2. ACCOUNTABILITY, ASSESSMENT, AND THE SCHOLARSHIP OF "BEST PRACTICE"

Alicia C. Dowd* and Vincent P. Tong
University of Southern California, and Gateway Community College

ACKNOWLEDGEMENTS

Research for this paper was conducted with funding from Lumina Foundation for Education. Earlier versions were presented at the Northeast Association for Institutional Research (NEAIR) annual conference, Portsmouth, NH, November 2004 and at the Measuring What Matters Symposium of the University of Massachusetts Boston's Community College Success Project (CCSSP), Roxbury Community College, Boston, MA, October, 2004.

The authors appreciate insightful discussions and feedback on earlier drafts from members of the CCSSP's National Advisory Council and New England Resource Center for Higher Education (NERCHE) Think Tank (Fall-Spring 2004–2005), the Cultivating Communities of Inquiry Symposium (January, 2006, Los Angeles, CA), and participants in the "New Era of Accountability" seminar at the Institute for Community College Development, Cornell University (August, 2006). The authors recognize valuable research assistance provided by Susan Dole, Shelley Fortin, John Grant, Linda Kent Davis, and Randi Korn. Insightful conversations with Estela Mara Bensimon, Robert Rueda, and other researchers at the Center for Urban Education at the Rossier School of Education were particularly valuable in informing our views. Ozan Jaquette and three anonymous reviewers provided extremely helpful critiques. The arguments expressed in this paper are those of the authors and do not represent those of Lumina Foundation, Community

* Rossier School of Education, WPH 702A, 3470 Trousdale Parkway, University of Southern California, Los Angeles, CA 90089–0031, USA. Email: alicia.dowd@usc.edu

J.C. Smart (ed.), Higher Education: Handbook of Theory and Research, Vol. XXII, 57–119.

College Student Success Project participants, or other reviewers. All errors remain our own.

Political concern for the productivity, efficiency, and quality of higher education is shaping the ways in which academic and institutional researchers evaluate the effectiveness of collegiate programs and institutions. The search is on to identify "best practices" that will increase educational productivity. To motivate that search, accountability standards tend to emphasize the importance of quantitative indicators of student learning outcomes and de-emphasize the need to understand educational processes and institutional contexts. Though a focus on student learning outcomes is essential to the study of institutional effectiveness, it is not sufficient in itself to spur productive change and innovation. In this chapter, we argue that the recent history of legislative accountability (and institutional responses to it), developments in educational research methodology, and theories of learning and professional practice demonstrate the need for equal attention under accountability to educational processes and contexts alongside the measurement of student outcomes.

To address our concern that the search for "best practices" in its current form will be ineffectual, we propose the creation of evidence-based inquiry councils (EBICs) as a central feature of a comprehensive accountability system designed to integrate knowledge of institutional context, educational processes, and learning outcomes for the purpose of increasing the educational effectiveness of colleges and universities. EBICs, as proposed here, are distinguished by an integrated scholarship involving academic and institutional researchers with an inter-related focus on institutional processes (including resource use) and student learning outcomes.

The evidence-based inquiry councils are intended to capitalize on existing features of assessment and accreditation systems, such as self-studies and campus review teams, in support of accountability goals. The proposed EBIC design addresses two primary weakness of results-based accountability plans, namely (1) lack of a clear strategy to understand "what works" in a variety of higher educational contexts and (2) a clear mechanism to promote the adoption by administrators and faculty of educational practices identified as what are often called "best practices." The educational programs and practices that prove to be effective in one setting may not be in another, or may not be effective in the same ways, so we prefer to call these "effective" or "exemplary" practices, meaning they are worthy of examination for potential adoption in other settings. Knowing that

educational contexts are highly variable, practitioners may be legitimately dubious or hesitant about adopting new approaches to improve student outcomes, even if those approaches have been warranted by rigorous academic research. The design of the inquiry councils, as proposed here, takes into account the different decision-making and information needs of policymakers and educational practitioners. We prescribe an integrated set of evaluation strategies to meet those needs.

The search for institutional "best practices" is evident today in federal and state accountability initiatives (Dwyer, Millett, & Payne, 2006; Erisman & Gao, 2006; U.S. Department of Education, 2006; "What Works Clearinghouse," 2006) and in the philanthropic priorities of foundations with a focus on higher education (see, for example, Dowd, 2005; Dowd et al., 2006; Lagemann, 2002; *Lumina Foundation*, 2006; *What we know*, 2006; Wyner, 2006). Our views are informed by participation in the Community College Student Success Project at the University of Massachusetts Boston, which was initiated with funding from Lumina Foundation for Education. The project involved higher education administrators, faculty members, and institutional researchers in a series of symposia and working papers which reframed the search for scientifically warranted best practices as a search for best practices *of assessment*. The distinction stems from our beliefs that the emphasis of accountability should be on creating a scholarship of effective assessment practices rather than on the identification of specific practices or programs as a toolkit for educational improvement. The entire accountability movement has been decidedly in the opposite direction, a point we explore by discussing the accountability environment for academic research and institutional research. In our view, the scholarship of best practice should be a scholarship of professional development and learning among higher education practitioners that integrates the work of practitioners-as-researchers and academic researchers as facilitators of learning. That said, we see the study of the causal-effectiveness of educational programs using traditional research methods as one of many essential elements of this broader research program.

Our consideration of the current state and future prospects of a "scholarship of best practice" centered on learning and the development of a "culture of inquiry" (Dowd, 2005) is presented to inform academic and institutional researchers, policymakers, and officials of funding agencies who are interested in increasing the quality and productive capacity of higher education. Our argument is presented in

five following sections, with references to the relevant academic and policy literature throughout. The first section provides a brief review of studies of the effectiveness of accountability in bringing about changes in institutional behavior. The second section discusses recent controversies in academic research concerning federal research standards that focus nearly exclusively on experimental and quasi-experimental methods as the kind of research needed to increase educational effectiveness. Consistent with our emphasis on practitioner learning, we argue that evaluation designs using a variety of methods are better suited to achieving accountability goals because ethnographic and case study methods are necessary to develop the "ordinary knowledge" (Lindblom & Cohen, 1979) and "practical wisdom" (Aristotle, cited in Polkinghorne, 2004) of effective practitioners.

Just as federal research policy places high value on quantitative methods and analyses, state accountability policies place great value on quantitative indicators of institutional performance. The accountability climate and the challenges of educational problem solving are similar for academic researchers and institutional researchers. In both settings, we perceive a risk that accountability policies will problematically obscure rather than illuminate the dynamic, context-sensitive nature of teaching, learning, and educational administration. The third section presents the basis for this concern in the realm of institutional research and assessment by comparing the methods and premises of performance, diagnostic, and process benchmarking. While accountability systems do not necessarily involve benchmarking, they often do, and peer benchmarking is the preferred strategy of policymakers hoping to spur institutional change and innovation (see, for example, U.S. Department of Education, 2006) The circumstances under which innovation is likely to occur through benchmarking activities are discussed, emphasizing concepts of individual and organizational learning. This section and the previous one demonstrate that both academic researchers and institutional researchers are responding to accountability requirements by emphasizing the importance of understanding educational processes within specific institutional contexts in order to determine "what works."

The fourth section describes the structure of the proposed evidence-based inquiry councils, which are characterized by the study of student learning and educational outcomes integrated with the self-study of practitioner knowledge and learning. The research and evaluation methods of the EBIC are intended to facilitate synthesis by practitioners of evidence from quantitative and qualitative data

analyses. The purpose of the EBICs is to understand how, why, and when educational practices are effective. Finally, a brief concluding section summarizes our main argument.

I. ACCOUNTABILITY TRENDS AND EFFECTIVENESS

Competing Priorities of Multiple Stakeholders The higher education news has prominently featured a national debate among legislators, business leaders, higher education officials, and higher education associations about the means by which colleges and universities should be held accountable for educating their students (Field, 2005a, 2005b; Selingo, 2006; Strout, 2004). As Congress drafted revisions to the Higher Education Act, which sets federal higher education policy, Republican leaders and the U.S. Secretary of Education's Commission on the Future of Higher Education argued that college accreditation reviews should become part of the public domain, with information about college effectiveness and quality made readily available to the public (Field, 2006; Fleming, 2004; Lederman, 2006; U.S. Department of Education, 2006).

Informed by business perspectives emphasizing consumer satisfaction and economic competitiveness, the watchwords of draft reports issued by the Secretary's Commission on the Future of Higher Education in the Summer of 2006 were "quality" and "innovation." Consumer access to college performance data was highlighted as an important lever for bringing about improvements in educational efficiency and effectiveness. In addition, the accreditation system was viewed as too focused on assessing educational processes, with insufficient attention to "bottom line" results (U.S. Department of Education, 2006). Both the Spellings Commission report, which called for a "culture of accountability," (p. 20) and a proposal for a comprehensive higher education testing system issued by the Eductional Testing Service, which called for a "culture of evidence" (Dwyer et al., 2006), advocated using the accreditation system to increase higher education accountability. In presenting evidence-based inquiry councils as an accountability structure, we also propose to link accreditation to accountability but in a manner that values the existing focus of accreditation on promoting effective assessment processes.

These proposals to subsume the accreditation system to the accountability agenda represent a striking change from the existing voluntary and confidential system of higher education accreditation (Bollag, 2004). While unwelcome to many higher education leaders,

these proposals were not unexpected. The advance of public accountability pressures into the previously private domain of accreditation was one step in an ongoing struggle to document the quality and productivity of colleges and universities (Zumeta, 2001). The tensions stem from contrasting values outside and within academia and from the inherent difficulty of the task of quality measurement. External values include institutional efficiency and quality, adherence to performance standards, comparability of measured outcomes, and public reporting to higher education consumers. Internal values also include institutional quality, but then diverge towards an emphasis on mission differentiation, a focus on process improvement rather than comparative outcome standards, and confidentiality of results for internal review (Burke et al., 2002; Burke & Minassians, 2003; Ewell, 1991, 2002; Moore, 2002).

The accountability debate raises questions regarding the appropriate reach of state and federal governments into the operation, management, and core educational activities of higher education. Trow (1996) defined accountability simply as "the obligation to report to others, to explain, to justify, to answer questions about how resources have been used, and to what effect" (p. 310). Based on this definition, he raised the following set of fundamental questions: "Who is to be held accountable, for what, to whom, through what means, and with what consequences?" (p. 310). Burke (2005a) drawing on Trow's framing questions described an "accountability triangle," with state priorities, academic concerns, and market forces as the three basic stakeholders of accountability. Like Clark's (1983) earlier work, which similarly identified state control, an academic oligarchy, and market models as the three forces of contention in the field of higher education, Burke's accountability triangle provides an apt instrument with which to survey the forces of contention in higher education accountability. For example, assessment (e.g. Kuh, 2005), accreditation (e.g. Wolff, 2005), and academic audits (e.g. Massy, 2005) can be located in the corner of academic concerns. State-by-state report cards (Callan & Finney, 2005), the "rules in use" of higher education governance (Richardson & Smalling, 2005), and performance reporting (Burke, 2005b) are in the corner of state priorities; and reputational ratings, such as those in U.S. News and World Report (Volkwein & Grunig, 2005) in the corner of market forces.

Under a paradigmatic shift towards a new form of governmental accountability, in the early 1980s legislators sought to create measurable goals and financial incentives to spur improved results,

as well as to motivate colleges to improved performance through concern for their public relations and market shares (Burke et al., 2002; Dougherty & Hong, 2005). While the external accountability movement stressed institutional efficiency and the establishment of performance standards, during the same era an internal assessment movement focused on the improvement of teaching and learning environments.

While the prevailing political ideologies of accountability has been strongly steeped in business perspectives concerning the need for efficiency, productivity, and economic competitiveness (Alexander, 2000; Ayers, 2005; Dowd, 2003; Leveille, 2005; J. S. Levin, 2001), higher education scholars have warned that business market models cannot be adopted wholesale to the regulation of higher education because they are essentially different organizational enviroments (Dill, 2003; Gumport & Pusser, 1995; Toutkoushian & Danielson, 2002; Zemsky, 2005). Dill (2003) and Zemsky (2005) pointed out that the current market conditions of higher education are not typical of business markets. For example, the higher education market is imperfect because consumers cannot differentiate the quality of institutions that are "reputation-" and "prestige-" oriented (Dill, 2003). Not surprisingly, the academic community has been hesitant to identify specific or uniform indicators of institutional performance as bottom line indicators of performance (Burke et al., 2002) and instead emphasized the diversity of institutional missions and broadly defined learning outcome goals (*Greater expectations*, 2002; *Our students best work*, 2004).

