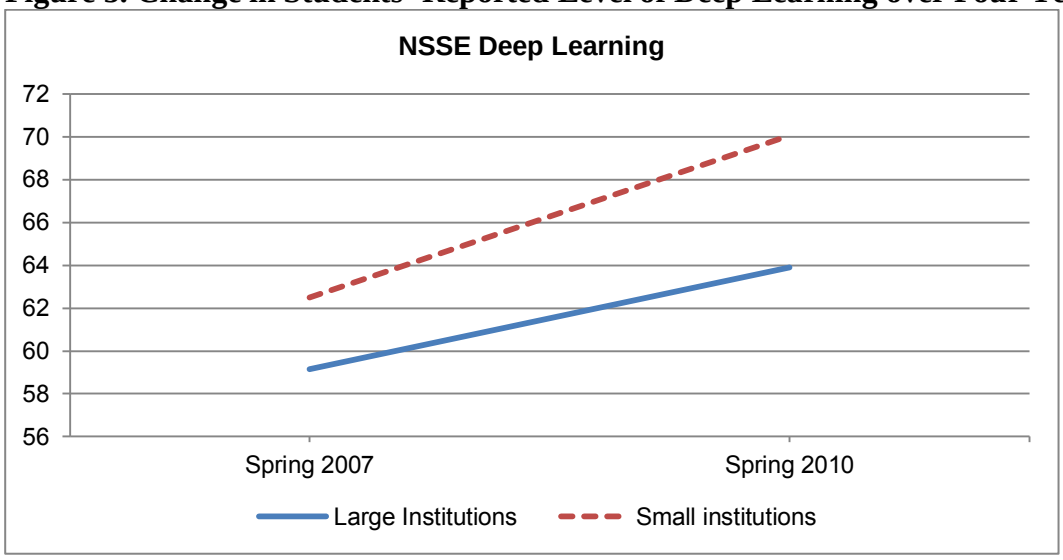


Figure 5. Change in Students' Reported Level of Deep Learning over Four Years of College



Appendix A: Wabash Study Research Methodology

The Wabash National Study of Liberal Arts Education (WNSLAE) is a longitudinal research and assessment project that was designed to deepen our understanding of the teaching practices, student experiences, and institutional conditions that promote the development of students' critical thinking, moral reasoning, leadership towards social justice, well-being, interest in and engagement with diversity, and interest in deep intellectual work. The Wabash Study was also designed to provide participating institutions with high-quality assessment evidence for institutional improvement and accreditation. Institutions entered the Wabash Study in three annual cohorts starting in 2006. The data in the previous NILOA paper are from the 2006 cohort of institutions.

Institutional Sample. The 2006 institutional cohort of the Wabash Study for this paper consisted of incoming first-year students at 17 four-year colleges and universities located in 11 different states from 4 general regions of the United States: Northeast, Southeast, Midwest, and Pacific Coast. There was also substantial variability in undergraduate enrollment, from institutions with entering classes between 3,000 and 6,000, to institutions with entering classes between 250 and 500. According to the 2007 Carnegie Classification of Institutions, 3 of the participating institutions were considered research universities, 3 were regional universities that did not grant the doctorate, and 11 were liberal arts colleges.

Student Sample. The individuals in the 2006 sample were first-year, full-time undergraduate students participating in the WNSLAE at each of the 17 institutions in the study. The initial sample was selected in either of two ways. First, for larger institutions, it was selected randomly from the incoming first-year class at each institution. The only exception to this was at the largest participating institution in the study, where the sample was selected randomly from

the incoming class in the College of Arts and Sciences. Second, for a number of the smallest institutions in the study—all liberal arts colleges—the sample was the entire incoming first-year class.