Uncertain Effects of Accountability Through a series of surveys of state higher education officials conducted at the Rockefeller Institute, Burke and colleagues have documented the changing forms and uncertain effects of accountability on institutional behavior (Burke et al., 2002; Burke & Minassians, 2003). Using the typology of state-required performance reporting (requiring assessments and reporting of performance indicators), performance budgeting (loosely tying institutional performance to budget priorities), and performance funding (allocating funds based on performance), they have shown that almost all states have experimented with some form of accountability and that legislative preference for these different accountability mechanisms have changed over time. Lawmakers attempted to strengthen early performance reporting requirements, which were ultimately viewed as ineffective, by tying performance to funding. However, the amount of money involved was typically small, in the neighborhood of one-half

to six percent of state funds, and states had a tendency to cancel or suspend performance funding when budgets were tight (Burke et al., 2002, pp. 27, 32–34). More recent state-level case studies of community colleges (Dougherty & Hong, 2005; Erisman & Gao, 2006) provide further evidence that performance budgeting and funding had little financial impact on colleges and that these requirements have lost favor in policy circles compared to performance reporting, which has reemerged as the favored accountability mechanism.

The uneven performance of state accountability requirements on institutional performance led Burke and colleagues to characterize them as "symbolic policies," which "appear to address problems, while having little substantive effect" (Burke & Minassians, 2003, p. 14). Similarly, a report from the Institute for Higher Education Policy, based on interviews with representatives of state higher education agencies and system offices in eight states, concluded that accountability data do not typically inform or drive state policy, largely due to a disconnection between performance indicators and policy goals (Erisman & Gao, 2006). Dougherty and Hong (2005) summarize their findings by noting that the effect of performance accountability on colleges is "uneven," "at best moderately strong," and perhaps undermined by "significant negative unintended outcomes" (p. 12). The policies affected colleges by raising their awareness of state priorities, increasing attentiveness to their own performance, and raising concerns about public perceptions of their quality. However, the case study results provide only weak evidence concerning the effect of performance accountability on the ultimate goal of raising student outcomes and those results were mixed. The possibility that increases in retention and graduation rates, where observed, were due to reduced academic standards could not be ruled out. The almost "mythical powers" of reform ascribed by some to the use of performance indicators as a mechanism for change have not been observed in institutional responses to accountability (Toutkoushian & Danielson, 2002, p. 206) These studies provide evidence that performance accountability as implemented to date in the U.S. has not been successful in improving the performance of colleges and universities. They signal the need for deeper examination of the means and goals of legislated higher education accountability.

According to Burke (2005c), the ideal accountability system will be located at the center of the accountability triangle, implying the equal role of all three types of stakeholders. The recent history of accountability systems adopted, implemented, modified, discontinued, and recycled is a strong sign of the search for that ideal as well as

the competing push and pull of stakeholders with oftentimes different values, goals, and cultural norms. Surveying this field of contention, a recent report by the National Commission on Accountability in Higher Education emphasized the need for new approaches to accountability, particularly ones based on democratic participation and shared responsibility among faculty, administrators, and legislators. (*Accountability for Better Results*, 2005). Our proposal for evidence-based inquiry councils incorporates the emphasis of accountability on evidence-based decisionmaking and the emphasis of assessment on professional judgment. Recognizing the very real challenges facing higher education, the EBIC design is intended to promote shared responsibility among stakeholders as the best strategy for addressing those problems.

Although tuition charges are an ongoing point of contention, the key problem facing higher education is how to educate large numbers of students with diverse levels of academic preparation (many significantly underprepared), speaking many different native languages, and often attending college part time while juggling home and work responsibilities. In many states demand exceeds capacity at the same time public resources are declining, so these educational challenges are to be met with the same or fewer resources. As a result, accountability perspectives have emphasized doing more with less, which implies efficient use of resources.

If news of poor performance in graduating or retaining students motivates practitioners to want to adopt different educational approaches or to change the way they interact with students, they also need to see how current practices waste resources in order to become more effective and efficient. This is difficult when it is a matter of seeing ones own culture, values, and behavior. Theories of practice in the "caring professions," such as education, indicate that practitioners learn through "intelligent inquiry" when faced with having to resolve errors in their own judgment (Polkinghorne, 2004). It follows that college administrators and faculty will need to discover how to educate more of the diversely prepared students more cost-effectively. Accountability can foster such discoveries, perhaps using many of the policies now in place, such as the collection of data on graduation rates, if it is not antagonistic to practitioner inquiry.

However, all research and evaluation is political (Chatterji, 2005; Lather, 2004; Weiss, 1975), and our culture currently places a high value on the technical knowledge of experts (Polkinghorne, 2004). The climate of accountability for academic research and institutional research, using the latter term for simplicity of exposition in a way

that encompasses institutional self-study and assessment initiatives, is similar because they exist in the same culture. Emphasis is placed on data, quantitative measures, and statistics as key mechanisms of accountability. The problematic aspects of this emphasis on technical rationality have been explicitly argued in the academic research literature, where academics schooled in other epistemological traditions point out that what people know and believe and how they behave is mediated through social interaction and is context-dependent (Tharp & Gallimore, 1988). Within the world of institutional research and assessment, this point of view is more often voiced as concern for a lack of attention to the diversity of educational contexts and students, which, it is argued, cannot be appropriately captured in uniform indicators of institutional performance.

In the following two sections, we take a broad view of the culture of accountability as we see it manifested in the worlds of academic and institutional research to explain what we perceive as a clear need for an accountability structure that combines evidence-based decisionmaking and practitioner inquiry. Our review emphasizes that in both realms the fundamental challenge is developing methods for understanding and evaluating dynamic, in the sense of highly interactive, educational contexts.

II. POLITICAL AND SOCIAL CONTEXT OF EVALUATING "WHAT WORKS"

As Lather (2004) has observed "science is, like all human endeavor, a cultural practice and practice of culture" (p. 28). The current political view of the culture of science is perhaps best epitomized by the fact that the No Child Left Behind Act of 2001 (HR1) reauthorizing the Elementary and Secondary Education Act references "scientifically-based research" 111 times, gaining it "acronym status inside the Beltway" as "SBR" (Feuer, Towne, & Shavelson, 2002, p. 4). The No Child Left Behind (NCLB) Act also requires that schools use educational methods shown to be effective through scientifically-based research.

Though higher education is not governed by NCLB, the Act and related definitions of scientific standards and rigor adopted and endorsed by the Institute for Education Sciences (IES) (*Identifying and Implementing Educational Practices*, n.d; *Scientifically-Based Research*, n.d.; "What Works Clearinghouse," 2006; "WWC Study Design Classification," 2006; "WWW Study Review Standards," 2006) affect higher

educational researchers and practitioners. The issues and priorities of the NCLB are clearly echoed in the draft reports of the Spellings Commission on the Future of Higher Education (U.S. Department of Education 2006), demonstrating the relevance of the research and policy environment surrounding NCLB to higher education. As indicated by recent calls for higher education research proposals issued by the IES, the same federal standards will shape what is funded. Experimental and certain forms of quasi-experimental research designs are endorsed as those most necessary for understanding the effectiveness of educational programs. Some view the emphasis on rigor and objectivity expressed in these new federal standards as an embrace of scientific reasoning (Feuer et al., 2002; Shavelson, Phillips, Towne, & Feuer, 2003) while others decry it as a narrow form of "scientism" (Lather, 2004, p. 28) bereft of insights from the broader world of research, inquiry, and scholarship (Berliner, 2002; Chatterji, 2005; Erickson & Gutierrez, 2002; Lather, 2004; St. Pierre, 2002).

These critics argue the federal standards are uninformed by important methodological and epistemological debates and insights of the past two generations of researchers in education, evaluation, sociology, cultural studies, and other fields. They argue that the claims made to objectivity and scientific rigor in the federal research standards are greatly overstated. Not only because "Time and again, political passion has been the driving spirit behind a call for rational analysis" (Cronbach, cited in Chatterji, 2005, p. 18), but because there are distinct challenges to the generalization of statistical results to educational settings, which are highly variable and constantly changing (Chatterji, 2005; Erickson & Gutierrez, 2002; Raudenbush, 2005). The academic mode of research driven by theory and hypothesis-testing of causal effects, is not sufficient to answer what are essentially evaluation questions of "how, when, and why a program works" (Chatterji, p. 20). Greene (2000), arguing that the content and method of program evaluation are "inextricably intertwined with politics and values" (p. 983), places the current emphasis on examining the effectiveness of social programs as part of a historically dominant tradition, which in its contemporary form she terms "postpositivism."

The omission from the federal definitions of scientific research of other forms of educational research, including the evaluation genres Greene (2000) refers to as utilitarian pragmatism, interpretivism, and critical social science is viewed as particularly problematic because "educational change is accomplished locally" in "local situations of complexity and contingency" (Erickson & Gutierrez, 2002,

p. 23). What "works" in one educational setting may not necessarily work in another. Therefore, practices shown to be effective through causal experimental analysis cannot be replicated in practice in a simple manner. In addition, even if programs can be demonstrated "scientifically" to be effective, the "push" by policymakers for replication of educational practices will require a "pull" from educational practitioners for innovation to actually occur (Zaritsky, Kelly, Flowers, Rogers, & O'Neill, 2003, p. 33). This implies that the federal investment of $18.5 million in the What Works Clearinghouse (Lather, 2004, note 15), which is a "push" approach to dissemination of effective programs, will require additional investment to put scientifically validated knowledge into use by practitioners. Cohen, Raudenbush and Ball (2003), for example, emphasize that knowledge gained through educational experiments will require consideration by "communities of practice" and well informed "professional conversations" to be put into practice. These insights are the basis of action research and practitioner inquiry (Bensimon, Polkinghorne, Bauman, & Vallejo, 2004), which emphasize the educational practitioner's role as agents of change.

Feuer, Towne and Shavelson argue that "decision makers *at all levels* are clearly thirsting for rational and disciplined evidence provided by science" (2002, p. 4, italicized emphasis added). However, scholars in diverse fields including psychology, philosophy, organizational behavior, and cognitive science have shown that decision making in educational practice is informed by a much broader range of ways of knowing (Bensimon et al., 2004; Polkinghorne, 2004; Sergiovanni, 1992; Simon, 1997; Tharp, 1993; Tharp & Gallimore, 1988). The dominance of "technical rationality," which elevates the knowledge of experts over the knowledge of practitioners, stems from the view that academic science can determine how to "understand and refine practice" and then "transmit" that knowledge to practitioners (Polkinghorne, p. 170). Yet, changes in educational practice come about through changes in the knowledge, beliefs, attitudes, and behaviors of educational practitioners. Effecting change is fundamentally an issue of professional development (Dowd, 2005) and learning, which is known to take place through social interaction (Bauman, 2005; Bensimon, 2004, 2005; Bensimon et al., 2004; Rueda, 2006; Tharp, 1993; Tharp & Gallimore, 1988). Understanding the feasibility of replicating or scaling up programs shown to be effective through scientific causal analysis is an issue distinct from that of promoting the adoption of effective practice among educators. Understanding what

motivates the adoption of "best practices" also requires the study of professional practice in education.

Audiences for Evaluation Results Though it is a critical issue, the way in which scientific knowledge is effectively transformed into practitioner knowledge and put to use in practice is not adequately addressed in the federal standards for educational research. To understand why, in addition to recognizing the role of politics, it is useful to distinguish the primary information needs of policymakers, educators, and academic researchers. As noted, the question of "what works" is essentially a matter of program evaluation. Whereas academic research is traditionally characterized by theory-driven questions and answers (Chatterji, 2005; Cobb, Confrey, diSessa, Lehrer, & Schauble, 2003; Feuer et al., 2002) and institutional, or user-driven, research is characterized by institution-specific investigations to inform administrative decision making and planning (McEwan & McEwan, 2003; Walleri, 2003), evaluation research occupies a middle ground and draws on both academic and institution-specific perspectives to inform different audiences. The field of evaluation uses diverse methods to provide formative and summative evaluations to practitioners and policymakers, recognizing their different decision making needs (Chatterji, 2005; Greene, 2000). Whereas practitioners are primarily concerned about effectively and cost-efficiently using resources at their disposal to operationalize program goals and objectives through specific educational practices, policymakers are rightly concerned with determining the best investment of public dollars among any number of educational programs or policies. The scope and specificity of interest of the two groups differs.

As Raudenbush points out (2005), policymakers cannot intervene directly in classrooms and, therefore, attempt to influence teaching and learning primarily through accountability and governance requirements and by providing resources in the form of incentives. Therefore, the historically dominant tradition of evaluation, with its focus on assessing effectiveness and cost-efficiency and quantifying the magnitude of program effects, is well oriented towards the views and needs of policymakers. Greene (2000) distinguished this set of values from that of organizational decision makers, who, acting as utilitarian pragmatists, focus on program effectiveness and improvement in terms of what is working at an operational level, including whether clients like a program and are satisfied with it. These methods meet the information and decision-making needs of midlevel program managers and

on-site administrators and are often carried out by them through interviews, focus groups, and surveys of program participants.

The Dynamic Educational Production Process Theory-driven academic research informs evaluations of higher education effectiveness conducted in these various evaluation traditions. Synthesizing the extensive academic literature in the area of college student outcomes by higher education scholars such as Pace, Astin, Tinto, Bean, Cabrera, Nora, Pascarella, Terenzini, and many others, Volkwein (2003, pp. 184–185) summarizes four "major assertions" of the higher education literature that are useful in analyzing institutional effectiveness. These four major research-based assertions are centered on (1) organizational characteristics, including mission, size, expenditures, complexity and selectivity, (2) pre-college characteristics, including academic preparedness and goals; (3) student-institution fit, depending on (a) social and academic integration, (b) student involvement and effort, (c) financial circumstances, and (d) competing demands of family, work and community; (4) campus climate, including perceptions of prejudice, discrimination, racial harmony, and tolerance of diversity. These families of assertions serve as models for academic and institutional research by characterizing and focusing on particular factors and processes that influence student learning and educational attainments, such as grades, degrees, certificates, and subsequent job placements. Under accountability, models of this type help to guide data collection and analysis, potentially making the processes of institutional assessment more efficient by avoiding "the expensive trap…of the need to measure everything all the time" (Volkwein, p. 184).

Conceptualizing institutional effectiveness with a basic "I-P-O" model where resources of all types are treated as inputs (I) used in a collegiate production process (P) to produce outputs (O), we can note that the families of research assertions focus attention on different elements of the production process, and thereby frame consideration of the problems and potential solutions of higher education effectiveness in different ways (For a review of the concept of a production process in higher education, see Toutkoushian & Danielson, 2002). Student pre-college ability and goals and certain organizational characteristics such as institutional expenditures can be viewed as measures of inputs, while concepts of student-institution fit emphasize the interaction of students and institutional characteristics in the educational production process. In these models, students themselves are viewed as important inputs, not only based on their levels of academic preparedness but on their level of motivation, effort, and time investments in their studies

as well. Rather than focusing on students as inputs, studies of discrimination and diversity focus on the way resources are used and people behave "inside the box" in educational processes to create supportive or denigrating climates for racial-ethnic minority students.

The difficulties of observing the use of inputs in educational processes and characterizing an institution's level of efficiency in transforming those inputs to desired educational outcomes are well known, and in fact are the basis for increasing calls for higher education accountability. Cohen, Raudenbush, and Ball (2003, p. 122) argued that understanding the central policy question concerning how the provision of additional resources affects instructional effectiveness has been highly elusive due to the fact that resources are not "self acting." As they pointed out, "The value of resources is likely to depend on the ways they are used" (p. 138) by teachers in instruction, by students in learning, and by the interactions between teachers and learners, which are mediated by the instructional environment. They observe that "What reformers term instructional 'capacity' is not a fixed attribute of teachers, students, or materials, but a variable feature of interaction among them" (p. 125).

Cohen, Raudenbush, and Ball (2003), therefore, advocate greater use of experimental designs, arguing that non-experimental causal analysis is inadequate to observe the highly interactive relationships between resources, users, and outcomes. It is not possible to statistically control for institutional characteristics and features "inside the black box" of the educational production function and relate those inputs to student outcomes in ways that can authoritatively inform policy decisions about resource allocation to improve educational effectiveness. Their call for experimental methods follows mounting critiques and consensus that commonly used multivariate regression analysis techniques have failed to adequately take into account unobservable characteristics influencing students' choice of educational programs as well as their performance in those settings. Within the field of higher education, major bodies of literature concerning key policy variables such as financial aid are thought to suffer from an inability to statistically distinguish the effects of student characteristics from the effects of program characteristics on important outcomes such as graduation and persistence, leading to biased estimates of program and policy effects (Alon, 2005; DesJardins, Ahlburg, & McCall, 2006; Dowd, 2004; Dynarski, 2002a; Rouse, 1998; Titus, in press).

These shortcomings of commonly used educational research methods have provided support for the NCLB research standards

and initiatives like the What Works Clearinghouse (WWC), which was created to identify and disseminate educational "programs, products, practices, and policies" demonstrated to be effective solely through experimental and quasi-experimental studies (see w-w-c.org, "What Works Clearinghouse," 2006) that are not subject to these estimation biases. The Clearinghouse limits its review of educational studies employing experimental designs with randomized treatment and control group assignment or quasi-experimental designs using regression discontinuity and matching techniques that statistically create equivalence in the characteristics of the treatment and control groups. These techniques are viewed as best for isolating the effects of specific programs on student outcomes. Unlike Cohen, Raudenbush, and Ball (2003), who advocated for the integrated use of ethnographic and experimental field research, the federal standards give minimal attention to non-experimental methods.

Observing Educational Practice in Institutional Context A primary weakness of the experimental and quasi-experimental analyses of program effectiveness emphasized as rigorous and scientific under the NCLB and IES standards is the lack of direct observation of educational processes and social contexts, a task for which other forms of research and evaluation are better suited (Chatterji, 2005; Erickson & Gutierrez, 2002; Feuer et al., 2002; Raudenbush, 2005; Rossi, Lipsey, & Freeman, 2004). Feuer et al. (p. 8), though major proponents of experimental methods, concede the need to understand the role of contextual factors in causal processes: "When a problem is poorly understood and plausible hypotheses are scant—as in the case of many areas of education—qualitative methods…are necessary to describe complex phenomena, generate theoretical methods, and reframe questions." This point is argued more forcefully by qualitative researchers. For example, Erickson and Gutierrez (2002, p. 23) criticizing the federal research standards wrote, "The variety and changeability of the hierarchically embedded contexts of social life are such that simple, consistent associations between generic cause and generic effect of the sort tested in formal social experiments are not likely to occur."

Given the relative strengths and weaknesses of experimental and non-experimental research, many argue for the use of mixed-methods of program evaluation (Chatterji, 2005; Cobb et al., 2003; Cohen et al., 2003; Design-Based Research Collective, 2003; Raudenbush, 2005). Cohen, Raudenbush, and Ball (2003) recommended the development and testing of social science experiments, or "regimes" as

they have termed them, using a combination of ethnographic and experimental designs (see also Raudenbush, 2005). Chatterji (2005, p. 17), argued that the quality of evidence concerning educational program effectiveness will be seriously compromised unless what she called "extended term mixed-method" (ETMM) designs are used. She emphasized that, within ETMM designs, theory-driven research informs academic perspectives, while pragmatic program-level research informs practitioners about client needs and social context, quality of program implementation, feasibility of goals, costs, and resource use. While small- and large-scale experimental tests of program effectiveness are necessary to inform policymaking, she argued, pragmatic evaluation is equally essential to inform decision making by practitioners. Similarly, the Design-Based Research Collective, recommends mixed-method use of hypothesis testing, experimental "engineering" of instructional settings, ethnographic observation, and theory development (Design-Based Research Collective, 2003). In design experiments, which have been conducted in secondary schools much more so than in higher education where the approach is relatively unknown, specific instructional technologies are studied within a "learning ecology," a complex, interacting social system (Cobb et al., 2003, p. 9). Design experiments are intended to address the shortcomings of traditional social experiments by directly examining and seeking to explain why instructional designs work and how they should be modified when implemented in new settings. These methodological debates and developments demonstrate the ways in which academic researchers are attempting to systematically examine and understand the dynamic processes of education in social context.

III. ASSESSMENT: EVALUATION OF INSTITUTIONAL EFFECTIVENESS THROUGH BENCHMARKING

Growing Interest in Performance Benchmarking Assessment is essentially an institutionally-designed mechanism of accountability that aims to improve learning and teaching in a higher education institution. According to Ewell (2005), assessment is "a program of locally designed and operated evaluation research intended to determine the effects of a college or university on its students, centered on learning outcomes, and engaged in principally for the purpose of improving teaching and learning." (p. 105). Assessment takes many forms, which have been amply documented by Banta, Ewell, Kuh, Peterson and colleagues, and others (Banta, 2004; Ewell, 1991, 2002; Kuh, 2001; Maki, 2004; Peterson

& Einarson, 2001). In this section, we focus narrowly on a particular form of institutional assessment, namely benchmarking through peer comparisons, which has proliferated under accountability perspectives infused with business-minded approaches to governing higher education (Bender & Schuh, 2002; Burke et al., 2002).

The perspective that benchmarking institutional performance will motivate innovation is reflected in the Spellings Commission report, for example, which sets the collection of "comparative institutional performance" data as its first priority for accountability. The purpose is to enable students, policymakers and others to "weigh and rank" institutions and use the information as a "vital tool for accountability, policy-making, and consumer choice" (U.S. Department of Education, 2006). The Commission envisions a central role for accrediting agencies in this effort, charging them to expand accreditation standards to "allow comparisons among institutions regarding learning outcomes and other performance measures" and to collect and disseminate these data as a "priority over inputs or processes" (p. 24).

Similarly, the Educational Testing Service (ETS) recommends the national development of a comprehensive data base measuring student learning outcomes through the use of standardized tests of general education and discipline-specific knowledge and skills in order to measure institutional performance and improvements in performance over time. In a far-reaching proposal, ETS recommends annual collection of standardized test results at all higher education institutions in pre- and post-college attendance administrations to allow comparisons of value added "across institutions or groups of peer institutions" (Dwyer et al., 2006, p. 23). Like the Spellings Commission, ETS would charge the regional postsecondary accrediting agencies with the task of integrating these test results into their accreditation review of colleges and universities as indicators of institutional performance. These proposals show that at many levels, the collection of student outcome data in forms enabling peer performance benchmarking is receiving serious attention and emphasis as a primary strategy to improve institutional effectiveness. Furthermore, accrediting agencies, whose standards set the context for a college's locally designed evaluation research (assessment), are being called on to play a major role in this effort.

Just as the prevailing federal legislative interest in educational research focuses almost exclusively on quantitative data and statistical analysis, accountability standards place great emphasis and value on the "hard evidence" (Dwyer et al., 2006, p. 1) and "bottom line results"

(U.S. Department of Education, 2006, p. 14) of quantitative student outcome indicators. The confidence placed in quantitative indicators and the rejection of assessment processes that are now an important aspect of accreditation requirements reflects the dominance of what Polkinghorne (2004) refers to as "technical-rational" views. Reviewing the historical development of different conceptions of rationality, Polkinghorne observed that "modern Western culture is dominated by means-end rationalization and technology. Technical-rational decision-making is presented by the culture as the only effective way of determining which actions to take in order to solve practical problems in both the physical and human realms" (p. 45). It is important to consider what is lost and gained through the current emphasis on performance benchmarking and to carefully consider the mechanisms by which higher education benchmarking can indeed be expected to increase the educational effectiveness of colleges and universities as a result.

Benchmarking is essentially a process of comparison for purposes of assessment or innovation (Bender & Schuh, 2002). The objective typically is for an organization to understand its own activities, achievements, shortcomings, and environment through comparison with carefully selected "peers." The peer group may be selected based on similar objective characteristics, such as enrollment size, or by the use of perceived best practices that are to provide a model for improved performance (Hurley, 2002). Benchmarking takes several forms and a number of classification systems exist to differentiate them. Yarrow and Prabhu (cited in Doerfel & Ruben, 2002) define metric, diagnostic, and process benchmarking in a manner that is relevant to the higher education context. Doerfel & Ruben (2002, p. 6) explain that the metric form of benchmarking is simplest and takes place through the straightforward comparison of performance data. This approach focuses "only on superficial manifestations of business practices" (p. 6). Diagnostic benchmarking is like a "health check," intended to characterize an organization's performance status and identify practices needing improvement. The third approach, process benchmarking, is the most expensive and time consuming. It brings two or more organizations into an in-depth comparative examination of a specific core practice.