Initial Data Collection. The initial data collection was conducted in the early fall of 2006 with 4,193 students from the 17 institutions. This first data collection lasted between 90-100 minutes and students were paid a stipend of $50 each for their participation. The data collected included a WNSLAE precollege survey that sought information on student demographic characteristics, family background, high school experiences, political orientation, educational degree plans, and the like. Students also completed a series of instruments that measured dimensions of intellectual and personal development theoretically associated with a liberal arts education. These are described in greater detail in the subsequent section on "WNSLAE Outcomes/Dependent Measures."

Follow-up Data Collection. The follow-up data collection was conducted in spring 2007 and spring 2010. This data collection took about two hours each time and participating students were paid an additional stipend of $50 each during each follow-up session. Two types of data were collected. The first was based on questionnaire instruments that collected extensive information on students' experience of college. Two complementary instruments were used: the National Survey of Student Engagement (NSSE) (Kuh, 2001) and the Wabash Study Student Experiences Survey (WSES). These instruments were designed to capture student engagement in, or exposure to, empirically vetted good practices in undergraduate education. These good practices included such dimensions as: exposure to effective teaching, quality of nonclassroom interactions with faculty, active learning, integrative experiences, influential interactions with

other students, high expectations, and the like (Pascarella, Cruce, Wolniak, & Blaich, 2004; Pascarella, Cruce, Umbach, Wolniak, Kuh, Carini, Hayek, Gonyea, & Zhao, 2006).

The second type of data collected consisted of follow-up (or posttest) measures of the instruments measuring dimensions of intellectual and personal development that were first completed in the initial data collection. All students completed the NSSE and WSES prior to completing the follow-up instruments assessing intellectual and personal development. Both the initial and follow-up data collections were administered and conducted by ACT.

Of the original sample of 4,193 students who participated in the fall 2006 testing, 2,953 participated in the spring 2007 follow-up data collection, for a response rate of 70.4%, and 2,212 participated in the spring 2010 follow-up data collection, for a response rate of 52.8%.

We used the following outcome/dependent measures in the Wabash Study:

- The Collegiate Assessment of Academic Proficiency (CAAP) developed by the American College Testing Program (ACT). The critical thinking test is a 40-minute, 32-item instrument designed to measure a student's ability to clarify, analyze, evaluate, and extend arguments. The test consists of four passages in a variety of formats (e.g., case studies, debates, dialogues, experimental results, statistical arguments, editorials). Each passage contains a series of arguments that support a general conclusion and a set of multiple-choice test items. The internal consistency reliabilities for the CAAP critical thinking test range between .81 and .82 (ACT, 1991).
- The Ryff Scales of Psychological Well-Being (SPWB) (Ryff, 1989; Ryff & Keys, 1995). The SPWB is a 54-item, theoretically-grounded instrument that specifically focuses on measuring six dimensions of psychological well-being: positive evaluations of oneself (Self-Acceptance), sense of continued growth and development as a person (Personal

Growth), belief in a purposeful and meaningful life (Purpose in Life), quality relations with others (Positive Relations with Others), capacity to effectively manage one's life and surrounding world (Environmental Mastery), and sense of self-determination (Autonomy) (Ryff & Keyes, 1995; Ryff, 1989; Keyes, Shmotkin, & Ryff, 2002). The six 9-item scales have internal consistency reliabilities ranging from .83 to .91. We averaged scores on the six scales to create a composite scale of psychological well-being. The composite scale has an internal consistency reliability of .88.

- The 18-item Need for Cognition Scale (NCS). Need for cognition refers to an individual's "tendency to engage in and enjoy effortful cognitive activity" (Cacioppo, Petty, Feinstein, & Jarvis, 1996, p. 197). The NCS has internal consistency reliabilities in the present study ranging from .87 to .90.
- The Positive Attitude Toward Literacy Scale (PATL). The PATL assesses students' enjoyment of such literacy activities as reading poetry and literature, reading scientific and historical material, and expressing ideas in writing, and has an internal consistency reliability of .71.
- The 15-item, short form of the Miville-Guzman Universality-Diversity Scale (M-GUDS). The M-GUDS measures an individual's universal-diverse orientation, which is defined as "an attitude of awareness and acceptance of both similarities and differences that exist among people" (Miville, Gelso, Pannu, Liu, Touradji, Holloway, & Fuertes 1999; Fuertes, Miville, Mohr, Sedlacek, & Gretchen, 2000). The instrument has a total scale score and three subscale scores: Diversity of Contact (interest and commitment to participating in diverse, intentionally focused social and cultural activities), Relativistic Appreciation (appreciation of both similarities and differences in people and the impact