Public higher education accountability systems initially relied primarily on metric benchmarking (Barak & Kniker, 2002), which is also called performance benchmarking, but have become more nuanced over time in response to objections from higher education practitioners

and leaders (Erisman & Gao, 2006). Elements of diagnostic benchmarking are beginning to emerge as accountability systems mature and are revised in reaction to the limitations of a sole reliance on metric indicators. As discussed below, this evolution has taken place in three ways: through greater attention to the identification of appropriate peer comparison groups, through the use of academic theories of student-institution fit as explanatory frameworks for student outcomes, and the development of practitioner inquiry teams designed to investigate the problems associated with metric indicators of poor performance.

Effects and Limitations of Performance Benchmarking Examples of systems designed to facilitate the comparison of higher education performance metrics are readily available. To enable comparisons among institutions by consumers, practitioners, and researchers, the federal government created the Peer Analysis System (PAS) and College Opportunities On Line (COOL) databases. These data analysis tools include a wide range of information from the mandated higher education data collection contained in the Integrated Postsecondary Education Data System (IPEDS). Subsequent to the Student Right to Know and Campus Security Act (1990), all colleges have been required to report the graduation rates of their full-time first-time fall student cohorts and community colleges have been required to report transfer rates (Bailey, Jenkins, & Leinbach, 2005). At the state level, the most common mandated outcome indicators of institutional performance are graduation, retention, transfer, and job placement rates (Burke et al., 2002; Dougherty & Hong, 2005; Erisman & Gao, 2006). The National Community College Benchmarking Project (NCCBP) is an example of a voluntary peer comparison effort that collects metric, as well as other, indicators, and involved 150 institutions and three state systems in 2005. In addition, non-profit agencies such as the Education Trust and The Institute for College Access and Success have created web-based data bases of college graduation rates and indicators of economic diversity, respectively, to enable institutional comparisons (see http://www.collegeresults.org/ and http://www.ticas.org/economicdiversity).

Graduation rates, the most prominent indicator in recent accountability news, have been assailed by college officials as particularly inappropriate and unfair as measures of performance for community colleges, with their diverse missions, (Bailey, Calcagno, Jenkins, Leinbach, & Kienzl, 2006; Dellow & Romano, 2002; Dougherty & Hong, 2005; Erisman & Gao, 2006), but the concern applies in the same

way to four-year colleges enrolling part-time students and students with "swirling" multi-institutional enrollment patterns (Borden, 2004; Pusser & Turner, 2004).

Bailey et al. (2005, 2006) summarize objections frequently heard among community college administrators and faculty, for example. These include the fact that many students are not aspiring to earn a degree, that many barriers to degree completion are beyond the control of the college, and that colleges don't earn credit under the Student Right to Know (SRK) graduation rate calculations for the many students who start at one campus and finish at another. Indicators based on rates of student progress or degree attainment are also undermined by the difficulty of identifying an appropriate cohort for comparison (Dellow & Romano, 2002). The federal graduation rate is based only on the outcomes of full-time students, which excludes a significant proportion of college students at institutions of lesser selectivity. In addition, it is based on a three- or six-year period, 150% of the traditionally expected time at two-year and four-year colleges, respectively, despite the fact that many students take much longer to graduate and are therefore excluded from the statistic.

Other challenges associated with the collection of accountability data include the fact that students may accumulate credits without applying for a degree or certificate; job placement, tenure, or performance information may not be available from employers; and subsequent enrollment at another institution may occur in another state or long after the student's initial enrollment, making it difficult to track. For community colleges, in particular, the indicators do not capture the multiple missions of the colleges and the extent to which they are responsive to their communities. The fundamental shortcoming of commonly used performance indicators stems from the fact that they are measures of college outputs, for example degrees awarded, rather than of the ultimately desired outcomes, in this example student learning, and therefore divorce attention from the true goals of accountability (Toutkoushian & Danielson, 2002)

Yet, Bailey et al. (2006) demonstrate that the SRK graduation rates nevertheless have value for several reasons. First, the majority of students enrolled do aspire to a degree and, given the high correlation between low aspirations and low socio-economic status, colleges also have a responsibility for raising students' aspirations. Second, though, community colleges and other open access colleges face challenges in educating academically under-prepared students who are often juggling work and family obligations at the same time they are enrolled, the

data show that some colleges are more effective in graduating students with similar characteristics. Therefore, low performing colleges may have something to learn from their higher performing peers, even if all institutions of lesser selectivity or open access admissions face distinct challenges in educating under-prepared students. In addition, although the SRK institutional graduation rates underestimate longer-term, system-wide graduation rates, this bias appears to be minimal and consistent across institutional peers in terms of their overall performance.

Although measurement issues may be addressed through more sophisticated data collection and analysis strategies, it is also important to consider the mechanisms by which the collection and public dissemination of collegiate performance indicators can be expected to bring about productive change at low performing institutions. Dougherty and Hong (2005) identify four policy levers, with the first, reduced funding, associated with performance funding and budgeting and the others associated with performance reporting. The threat of reduced funding has the potential to be a powerful lever of change, as illustrated by a case study of performance funding in England's further education colleges. Under an extensive and far reaching reform, the budgeting and administration of the further education colleges, which are similar to community colleges in the U.S., were restructured such that ten percent of their total funding depended on student outcomes. Along with other policy tools, this incentive led to a ten percent increase over a five-year period in the rates of student completion of short-course diploma programs (Jaquette, 2006). Though the context of this case study differs in many ways from accountability in the U.S., particularly in regard to the high degree of centralization of authority at the national level, the case demonstrates the manner in which changes in government policy can function as key determinants of higher education performance. Richardson and Ramirez (2005) reached a similar conclusion by examining specific policy components of accountability in six states and rating each of the states against each component. Comparing the relationship between these ratings and performance indicators, they concluded that the governmental "rules in use" do affect institutional performance.

However, as noted above, performance funding and budgeting have declined in popularity and were never associated with significant proportions of institutional budgets in the U.S., with the exception perhaps of Florida (Dougherty & Hong, 2005). Policymakers appear to be hesitant to increase the complexity of accountability plans

with elaborate budgeting schemes, unable to settle on stable and fair performance-based resource allocation criteria, and unwilling to increase higher education budgets sufficiently to fund performance incentives. Therefore, the remaining three policy levers associated with performance reporting are those primarily in use in higher education today. In the absence of financial incentives, the collection and reporting of student outcome data can motivate more effective performance through three mechanisms: by increasing awareness of state priorities at the college level, enhancing knowledge on campus of student outcomes, and promoting status competition. Dougherty and Hong (2005) found that these changes in practitioner knowledge had a more significant impact on colleges than threats of or real changes in funding. This was in part due to the fact that state officials themselves were required to clearly articulate state-level priorities and goals in setting the accountability standards.

However, this new knowledge about and concern for student performance had intended and unintended consequences. Dougherty and Hong (2005) observed positive changes in educational practice in the areas of institutional planning, curriculum, advising, outreach, and student outcomes assessment. But they also found evidence that colleges had reduced their academic standards and adopted policies restricting admission to more prepared students in order to improve their graduation and program completion rates. Their study shows that accountability pressures can lead colleges towards improved data collection, a more nuanced understanding of institutional performance, and decision-making based on data analyses (see also Dougherty, 2002), but they can also create "perverse" incentives that undermine the goals of accountability (Dougherty & Hong, 2005, p. 8). Reviewing the effects of accountability programs, such as mandatory testing, on secondary schooling, Bowen, Kurzweil, and Tobin (2005, p. 240) also conclude that powerful government incentives can have "good and bad results." The negative effects can be particularly problematic, they emphasize, because "the worst subversions" are often found at schools educating the most disadvantaged students.

Individual and Organizational Learning about Institutional Effectiveness If changes in beliefs and behaviors come about through experiential knowledge, the skepticism of practitioners regarding the validity and fairness of the accountability metrics is problematic. Applying the "carrot and stick" metaphor, learning theory demonstrates that simply using performance indicators as a stick to prod educators is unlikely to be effective in achieving accountability goals. A study of schooling

under accountability involving extensive field research (Abelmann & Elmore, 1999) illustrates why this is the case. In their study, Abelmann and Elmore found that the external accountability system had very little influence on actual problem-solving by school teachers and administrators. They concluded that the personal attitudes, values and beliefs of these practitioners regarding the learning enterprise in school (i.e, what students can do, what teachers expect from each other, and how much student, family, community, and school influence student learning) were the key factors in determining the solutions for the problems targeted by accountability.

Adult learners and experienced practitioners possess established knowledge of their educational contexts and students and this knowledge must be challenged or disrupted in some way before new professional practices will be adopted. Polkinghorne (2004), drawing on Dewey's description of the learning that takes place through reflective problem solving and "experimental inquiry" in "indeterminate situations" stresses the importance of experiences where "one's actions fail to produce the desired result" (p. 121). Based on this insight, Polkinghorne outlined four steps of "intelligent inquiry" necessary for the development of professional expertise and increased efficiency in problem solving. Notably, the inquiry process starts with "an indeterminate situation" (p. 123)—the experience of uncertainty or errors in judgment. Once the practitioner realizes certain practices are ineffective, she or he will subsequently be motivated to identify the problem, determine a solution, and carry out the solution. As Bauman (2005) concluded based on extensive field research involving higher education practitioners engaged in an organizational learning initiative, "When organizational actors doubt what they have traditionally believed to be true, an opportunity for learning arises" (p. 27). In fact, she defined "high learning groups" as those "prepared to doubt and question their own knowledge and practices" (p. 29).

Similarly, Tharp and Gallimore (1988), synthesizing a large body of socio-cultural, psychological, and cognitive research, emphasize the importance of the "de-automatization," of "fossilized" knowledge—what might be referred to more crudely as the "unlearning" of established knowledge—to instigate openness to new learning. When learners, including professionals, realize their established competencies and knowledge are ineffective or incorrect, they become motivated to seek assistance from "more capable others" and to consciously assist themselves in gaining new competencies.

Among the means of assisting learning and the performance of new competencies are strategies in use in the accountability field. These include modeling (e.g. publicizing "best practices"), contingency management (legislative incentives and sanctions), explanation (journal articles, research briefs, and policy reports), and instruction (presentations at professional association meetings). An additional and critically important strategy is creating "activity settings" in which learners are called on to perform actions requiring new competencies. As Thomas, Clark and Gioia (1993) express it, citing Weick, "cognition often begins with action." Ideally, these activity settings are structured in ways that both draw on and challenge established competencies, placing learners in the "zone of proximal development" where the new competencies are in reach (Tharp, 1993; Tharp & Gallimore, 1988) and obtainable through a process of "situated" inquiry (Bauman, 2005; Pena, Bensimon, & Colyar, 2006). There are two important implications of these theories of professional practice and socio-cultural learning. First, performance indicators must have face validity, or what we might also think of as "experiential" validity, among practitioners before they can be expected to motivate changes in behavior substantive enough to improve student learning outcomes. Second, the learning necessary to improve institutional effectiveness begins with the "cultivation of doubt" (Bauman, 2005) among organizational actors about the effectiveness of their current practices.

Academic Resistance and Engagement in Accountability Many in academia are far from being swayed by the "rational" and "objective" meaning of performance metrics. Miller (1998), for example, who takes a Zen approach to teaching and believes that teaching and learning are deeply personal endeavors, argued that great teachers are those immersed completely in the compassion of teaching and the "intoxication" of their subject matter. Arguing "that's what counts," he notes that teachers will "die" under the outcomes assessments and learning objectives of assessment (p. 15). The spiritual nature of teaching and the importance of passion and compassion to the art of teaching have also been emphasized by Palmer (1998), for example.

In contrast, others in academia support the contention that faculty members have been negligent in ensuring successful student outcomes. In an in-depth, historical analysis, Lazerson (n.d.) criticized the professoriate for derailing the issue of student learning. In fact, he suggests that as academic disciplines become more prosperous, their intellectual engagement with student experiences decreases, a view also advanced by Zemsky and colleagues (Zemsky, Massy, & Oedel, 1993).

Bogue (1998), as well as others, advocated self-leadership and responsible partnership with other stakeholders to strive for effective reorganization. This range of responses demonstrates it is not inevitable that accountability will be met by resistance from the professoriate. However, as Hirschhorn observed (1990), social defenses will be mobilized when stakeholders feel threatened and accountability is not managed with care. The axiom "we account for what we choose and what we claim as our own" (Koestenbaum & Block, 2001 p. 10) is extremely relevant to examining educators' responses to accountability. It leads to the conclusion that practitioners must be integrally involved in designing, adopting, and applying the knowledge acquired through accountability plans in order for accountability to have the desired influence on institutional quality and efficiency.

Furthermore, authentic faculty and administrative participation is essential because accountability systems leave the job of identifying effective processes to the colleges themselves. Metric benchmarking on output indicators assumes an underlying optimal "production function" (*Higher education revenues and expenditures*, 1998), in which resources are used to produce the highest quality outputs possible. The benchmarking process is intended to spur colleges to search for and adopt optimal practices once they become aware of their poor performance relative to peers. There is an implicit assumption that these strategies are knowable and attainable.

Therefore, learning must take place "inside the black box" of higher education. The ease with which practices effective in one setting can be adopted into another is debatable, as shown by our review of methodological debates in the educational research and evaluation literature. This is one reason state-level accountability systems have been under constant revision, evolving over time to be more sensitive to differences in institutional contexts (Erisman & Gao, 2006). "Contextualized problem solving" is important to develop "local knowledge" of a problem (Pena et al., 2006, p. 50), without which a college may copy the practices of its peers in a manner inappropriate to their context, leading to negative kind of "institutional isomorphism" (DiMaggio & Powell, 1983). If colleges bypass problem-framing, they may fail to identify the root causes of the problem on their campus and lose out on opportunities to prioritize potential solutions. Bensimon and colleagues (Bauman, 2005; Bensimon, 2005; Bensimon et al., 2004; Pena et al., 2006), drawing on work by Argyris and Schon, have emphasized the need for "double-loop" rather than "single-loop" learning to bring about transformative changes in higher education:

> The difference between single- and double-loop learning is that the former encourages individuals to view a problem functionally and search for structural or programmatic solutions. In contrast, double-loop learning entails the ability to reflect on a problem from within, in relations to one's own values, beliefs, and practices. Simply put, the difference is that the single-loop learner locates the problem externally and seeks to change others. Conversely, the double-loop learner is more apt to start from the self and engage in reflection that brings about self-change in values, beliefs, and practices (Pena et al., 2006, p. 50).

If we accept the empirical findings that over and above policymakers' "rules in use" (Richardson & Ramirez, 2005) the personal attitudes, values and beliefs of educators are the key determinants of the solutions practitioners adopt to address the problems targeted by accountability (Abelmann & Elmore, 1999), these distinctions between double- and single-loop learning are critical and cast the entire enterprise of identifying universal "best practices" for higher education in doubt. Investments in the collection of vast amounts of assessment data will not serve to create comprehensive "national resources" (Dwyer et al., 2006) in the absence of equally large investments in professional development activities that engage higher education practitioners in actively identifying and learning about the causes of inadequate institutional performance. This point distinguishes arguments for an accountability "culture of inquiry" (Dowd, 2005) from those advancing a "culture of evidence" (Dwyer et al., 2006).

The Evolution of Higher Education Benchmarking Causal analysis is inevitably necessary if the primary accountability interest is on institutional improvement, but performance benchmarking in and of itself does not provide any data for causal analysis. Though it may be tempting to interpret the higher scores of one institution relative to another in a straightforward way as a sign of a higher quality, more productive college, such an approach is flawed. If an institution (X) scores lower than a peer institution (Y), the comparison has many potential meanings, including that (a) X is, in fact, a lower quality institution than Y, or that (b) X is actually more productive than Y, but its outcomes are also lower because it uses a lower level of inputs, (c) X is better than Y on other valuable outcomes that are not measured, or (d) X obtained a lower score than Y in any given year due to measurement error. In terms of causal validity, a linear I-P-O production function model is inadequate as an explanatory framework for improving college quality. The fundamental stumbling block is the

fact that students themselves are both inputs and outputs in educational processes (Boulding, 1996; H. Bowen, 1996; Toutkoushian & Danielson, 2002), making it difficult to estimate the college's effect on its primary output: students. This is recognized both by educators who object to performance criteria that fail to take entering student characteristics into account and by educational research methodologists (Cohen et al., 2003). Therefore, in addition to collecting outcome data, a diagnostic strategy for improvement must also be adopted.

Informed by such methodological debates and controversies, higher education accountability has evolved in two primary ways. The first entails renewed attention to the importance of identifying appropriate peer groups for benchmarking performance indicators, which is necessary to achieve campus "buy in" that the comparisons are legitimate. The second involves the use of different strategies to contextualize benchmarking results. These strategies add a "diagnostic" element to benchmarking by providing a theory or process for interpreting performance data. As discussed below, diagnostic benchmarking as implemented in higher education often links academic research to institutional research and assessment, using diagnostic frameworks or diagnostic processes based on academic theories.

Identifying Appropriate Peer Groups for Benchmarking Selecting peer groups for appropriate comparative analysis is an important first step in benchmarking and is in itself challenging. The peer selection process is a political one. Naturally, college leaders seek to position their institution well in relation to the peer group for subsequent performance reports and general publicity. Therefore, the selection process often combines objective data analysis and political wrangling. Administrators, institutional researchers, state system analysts, and external consultants are among those who may be involved in identifying a group of potential peers, with administrative leaders, in particular, keeping an eye on potentially negative funding and public relations implications.

Throughout this process, each institution remains aware of its own unique characteristics and those of the communities and students it serves. This is true even when an entire state system is identified as the peer group, because inevitably important differences in institutional characteristics are present. Further, even if peer institutions are similar in terms of structural and environmental characteristics at the time they are chosen, this equivalence may not hold throughout the life of an accountability plan. It is likely no peer group will ever be considered perfect by all interested parties for accountability purposes,

especially when subsequent student outcome indicators will mark some institutions as "underperforming." These aspects of peer comparison under accountability create incentives to aim low in the selection of peers, rather than including "aspirational" institutions as members of the comparison group (Hurley, 2002).

Grouping colleges by state and institutional characteristics such as size and degree of urbanization are natural starting points for peer comparisons under public accountability. However, concerns that these groupings fail to take into account differences in student characteristics or the economic vitality of the surrounding communities or region have led to the proposed use of statistical modeling to predict a college's expected performance on an outcome measure such as graduation rates, controlling for such institutional and external factors (Bailey et al., 2006). Colleges would then be held accountable for levels of student outcomes based on statistically predicted performance targets (Dougherty & Hong, 2005).

Statistical controls of this type represent an important development because they can take into account two important types of "inputs" determining a college's productivity: student quality and financial resources, which particularly in the case of community colleges may differ quite significantly due to variations in local resources and tax support. While results-based accountability has always emphasized outcomes over the provision of resources and politically shifted attention away from questions of equity in resource allocation (Alexander, 2000; Burke et al., 2002; Dowd, 2003; Leveille, 2005), recent reports are again emphasizing the importance of understanding differences in institutional inputs, including governmental funding, on college productivity (Dougherty & Hong, 2005; Dwyer et al., 2006; Erisman & Gao, 2006; Jaquette, 2006).

Contextualizing Performance Results While these developments begin to connect the inputs and outputs of higher education, other provisions of accountability plans begin to look inside the "black box" of the "production function" of higher education. In a simple way, this is reflected by the fact that most states now allow colleges to provide a narrative explanation of problems or inconsistencies suggested by their outcome indicators (Erisman & Gao, 2006; Wellman, 2002). This is arguably an essential step for reducing unintended negative effects of selected performance measures that create disincentives to pursuing other desired goals. Erisman and Gao (2006, p. 15) provide the example that a community college should not be dissuaded from providing a needed GED preparation program because it might

"consume resources but not produce outputs for several years." To some extent, such undesirable trade-offs can be avoided by including a broader range of goals, such as increased access or remedial education, among the adopted performance indicators, as suggested by Dougherty and Hong (2005), or by allowing colleges to select their own customized indicators consistent with a state's strategic planning and long-term goals, a trend observed by Burke et al. (2002). These options move accountability plans towards a diagnostic benchmarking approach, in which campuses report indicators they consider most consistent with their mission and priorities for improvement. They also represent a compromise between the external accountability values of standardized reporting of outcome indicators and the internal academic values of mission differentiation and local autonomy of campuses.

However, these developments also have limitations in that states and institutions themselves would then have to deal with a larger number of indicators, which might proliferate and diversify by institutional sector. Wellman (2002), who studied the accountability systems of five states intensively, described seven principles for accountability effectiveness, including the need for comparability, simplicity, and visibility. She recommended that states link indicators to goals established in statewide plans. This theme is echoed in the Institute for Higher Education Policy's accountability study, which found a "frequent disconnect" between indicators and goals, as well as inconsistent use of the data collected to inform policy decisions. These studies suggest that indicators can be contextualized not only through narrative explanations of performance indicators, but also by relating specific indicators to broader goals for higher education and allowing colleges to demonstrate how trade-offs in their results contribute to those goals.

Diagnostic Frameworks and Processes Through institutional assessments driven by or endorsed as part of accountability requirements, academic theories also provide the basis for contextualizing results on performance indicators. A number of assessment instruments commonly administered by colleges draw on theoretical models of student-institution fit. These models conceptualize the production process of higher education as a set of interactions between students and learning environments. For example, in Astin's theory of student involvement, where the "process" component of the "I-P-O" is replaced by "environment," the model is characterized as "Input-Environment-Output."

Assessment instruments grounded in the student-institution fit literature include the National Survey of Student Engagement (NSSE), for four-year colleges, and its counterpart the Community College Survey of Student Engagement (CCSSE); the College Student Experiences Questionnaire (CSEQ) and its two-year counterpart (CCSEQ), and the Cooperative Institutional Research Program (CIRP) survey. The growth under accountability initiatives in use of these assessment tools illustrates the demand for diagnostic benchmarking strategies designed to assist practitioners in identifying solutions to the problems suggested by outcome indicators (Barak & Kniker, 2002). Other popular instruments such as the Noel-Levitz Student Satisfaction Inventory and the ACT Student Opinion Survey are essentially surveys of consumer satisfaction. These can inform administrators' understanding of student satisfaction with institutional support services, such as parking, dining, and access to timely advising, but do not offer solutions at the heart of the pedagogical process.

One such assessment tool, the CCSSE and its College Student Report, provides an example of the application of the student-institution fit model to diagnostic benchmarking. The CCSSE has been adopted system-wide or at the majority of colleges in five states (Connecticut, Florida, Hawaii, Maryland, and New Mexico). Several consortia, including the Hispanic Serving Institutions/Hispanic Association of Colleges and Schools, have adopted the CCSSE as a shared assessment strategy (CCSSE, 2004) and the CCSSE as well as the NSSE were recommended for institutional adoption by the Educational Testing Service's in its discussion of a comprehensive national system for assessing student learning. Student responses to survey items are grouped into five areas of collegiate experiences—student effort, active and collaborative learning, academic challenge, student-faculty interaction, and support for learners—and standardized benchmark scores obtained for each participating college. The concepts of student "engagement" and "effort" provide explanatory frameworks, based in the academic research literature. These constructs are a lens for interpreting the benchmark scores in relation to a small peer group present in the data and examining institutional practices to determine how institutions can increase their effectiveness.

For example, the survey asks how often a student "Talked about career plans with an instructor or advisor." The student's response to this and other similar questions about behaviors in and out of class contributes to the college's score on a Student-Faculty Interaction scale (Marti, 2004). Remedies for colleges scoring low on this scale might

include revised faculty advising policies or assignments, with the goal of improving student retention. This benchmarking strategy offers an explanatory framework for "what matters" to student success in ways that offer guidance to administrators about effective practices. College results on the benchmarks are comparable nationally against other institutions participating in the survey, within state systems that adopt the CCSSE as a performance indicator, and over time at individual colleges (CCSSE, 2004).