of these in one's self-understanding and personal growth), and Comfort with Differences (the degree of comfort with diverse individuals). The internal consistency reliability for the total M-GUDS score in the present study was .85, while reliabilities for the three subscales ranged from .77 to .78.

- The seven-item Openness to Diversity/Challenge (ODC) scale. This scale measures one's openness to cultural and racial diversity as well as the extent to which one enjoys being challenged by different perspectives, values, and ideas (Pascarella, Edison, Nora, Hagedorn, & Terenzini, 1996). The ODC has internal consistency reliabilities in the present study ranging from .83 to .87.
- The 68-item, revised version II of the Socially Responsible Leadership Scale (SRLS). The SRLS measures the eight dimensions of Astin's Social Change Model of leadership development (Astin, A., Astin, H., Boatsman, Bonous-Hammarth, Chambers, Goldberg, et al., 1996). According to this model, leadership is a collaborative group process directed toward promoting positive social change in an organization or community (Tyree, 1998). A person who demonstrates strong socially responsible leadership capabilities is self-aware, acts in accordance with personal values and beliefs, invests time and energy in activities that he or she believes are important, works with diverse others to accomplish common goals, has a sense of civic and social responsibility, and desires to make world a better place. The SRLS was developed specifically to measure leadership in college students. The instrument has eight scales corresponding to the eight dimensions of leadership specified in the Astin model (Astin, et al., 1996; Dugan, 2006). The eight scales are: Consciousness of Self (being aware of the values, emotions, attitudes, and beliefs that motivate one to take action); Congruence (thinking, feeling, and behaving

with consistency, genuineness, authenticity, and honesty toward others); Commitment (intensity and duration in relation to a person, idea, or activity—the energy and passion that propels one to act); Collaboration (working with others in a common effort); Common Purpose (working with others within a shared set of aims and values); Controversy with Civility (recognizing two fundamental realities of any group effort, that (a) differences of viewpoint are inevitable and valuable, and (b) such differences must be aired openly and with respect and courtesy; Citizenship (believing in a process whereby a person or group is responsibly connected to the environment and the community); and Change (adapting to continuously evolving environments and situations, while maintaining the primary functions of the group). We averaged scores on the eight scales to create a composite scale of socially responsible leadership. The composite scale has an internal consistency reliability of .84.

- The Defining Issues Test 2 (DIT2). The DIT2 is a revised version of James Rest's original DIT from 1979 that measures one component of moral development, known as moral judgment or reasoning (Rest, Narvaez, Thoma, & Bebeau, 1999). Our analyses focused on the N2 score. The internal consistency reliability for the N2-score ranges from .77 to .81.
- Three scales based on 16 items from the Cooperative Institutional Research Program (CIRP) Survey developed by the Higher Education Research Institute at the University of California at Los Angeles.
    - Contributions to the Arts Scale—a three-item scale in which respondents identify how important (ranging from not important to essential) it is for them to contribute to the arts. Contributing to the arts included "becoming accomplished

in the performing arts," "writing original works," or "creating artistic work." The internal consistency reliability for the scale is .69.

  - Political and Social Involvement Scale—an eleven-item scale in which respondents identify how important (ranging from not important to essential) it is for them to be involved politically and socially in communities. Political and social involvement included such activities or goals as "influencing the political structure," "influencing social values," "becoming a community leader," and "volunteering in my community." The internal consistency reliability for the scale ranges from .80 to .83.
  - Contribution to Science—a two-item scale which asks respondents to indicate how important it is for them to contribute to advances in scientific fields. Contribution to science includes "making a theoretical contribution to science" and "working to find a cure for a disease or illness." The internal consistency reliability for the scale ranges from .70 to .76

- The eight-item Academic Motivation Scale in which respondents were asked to indicate the extent to which they agree or disagree (ranging from strongly agree to strongly disagree) with statements about their academic motivation. The internal consistency reliability for the scale ranges from .69 to .74.