Whereas assessment instruments based in the student-institution fit literature provide a diagnostic framework for interpretation of outcome indicators, another approach, taken by the Equity Scorecard projects structure a diagnostic process. With its focus on inequities in student outcomes by race and ethnicity, the interpretive lens is provided by the fourth "family" of academic perspectives described by Volkwein (2003). Practitioner "evidence" or "inquiry" teams (Bensimon, 2004, 2005; Bensimon et al., 2004) are instituted to "cultivate doubt" among team members about the equity and effectiveness of current practices (Bauman, 2005) through close examination of student outcome data, such as graduation and persistence rates, disaggregated by race and ethnicity. Institutional barriers to student attainment are conceptualized as stemming from socially constructed beliefs, held consciously or unconsciously by faculty and administrators, about lack of student ability, motivation, or aspiration, particularly among African American and Latino students.

The Equity Scorecard Project, and its predecessor the Diversity Scorecard, is modeled on the business world's Balanced Scorecard Framework (Kaplan & Norton, 1998) and has been implemented in over 35 four- and two-year colleges in six states. In California and Wisconsin, the use of the Equity Scorecard process by groups of colleges was endorsed by state system leaders to meet state accountability requirements and in Colorado it was implemented in collaboration with the Western Interstate Commission for Higher Education (WICHE). The benchmarking process is designed to create an "activity setting" for learning (Tharp, 1993; Tharp & Gallimore, 1988) and professional development among the faculty, administrators, and institutional researchers who constitute the inquiry teams as well as among their colleagues with whom they interact in professional settings.

The Scorecard process engages team members in a series of meetings that initially revolve around "vital signs" (indicators of student degree progression and student outcomes) but proceed to examination of "fine-grained measures" of the same types of indicators.

An important difference between the vital signs and the fine-grained measures is the fact that the teams themselves select the latter indicators and request that they be provided for discussion to the team by the institutional research office. The selection of "fine-grained" indicators often involves narrowing the focus to student progression between particular course sequences, for example from developmental to college-level coursework, or to particular groups of students, for example African American males, who may have progressed through the curriculum at lower rates. The process of defining the indicators and specifically requesting the data is intended to promote problem framing and ownership of the assessment results. Theoretically, the plausibility of this assumption is supported by sociocultural and organizational learning theories that indicate that learning is essentially a social process that takes place in "communities of practice" through engagement in collaborative and productive activities (Pena et al., 2006; Rueda, 2006; Tharp & Gallimore, 1988). The Equity Scorecard project provides "assisted performance" to the inquiry team members by structuring the inquiry process and providing instruction to team members in the dynamics of productive group processes and social learning (Bauman, 2005; Rueda, 2006).

These examples illustrate the diverse approaches to higher education benchmarking intended to assess and improve collegiate educational quality, productivity, and efficiency. They reflect different strategies for finding a model or explanation for "what matters" in achieving positive student outcomes as well as different assumptions about the best strategies for bringing about institutional change and improvements in educational effectiveness. That some type of diagnostic process is needed to achieve accountability goals is suggested by the uneven performance of various accountability indicators and plans to date, which empirical studies of accountability have shown to have effects that are moderately positive at best, inconsequential or uncertain at the mode, and negative at worst due to unintended consequences that undermine other valued higher education goals. Further, although the value of diagnostic benchmarking is supported through educational and learning theories, empirical studies of the effects of these types of assessments on organizational learning and change are also needed to demonstrate their capacity to improve student outcomes.

None of these benchmarking strategies is designed to measure the "productive efficiency" (DesJardins, 2002) of educational processes, which would entail understanding an organization's ability to convert inputs to outputs, or resources of many different types to student

outcomes in the form of learning and degree completion. Process benchmarking, which involves the in-depth comparison of core "production" processes for the purpose of improvement and innovation among two or more organizations (Doerfel & Ruben, 2002), has been largely absent from accountability plans. While the accountability movement has primarily asked institutions to report metric indicators of student outcomes, the assessment movement has focused on characterizing specific aspects of instructional environments and institutional practices that support learning. The design of the EBICs, discussed in the next section, is intended to be comprehensive in capitalizing on recent developments in the theories and methods of accountability and also to introduce a process benchmarking dimension. Process benchmarking is an important missing piece of what is needed to identify and disseminate effective educational practices in such a way that they will function as "best practices" not only under empirical study or in the eyes of individual practitioners who promote them based on their experiential knowledge, but at colleges of varying characteristics and institutional contexts where they might be adopted.

IV. EVIDENCE-BASED INQUIRY COUNCILS

In the previous section, we have argued that practitioner inquiry is an essential component of an accountability-driven research agenda. Furthermore, we contend that the theories and empirical results of the organizational and sociocultural learning literature demonstrate that much is known about the kinds of accountability structures needed to increase practitioner knowledge, motivation, and efficiency to meet accountability goals. To create a "pull" for change and innovation complementary to the "push" created by accountability policy (Zaritsky et al., 2003), educational practitioners should be involved in assessment activities that will cause them to question the effectiveness of their current practice and that will offer new information as the starting point for collaborative learning and increased productivity.

The design of the Evidence-Based Inquiry Councils (EBICs) is particularly attentive to the practitioner role within an integrated strategy of assessment and accountability and provides structures for user-driven research that will promote innovation. It treats the collection of student outcome data and the institution of practitioner inquiry teams as equally essential aspects of accountability. In addition, recognizing direct study of the efficient use of resources as a missing piece of accountability, the EBICs incorporate processes to audit the use

of resources within colleges and to compare those resource allocation decisions among peer institutions. The comprehensive research and evaluation agenda of the inquiry councils is intended to promote practitioner knowledge, organizational learning, instructional design innovation, and the integrated study of causal effects and mediating contexts.

As outlined in Table 2.1, the EBIC structure and activities are conceptualized as four inter-related phases, with different types of research and evaluation in the foreground at different points in the assessment and accountability cycle. Phase I involves determining a content focus for the council, such as developmental mathematics education, writing, critical thinking, engineering, or nursing, and convening member institutions. Phase II draws on multiple approaches including statistical data analysis, ethnographic observation, resource audits, and descriptive data analyses to reach a collective understanding of the nature of the problem to be addressed.

At Phase III the inquiry team adopts a "programmatic intervention," broadly defined as changes in the instructional materials, methods, settings, and interactions, as well as changes in the preparation and ongoing professional development of instructors, administrators, and student services personnel. The intervention, for example the adoption of computerized software and peer tutoring to teach developmental mathematics is implemented and the member colleges begin to test the effectiveness of the instructional changes through a design-based experiment (Cobb et al., 2003; Design-Based Research Collective, 2003) and, where warranted, experimental field research. Phase III is a period of formative evaluation, which also yields valuable information about how to implement the solution in a variety of contexts and how to achieve a high degree of uniformity in the "instructional technology" for testing in an experimental "regime," a highly structured instructional program conducted among peer colleges with variation in institutional contexts and resources (Cohen et al., 2003).

Phase IV provides a summative evaluation in three main forms. First, the results of the design-based research are presented as narratives providing rich descriptions of program design, implementation, and outcomes. These enable "naturalistic generalizations" (Stake, 1995) by practitioners to inform their decision to adopt the solution in other settings. The design-based research results also inform theories of teaching and learning that explain how, when, and why the intervention is expected to be effective. Second, quasi-experimental statistical analyses of the factors affecting student outcomes measured in

Table 2.1: Structure and Activities of Evidence-Based Inquiry Councils (EBICs)

Structure and Activities	Primary Mode of Inquiry and Research Methods
Phase I: EBIC Request for Proposals and Council Formation 1. Call for participation in one or more EBICs focused on a specific domain of educational practice (e.g. developmental mathematics or two- to four-year college transfer) issued by a state higher education coordinating body, a consortium of colleges, the IES, an accrediting agency, or a philanthropic foundation. 2. Colleges submit proposals to demonstrate preparedness for participation. Criteria for participation include data analysis capacity, commitment of personnel (including institutional researchers, faculty members, and administrative leaders), endorsement of the academic governing body and (where applicable) faculty and administrative unions, and allocation of physical and financial resources. 3. Academic and evaluation researchers submit proposals to serve as evaluators and facilitators.	*Mode of Inquiry* User-driven institutional research *Methods* Document review; Descriptive statistical data analysis.
Phase II: Problem-Framing 1. Analysis of course-level descriptive quantitative student enrollment, progression, and outcome data disaggregated by race/ethnicity, gender, and income. 2. Audit of current practice and resource allocation. 3. Self-study of institutional culture. 4. Cross-institutional metric benchmarking of inputs (e.g. financial resources, faculty qualifications, student characteristics) and student learning outcomes using available institutional and state-level data. 5. Review of extant literature characterizing effective practices in the educational domain of the EBIC, including studies of causal effectiveness in peer-reviewed journals, web-based clearinghouses, and archival sources. 6. Identification of a small number of promising practices as exemplary and review of the program logic (or theory) for consistency with institutional missions. 7. Advice gathered through expert testimony regarding evidence of effectiveness of the exemplary programs in the peer-reviewed academic literature. 8. Comparison and benchmarking of self-study findings among EBIC colleges.	*Primary Mode of Inquiry* Action research Practitioner inquiry Ethnographic case study *Methods* Document review; Descriptive statistical data analysis; Unobtrusive observation; Interviews

Table 2.1: (*Continued*)

Structure and Activities	Primary Mode of Inquiry and Research Methods
Phase III: Adoption of a Programmatic Intervention and Formative Evaluation 1. Council adoption and detailed description and of a programmatic intervention warranted by existing research and practitioner knowledge as an exemplary educational practice in the content domain of the EBIC. 2. Adoption of learning assessment instruments. 3. Field research by external evaluators examining the process and fidelity of program implementation. 4. Where warranted by consensus within the council of the potential effectiveness and fidelity of implementation of the instructional interventions, design and assignment of colleges to treatment and control groups for a small-scale randomized clinical trial. 5. Dissemination of findings and practices to other campuses through mobility of experienced faculty and administrators, presentations at professional association conferences, peer-reviewed journals, archival data bases, and EBIC web sites.	*Primary Mode* Extended-term mixed methods evaluation; Design-based research and social science experiments. *Methods* Design experiments; Randomized assignment field experiments; Ethnographic observation; Interviews; Document review; Narrative analysis.
Phase IV: Summative Evaluation 1. Statistical analysis of student learning outcome data to determine treatment effects. 2. Institutional self-study reports of program effectiveness and costs, including social context. 3. Cost-benefit analyses. 4. Dissemination of reports on EBIC member campuses. 5. Dissemination of findings and practices to other campuses through mobility of experienced faculty and administrators, presentations at professional association conferences, peer-reviewed journals, archival data bases, and EBIC web sites.	*Primary Mode* Inferential statistical analysis; Evaluation *Methods* Statistical analysis; Narrative analysis; Thematic and categorical analysis.

terms of course grades, tests scores on standardized assessments, and completion rates in courses and degree programs provide evidence of the relationship between student characteristics and outcomes in the programmatic intervention. When there is strong consensus in an EBIC

that the adopted intervention is expected to be effective and can be implemented with a high degree of uniformity and consistency, experimental research is conducted with volunteers from EBIC member or non-member colleges. A comparison of the outcomes of students in the experimental treatment and control groups provides evidence of a program's causal effectiveness.

Phase I: Call for Participation and Council Formation Under an integrated accountability, accreditation, and assessment system with the EBICs as a central feature, both private and public institutions would be required to participate in at least one EBIC per accreditation cycle by their accrediting associations. However, they would have the opportunity to choose among multiple possibilities which EBIC to participate in. For public institutions, EBIC participation would also meet state requirements for performance accountability. States and regions would have reciprocal recognition of each other's EBICs to enable the in-depth, sustained study of the most pressing problems facing higher education. When states require reporting of numerous outcome indicators, there is an implicit assumption that reporting on those indicators will motivate colleges to adopt solutions to improve performance across the board. This expectation is not well supported by empirical studies of the effects of accountability on institutional behavior, which is not surprising given that progress on longstanding problems is likely to require focused and collaborative effort. Therefore, states would coordinate the content focus of their EBICs with the assistance of the regional accrediting associations to enable a number of problems to be addressed simultaneously but without redundant effort in each state.

For example, if one state were to commission an EBIC to identify effective practices in developmental mathematics education, another an EBIC focused on writing-across-the-curriculum, another science education, and so on to address a wide range of issues including civic education, critical thinking, service learning, nursing, teacher education, etc., resources would be better directed both towards problem-solving and dissemination of evidence from the EBIC evaluation of program effectiveness. A call for participation in an EBIC would come from a state or regional higher education coordinating body, at the federal level from the IES (for example in critical national security areas such as foreign language study), from consortia of colleges and universities, or from a foundation seeking to improve institutional performance in a specific area of practice (for example, as in the Jack Kent Cooke Foundation's focus on increasing transfer from community

colleges to highly selective institutions (Wyner, 2006). In issuing the call for participation, the coordinating organization would name a steering committee and specify the content focus. The types of institutions eligible for participation would be specified, for example, all public or private, all four-year or two-year, or a cross-sample by sector and type. The number of participating colleges would vary, with a maximum number established at about twenty to allow for coordination and communication among the colleges. Conceivably, multiple EBICs would be established nationwide on the same topic to accommodate both the demand for participation and the need for an intensive problem-solving effort. The EBIC steering committee would seek to balance the participant group to enable peer and aspirational-peer comparisons among institutions with a range of existing practices.

In responding to the EBIC call for participation, each college would be asked to describe their current curriculum and assessment activities in the focal area, their data analysis capacity, and the titles and backgrounds of those who would be appointed as member of the college's inquiry team. All members must be willing to participate as collaborative learners, as the team is intended to support inquiry within a community of practice focused on understanding institutional effectiveness in the EBIC's content area. This team would coordinate activities on their campus both to gather input to the EBIC and to disseminate its findings. Each college would also submit a budget identifying their in-kind contributions of human and physical resources, as well as their resource needs, particularly in the area of data management and analysis. To ensure faculty participation, endorsement of the proposal to participate in the EBIC would also be required of the governing bodies of the college's faculty.

The EBIC will be effective only if it has the full support of the college leadership and it is allocated sufficient resources, including data, information, and time. The formal charge to the council to leverage resources for problem-solving in the EBIC content area should set high expectations for performance. Although there is no magic number for the size of the group, to facilitate effective participation at the college level, it should be relatively small and the members must have relevant and complementary expertise. Lipman-Blumen & Leavitt (1999) indicated that "hot groups" can range in size from three to thirty, but the optimal size depends on the degree of complication of the group task. The designation of advisory teams or sub-committees of the campus-level inquiry groups would be desirable, particularly for conducting specific tasks such as the cultural audits or cost analyses

of Phase II (discussed below). Goal clarity is essential for effective work groups, so the initial EBIC goals should be well specified by the steering committee and endorsed at the college-level at the proposal phase.

The entire multi-campus inquiry council, comprising all the campus-level inquiry team members, would be considerably larger than an effective work group and would function more as the EBIC's governance, communication, and decision-making body. At council-wide meetings, college-level teams would report the findings of their research concerning effective educational practices, benchmark their initial and ongoing educational processes, interpret outcomes of quasi-experimental and experimental tests of programmatic interventions, and share strategies for disseminating findings. In Phase III, where the council is asked to adopt and test a programmatic intervention warranted by existing research as effective or designed by the EBIC based on the extant literature, it might well be difficult to reach a full consensus among competing programmatic solutions. Therefore, the full inquiry council would also require a governance structure similar to an academic senate to vote on proposed solutions put forward by the teams. This governance process should stimulate debate about the range of options for selecting an intervention, because the debate itself, in promoting rigorous professional conversations (Cohen, Raudenbush, & Ball, 2003) and intelligent inquiry (Polkinghorne, 2004), may be as valuable in promoting organizational effectiveness as the identification of effective practices in the instructional regime. Both at the campus level and at the cross-campus EBIC level, the goals and tasks of the EBIC are intended to create structured "activity settings" (Tharp, 1993; Tharp & Gallimore, 1988) for organizational learning about institutional effectiveness.

The call for participation in an EBIC would also solicit proposals from academic researchers and professional evaluators to serve as facilitators and evaluators of the EBIC inquiry process and of the effectiveness of programmatic interventions adopted by the EBIC. Although studies of sociocultural learning have most typically involved subjects in experimental tasks rather than in everyday decision making, the insights of this field provide support for the integration of certain types of "assisted performance" (Tharp, 1993; Tharp & Gallimore, 1988) that would structure the inquiry councils as groups with a capacity for peer-assisted teaching and learning surpassing that of ordinary committees and task forces. EBIC evaluators external to the member colleges would serve as facilitators to the inquiry council to create task structures and

inquiry team member roles necessary for professional development and learning. Researchers with methodological expertise in ethnographic case study, design-based research, evaluation, and statistical analysis would also be needed to carry out the formative and summative evaluations of the EBIC's programmatic intervention as well as to disseminate generalizeable results of the study.

In addition to being in a better position to reach summative judgments of program effectiveness, external researchers and evaluators are needed at times to assist practitioners in the iterative process of identifying problems and evaluating solutions. Tharp and Gallimore (1988) describe several inhibitors of learning in professional settings: practitioners do not always see their own social (eco-cultural) context; supervisors and those with bureaucratic authority mistakenly focus on assessing rather than assisting performance; practitioners face real or perceived constraints on professional development and learning from authorities in their professional life; habits of interaction ("interaction scripts") are unconscious, deeply embedded in professional culture, and taken as a given; errors or weaknesses are not well tolerated as opportunities for learning in everyday professional life; and in-house training programs may simply perpetuate the existing culture and strengthen counter-productive entrenched knowledge. Based on these observations, the key principle for designing effective professional development programs is to ensure that effective assistance for learning and professional development occurs among peers, among authorities and those whose professional actions are regulated, and between external facilitators and participants in the activity setting. Some of these positive features already exist in the self-studies and external review team visits of accreditation, which can serve as a starting point for integrating assessment and accountability.

Phase II: Framing the Problem and Identifying Potential Solutions
Once the focus and goals of the EBIC have been established and the inquiry teams convened, Phase II involves the inquiry teams first in self-study and problem-framing at the campus level and then in cross-institutional benchmarking of resources, educational processes, and student outcomes among all the EBIC member colleges. The self-study involves descriptive data analyses, audits of resource use, and ethnographic observation. The use of a variety of audit instruments serves several purposes. First, they collect initial data for benchmarking change in practices through the course of the EBIC. Second, they help the inquiry team see and consider their own practices in new ways, in order to introduce the new ideas that stimulate learning.

And third, they structure the activity settings of the inquiry group, as cultural "artifacts" or mediators of interaction. (Design-Based Research Collective, 2003; Rueda, 2006).

In an EBIC focused on developmental mathematics education, for example, the college inquiry team would first examine student enrollment, performance, and outcome data in the mathematics curriculum. The relevant data, when available, include high school mathematics grades, standardized placement test scores, college course grades and completion rates, both in the developmental course sequence and in college-level courses in the curricular sequence, and the ultimate transfer (for two-year colleges) and graduation rates for students starting at the developmental course level. These data would be disaggregated by race and ethnicity and socio-economic status to observe inequities in participation and performance (Bensimon, 2004; Bensimon et al., 2004). The correlation among a variety of assessments of student learning (for example instructor-designed examinations, standardized placement or achievement tests, course grades, progression to higher level courses) would also be examined at this stage and disparities examined to determine if certain outcome indicators function better than others or can be improved through modification.

The resource audit process essentially involves creating an inventory of current personnel and physical materials allocated to instruction in the EBIC area, where "instruction" is broadly defined as inclusive of administration, teaching, and student support services. The costs associated with these resources would be estimated using an inventory tool such as the "ingredients method" (H. L. Levin & McEwan, 2001, 2002), which requires a systematic accounting of all the ingredients of instruction. Design-based researchers should be called on during Phase I when the EBIC is being convened to devise assessment instruments to facilitate the inquiry teams' work on this task. Through a case study of a small number of purposefully sampled institutions, the researchers could, for example, identify both typical and atypical, but potentially exemplary, ingredients of instruction in the EBIC content area.

A number of examples exist in higher education to provide models for the development of assessment instruments designed to benchmark educational processes. These include the Campus Compact's Indicators of Engagement Project (*IEOP*, n.d.), the Council for Adult and Experiential Learning's (CAEL) Adult Learning Focused Institution (ALFI) Assessment Toolkit (*ALFI Toolkit*, n.d.; *Serving Adult Learners*, 2000), and the Transfer Access Self-Assessment Inventory developed

in support of the Jack Kent Cooke Foundation's Community College Transfer Initiative (Dowd, Bensimon, & Gabbard, 2006; Gabbard et al., 2006).

The Campus Compact's indicators, which are derived from survey data, document exemplary practices for colleges whose mission and values include service learning and civic engagement. Based on a process benchmarking study of six institutions identified as highly focused on adult learners, CAEL's Assessment Toolkit includes a series of practices identified as particularly effective for serving adult learners. The Transfer Access Self-Assessment Inventory was developed through a literature review, document analysis, and case study of eight pairs of community colleges and highly selective colleges which appeared to have exemplary practices in the area of transfer. As recommended for identifying peer benchmarking groups and performance goals (Bailey et al., 2006; Dougherty & Hong, 2005), the initial case study sample of exemplary colleges was identified through statistical regression analysis comparing the predicted and actual number of transfer students (Dowd & Cheslock, 2006).

These standardized assessment tools provide examples of audit instruments needed to delineate core educational practices, which is an essential component of process benchmarking. These tools can be paired with cost reporting using the ingredients method of cost analysis to estimate the instructional costs per student expended at each EBIC college. In addition, they should be supplemented with self-study activities using the ethnographic methods of interviews, focus groups, and observation to enable the inquiry team to systematically study their own institutional culture. Examples of inquiry team activities conducted as part of the cultural audit include sitting in on classes, participating in students' study groups, observing patterns of use and interaction in the mathematics tutoring center, reading the course catalog to look at information about the curricular sequence from the student's point of view, and interviewing students either individually or in focus groups.

In this phase, EBIC member colleges would also exchange campus visits and observations to initiate the process of seeing their own campus from a new perspective. In addition, this evaluation would be informed by a review of the research literature concerning effective practices, factors affecting student outcomes, and variation in learning and outcomes by student characteristics. Peer-reviewed journals and archival data bases of "what works" will be useful at this stage to gain a comprehensive understanding of current educational theories and practice. This research review may be conducted by a college-level

subcommittee or advisory committee of the inquiry team who may be assisted by academic researchers, but it should not be conducted primarily or solely by external researchers. The inquiry team members must be knowledgeable decision makers regarding the design of the programmatic intervention to be adopted in Phase III and must also take ownership of the research and evaluation process to effectively implement the intervention and disseminate the findings of the EBIC throughout its life cycle.

The full inquiry council should convene at two points in Phase II, once to plan the processes and define the desired outcomes of the resource and cultural audits and once to compare the findings of their audits and research. The manner of benchmarking the inputs, processes, and "outputs," or student outcomes, of the colleges should be specified by the adoption of audit instruments and reporting formats. At either or both of these meetings, a panel of content and methodological experts should be convened to advise the EBIC members in the inquiry and evaluation process. Consistent with theories of sociocultural learning and the conceptualization of the inquiry teams as learning teams, the role of the expert advisory panel and of the external team of facilitators is to assist the performance of the inquiry teams by helping them to acquire new ideas and cultivate doubt in their current practice.

For the inquiry council to function as a learning group, the content experts and academic researchers serving as facilitators and evaluators should not impose an educational "solution" on the EBIC, but rather assist decision making by using all "seven means of assisted performance" (Tharp, 1993). As stated, these include explanation and instruction, common forms of communication by experts interacting with practitioners. However, other necessary modes of interaction include modeling processes of program evaluation, questioning decision makers about their rationale and evidence for adopting new programs or practices, and helping to segment decision processes to bring appropriate evidence to bear on decisionmaking.

The benchmarking process of Phase II should move the EBIC towards adoption among member colleges of an instructional "regime" in Phase III, where following Cohen, Raudenbush, and Ball (2003), a regime is defined as "systematic approaches to instruction in which the desired outcomes are specified and observed, and in which the intended outcomes are rationally related to consistent methods of producing those outcomes" (p. 133). It is important to note that instruction is construed broadly as a "collection of practices, including

pedagogy, learning, instructional design, and managing organizations" (p. 124). Therefore, the intervention is conceptualized not solely as the adoption of a specific practice or program, but also as the development of a learning system with its own complex ecology (Cobb et al., 2003; Design-Based Research Collective, 2003). Therefore, the underlying theories of program effectiveness also need to be articulated in order to ensure "authentic" (Chatterji, 2005) implementation of the intervention in a variety of sociocultural settings. In order to enable process benchmarking and improvements in Phase III, the resource and cultural audits of Phase II should, therefore, precisely document the existing "production" processes, resource use, and contexts of instruction among the member colleges. These activities provide the foundation for formative and summative evaluation of the educational effectiveness of the practices of the EBIC colleges.