With two exceptions, each Wabash Study outcome/dependent measure discussed above was completed by all participants during both the initial data collection in fall 2006 and the follow-up data collections in spring 2007 and spring 2010. The two exceptions were the CAAP Critical Thinking Test and the Defining Issues Test. Each of these instruments took at least 40

minutes to complete, and because we were concerned with the amount of time required of students during each data collection, the CAAP Critical Thinking Test and the Defining Issues Test were not taken by all participants. Rather, during the first data collection the study participants were randomly divided into two approximately equal samples. The CAAP Critical Thinking Test was then taken by one random sample during all three data collections and the Defining Issues Test taken during all three data collections by the other random sample.

Consistent with the data-reduction procedures of Cruce, et al., (2006), we submitted the scales and items from NSSE and WSES to similar factor analytic procedure. Six factors appeared to underlie the individual measures of good practices/liberal arts experiences in the WNSLAE. They were titled: "Good teaching and high quality interactions with faculty," "Academic challenge and high expectations," "Diversity experiences," "Influential interactions with peers," "Frequency of interactions with faculty/professional staff," and "Cooperative learning." Scales were constructed for each factor by first standardizing each item from the individual good practice/liberal arts experience dimensions loading on the factor, and then computing the mean score. In this paper we focus on the first three of these scales.

- Good teaching and high quality interactions with faculty is a 23-item scale that combined items from four subscales: Faculty interest in teaching and student development (e.g., the extent to which faculty are interested in helping students grow in more than just academic areas, the extent to which faculty are generally interested in teaching, and the extent to which faculty are willing to spend time outside of class to discuss issues of interest and importance to students); Prompt feedback (e.g., how often faculty informed students of level of performance in a timely manner, how often faculty checked to see if students had learned the material well before going on to new materials); Quality and impact of

nonclassroom interactions with faculty (e.g., extent to which nonclassroom interactions with faculty have had an impact on: intellectual growth and interest in items; personal growth, values, and attitudes; and career goals and aspirations); and Overall exposure to clear and organized instruction (e.g., frequency that faculty give clear explanation, frequency that faculty make good use of examples and illustration to explain difficult points, frequency that class time was used effectively, frequency that course goals and requirements were clearly explained). The internal consistency reliability for the 23-item scale is .92.

- Academic challenge and high expectations was a 31-item scale that combined items from four subscales: Academic challenge and effort (e.g., how often one worked harder than one thought he or she could to meet an instructor's standards or expectations, number of hours a week spent preparing for class, extent to which one's institution emphasizes spending significant amounts of time studying and on academic work, number of assigned textbooks, books, or book-length packs of course readings one read during current year); Frequency of higher-order exams and assignments (e.g., how often exams or assignments require students to: write essays, compare or contrast topics or ideas from a course, argue for or against a particular point of view and defend an argument); Challenging classes and high faculty expectations (e.g., how often faculty: asked challenging questions in class; challenged students' ideas in class; asked students to argue for or against a particular point of view; asked students to point out any fallacies in basic ideas, principles, or points of view presented in the course); and Integrating ideas, information, and experiences (e.g., extent to which one agrees that courses have helped him or her understand the historical, political, and social connections of past events; how

often one has worked on a paper or project that required integrating ideas or information from various sources; how often one has put together ideas or concepts from different courses when completing assignments or during class discussions). The internal consistency reliability for the 31-item scale is .88.

- Diversity experiences was a 9-item scale that combined items from two subscales: Diversity experiences (e.g., extent to which one's institution encourages contact among students from different economic, social, and racial or ethnic backgrounds; how often one had serious conversations with students of a different race or ethnicity than one's own; how often one participated in a racial or cultural awareness workshop during the academic year); and Meaningful discussions with diverse peers (e.g., how often one had meaningful and honest discussions about issues related to social justice with diverse students, how often one had discussions regarding intergroup relations with diverse students). The internal consistency reliability for the 9-item scale is .80.

We also used the 12-item scale from the NSSE called the Deep Learning Scale. This includes information from three subscales: Higher order learning (e.g., how much one's coursework emphasized analyzing the basic elements of an idea, experience, or theory), Integrative learning (e.g., how often one has worked on a paper or project that required integrating ideas or information from various sources), and Reflective learning (e.g., how often one has examined the strengths and weaknesses of his or her own views on a topic or issue). The internal consistency reliability for the 12-item scale is 0.82.

We have converted all scores on our Good Practice and Deep Learning Scales to 100 point scales by calculating the following score for each item:

((Response on Likert Scale – 1)/(Number of Response Options – 1))*100

In our previous research (see for example, Seifert, Goodman, Lindsay, Jorgensen, Wolniak, Pascarella, et al., 2008; Seifert, Pascarella, Goodman, Salisbury, & Blaich, 2010), we have found these four scales to be correlated with growth on most of our outcomes even after controlling for students' background characteristics and their scores on our outcome measures when they entered college.

References

American College Testing Program (ACT). (1991). *CAAP technical handbook*. Iowa City, IA: Author.

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.

Cacioppo, J., Petty, R., Feinstein, J., & Jarvis, W. (1996). Dispositional differences in cognitive motivation: The life and times of individuals varying in need for cognition. *Psychological Bulletin, 119,* 197-253.

Cruce, T., Wolniak, G., Seifert, T., & Pascarella, E. (2006). Impacts of good practices on cognitive development, learning orientations, and graduate degree plans during the first year of college. *Journal of College Student Development, 47,* 365-383.

Dugan, J. (2006). Explorations using the social change model: Leadership development among college men and women. *Journal of College Student Development, 47,* 217-225.

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.

Keyes, C., Shmotkin, D., & Ryff, C. (2002). Optimizing well-being: The empirical encounter of two traditions. *Journal of Personality and Social Psychology*, *82*, 1007-1022.

Kuh, G. (2001). Assessing what really matters to student learning: Inside the National Survey of Student Engagement. *Change*, *33*(3), 10-17, 66.

Miville, M., Gelso, C., Pannu, R., Liu, W., Touradji, P., Holloway, P., Fuertes, J. (1999). Appreciating similarities and valuing differences: The Miville-Guzman Universality-Diversity Scale. *Journal of Counseling Psychology*, *46*(3), 291-307.

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., Cruce, T., Umbach, P., Wolniak, G., Kuh, G., Carini, R., Hayek, J., Gonyea, R., & Zhao, C. (2006). Institutional selectivity and good practices in undergraduate education: How strong is the link? *Journal of Higher Education*, *77*, 251-285.

Pascarella, E., Cruce, T., Wolniak, G., & Blaich, C. (2004). Do liberal arts colleges really foster good practice in undergraduate education? *Journal of College Student Development*, *45*, 57-74.

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.

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

Ryff, C., & Keyes, C. (1995). The structure of psychological well-being revisited. *Journal of Personality and Social Psychology, 69,* 719-727.

Seifert, T. A., Goodman, K. M., Lindsay, N. K., Jorgensen, J. D., Wolniak, G. C., Pascarella, E. T., et al. (2008). The effects of liberal arts experiences on liberal arts outcomes. *Research in Higher Education, 49,* 107–125.

Seifert, T. A., Pascarella, E. T., Goodman, K. M., Salisbury, M. H., & Blaich, C. F. (2010, January/February). Liberal arts colleges and good practices in undergraduate education: Additional evidence. *Journal of College Student Development, 51,* 1–22.

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.