Phase III: Adoption and Formative Evaluation of an Instructional Regime While Phase II is a period of "informed exploration," Phase III is an "enactment phase," (Bannan-Ritland, 2003) in which the EBIC adopts, refines, and tests an instructional regime. This may center on a new curriculum, a tutoring program, a learning community of integrated teaching and advising, or computer-aided instruction, but also requires attention to the social context of implementation, including the values and beliefs of practitioners at the colleges regarding student success. The adoption of an intervention entails a commitment from EBIC members to follow a specific instructional program with a high degree of uniformity and consistency and to allocate resources in particular ways (Cohen et al., 2003). The regime is broadly conceived in a way inclusive of its human, physical, and social resources. It is implemented as an intervention in this phase with an initial design that is subsequently modified through formative evaluation and theoretically informed "engineering" of the instructional program and learning ecology (Cobb et al., 2003; Design-Based Research Collective, 2003).

Developing Communities of Practice with Evaluative Capacity The benefits of this iterative and collaborative process of program definition are not only the development and testing of the specific set of practices that constitute the regime, but the development of communities of practice involved in evidence-based decisionmaking. EBIC activities in Phase III contribute to practitioner knowledge about program effectiveness in local settings, provide comparative estimates of cost-effectiveness across institutions, and aid in developing audit instruments for comparing resource use and institutional cultures. They are also intended to motivate practitioners to see their own practices and

sociocultural contexts in order to stimulate learning and a willingness to change. The simple fact of closely observing ineffective practices can create the "indeterminate situation" (Polkinghorne, 2004) that challenges established practitioner knowledge and spurs new learning.

The dual objectives of the EBIC are to generate evidence of effective educational practice to inform policymaking and to increase organizational capacity for evaluation in order to bring about regular use of multiple forms of evidence to inform practitioners' everyday decision-making. Cohen, Raudenbush, and Ball (2003) describe this approach as "developing professional knowledge and norms around a skeleton of objectives and tasks" in a community of practice. They describe the purpose of developing communities of practice organized around an instructional regime as fostering "professional conversations" as a form of rigorous evaluation (p. 138).

The audit processes of Phase II are intended to help EBIC member colleges observe their own institutional culture, particularly the ways in which decisions about resource use affect student learning. In Phase III, colleges are asked to adopt more uniform use of resources in order for systematic comparisons to be made across EBIC members about the effectiveness and efficiency of resource allocation decisions. Evaluation and research in Phase III is intertwined to determine both what works in the local context of the participating colleges and to develop theories of teaching and learning that explain how, when, and why the instructional program works in order to enable successful innovations at other colleges. The evaluation is focused on determining if the intervention works as designed and if it works in ways that make sense to the participants in the assessment process, which Bannan-Ritland refers to as its "ecological validity" (2003, p. 23). Undoubtedly, variations will be observed in the extent to which the practices selected for evaluation are effective. These variations are themselves instructive in clarifying why certain practices are effective in certain settings. "Lethal mutations" of innovations, which superficially share program characteristics but deviate upon implementation from the underlying principles of effectiveness (Zaritsky et al., 2003), can be instructive in emphasizing the how, when and why of "what works." In addition, the logic of program effectiveness becomes better specified through structured observations of program implementation in multiple settings.

Obtaining evidence of effectiveness and ineffectiveness through the problem-framing and formative evaluation stages of the EBIC, faculty members and administrators would begin the dissemination of the EBIC questions and findings through presentations by inquiry team

members in their departments, colleges, universities and professional association meetings. In addition, faculty members might very well inform their research, writing, and teaching in other settings through their experiences in the EBIC. The involvement of practitioners in user-driven research is more likely to spur the "word of mouth" dissemination that is critical to creating a demand for information about innovative practices (Zaritsky et al., 2003).

EBICs and Experimental Field Research The quality of the summative evaluation possible in Phase IV will be affected by the types of formative evaluations conducted in Phase III, including whether the EBIC conducts experimental field research. Although a well designed experiment is considered the "gold standard" for evaluating the causal effectiveness of educational programs, the majority of EBICs would not likely include a "true" experimental component, because it is administratively demanding and expensive to conduct experimental research with random assignment (Feuer et al., 2002). Nevertheless, valuable information about program effectiveness would be obtained to meet the dual objectives of generating results to demonstrate "what works" to both policymakers and practitioners.

Even in the absence of experimental field research, the formative evaluations of Phase III would be valuable to inform the experimental research agenda. Arguing that randomized clinical trials are a necessary but insufficient component of a research agenda focused on the effective use of educational resources, Raudenbush (2005) described the critical value of formative evaluation for eventual causal analysis of effectiveness. For example, the design stage identifies promising interventions and reduces the number of candidates for testing in experimental research. Formative assessments also precisely specify the instructional innovations worthy of testing and can do so in a manner targeted towards specific learners in specific settings. Given that large-scale experimental interventions are expensive, it is important to know if an intervention can be implemented as conceived on a smaller scale before proceeding to a large-scale randomized assignment intervention. The results of poorly conceptualized experiments are not only wasteful but may be misleading, Raudenbush observed, given that "Testing good ideas that are poorly implemented does not tell us 'what works'" (p. 29).

In contrast, a well constructed non-random or small-scale randomized experiment can show whether an innovative practice produces an effect in the expected direction. Such findings provide valuable information regarding instructional innovations even if the effect is not representative in broader populations. In addition, certain

innovative practices that worked in one setting may ultimately be ineffective due to resource constraints in other settings, so it is important to systematically observe resource use in an educational change process in multiple settings (Raudenbush, 2005).

The formative evaluation also assists in ensuring fidelity of program delivery, defining the specific nature of a "treatment" and identifying confounding or interaction variables that can mask effects in experimental field research, should large- or small-scale experiments be conducted as part of the EBIC or subsequently based on its results. The measurement of predictor and outcome variables for quasi-experimental and experimental studies can be specified and validated through formative evaluation (Chatterji, 2005; Cohen et al., 2003; Raudenbush, 2005). Through their discussions in Phase III, for example, EBIC members might adopt uniform ways of measuring student characteristics, such as race and ethnicity, socioeconomic status, and prior academic achievement, which would serve as control variables in analyses of student outcomes. Given that "innovative thinking often entails new goals for student learning" (Raudenbush, p. 29), alternatives for defining the dependent variable would also be explored at this stage. Assessment might take place through available standardized tests or instructor-designed assessments. The use of formative evaluation to specify key predictor, control, and dependent variables is an essential step for designing large-scale randomized field experiments (Chatterji, 2005).

An Example of an EBIC Instructional Regime An example helps illustrate the multiple research methods and tools that would inform understanding of "what works" under this proposed application of Cohen, Raudenbush's and Ball's "regimes" to the institution of EBICs as part of the accreditation and accountability system. As before, consider an EBIC investigating effective practice in developmental mathematics education. The members conclude through the problem framing, literature review, and expert panel discussions in Phase II that a particular curriculum involving computer-aided instruction (CAI) was a promising practice for improving student learning in developmental mathematics education. Therefore, the EBIC members decide to adopt the pedagogical theories, instructional software, textbook, instructor's manual, and standardized assessment tests of this curriculum at their colleges. The adoption of this program as an innovation is expected to be beneficial and, ultimately, scaleable as a cost-effective approach to increasing higher education productivity.

At the beginning of Phase III, the council faces the task of determining the physical, human, and social resources necessary to implement the new curriculum and specifying the program "theory" or "logic" of its effectiveness. The computerized component of the instructional program might be expected to increase student learning by allowing students to progress at their own pace and increase their time in class spent on solving mathematical problems. In addition, the EBIC members expect that with CAI instructors will spend more time in one-on-one interaction with students, which will help them individualize their instruction. The existing variation in resources among the EBIC colleges, such as in teacher experience and credentials, tutoring capacity, computers, and classroom space, would inform understanding of the resources necessary to implement the curriculum.

Different models of independent and collaborative learning, resulting from differences in available resources, could be explored. Through class observations, learning at one college with enough computers for each student might be compared to another college where students are grouped in pairs or trios due to a shortage of computers, with an opportunity to determine if individualized computer access is essential to the curriculum. The effects of differences in physical space use, such as the placement of computers in rows looking towards an instructional console at the front of the room or around the periphery of a group work space at the center could also be explored. Differences in student experience of the curriculum by characteristics such as age, enrollment intensity (full time or part time), and native language would be explored through interviews, focus groups, and surveys. The assumption of increased time on mathematics problem-solving might not be observed other than in classrooms with peer tutors, for example, who might have helped in a critical way to reduce the time students spent waiting for the instructor's assistance.

Through a combination of evaluative activities of this type involving cross-institutional observations, interviews, surveys of faculty and students in multiple developmental mathematics sections at multiple colleges, the EBIC members might then arrive at a preferred implementation plan involving the computerized instructional materials, one peer tutor, and one instructor per classroom with collaborative work groups of three students per computer. This configuration of resources could be benchmarked at member EBICs in terms of cost and feasibility using the cost analysis instruments from Phase II, which could also be refined based on observations of key "ingredients" that may have previously escaped notice.

Recognizing that the computers and educational software of the computer-aided intervention are not "self-acting" (Cohen et al., 2003), the inquiry teams would also investigate the extent to which the provision of resources in the regime motivates faculty effort, student effort, and facilitates or impedes interactions between them. Interviews with students, for example, may show that some feel embarrassed to ask instructors for help (perhaps with variation observed by the student's native language) while others have trouble navigating the software or seeing the screen (with variation by age). These results would lead to a revised implementation, in which the instructional program is "reengineered" to include a formal question and answer period and larger screens are made available in classrooms for older students. Native language and age are documented as interaction variables that affect students' experience of the curriculum.

The observed negative effects among some groups of students might be significant enough to reject further investigation of the CAI curriculum as a scaleable intervention, or the treatment could become better specified in terms of the required social resources (e.g. peer tutors) or physical resources (e.g. larger computer screens). A period of faculty and tutor orientation to provide instructional training in CAI might also be introduced as an essential aspect of the regime.

Through this process of specifying the CAI treatment, which should be long enough to include at least one feedback loop (Chatterji, 2005), the EBIC may have enough member colleges interested in conducting an experimental test of the effectiveness of the software and curriculum in comparison to traditional classroom approaches. If colleges did not feel it was ethical to assign students randomly, student characteristics in the CAI and traditional classrooms could be matched across colleges to obtain quasi-experimental statistical estimates of the impacts of the CAI curriculum. Colleges opting out of the experimental phase of implementation could participate as control classrooms or conduct observations of the fidelity of program implementation in the treatment classrooms.

Chatterji (2005) gives the example of a small scale field experiment involving one school, 16 classrooms, and approximately 250 students that provides a model for quasi-experimental in small numbers of EBIC colleges. In Chatterji's study, administrative constraints ruled out random assignment, but teachers in eight classrooms volunteered to participate in the field experiment and their classrooms were matched with eight others as a control group. Student outcomes were compared in matched-pairs by grade level and demographic characteristics. Prior

to the summative evaluation of the program effects, threats to the validity of the causal analysis from non-equivalent student characteristics on these moderating factors were evaluated and ruled out. Similarly, regression techniques analyzing a treatment group in one semester and a group of students from a previous semester can take advantage of the "discontinuity" of program characteristics amid the continuity of student characteristics to arrive at estimates of new program effects (see for example, Dynarski, 2002b; Linsenmeier, Rosen, & Rouse, 2001). As an alternative approach to obtaining rigorous estimates robust to self-selection and endogeneity bias, Titus (forthcoming) recommends the use of the propensity score matching technique to simulate the comparison of outcomes among groups of "treated" and "untreated" students.

Phase IV: Summative Evaluation Four sets of questions would be asked of the EBIC at the summative evaluation stage regarding the effectiveness, efficiency, and productivity of the tested educational practices and the effectiveness of the EBIC itself in instituting evidence-based decision making. First regarding effectiveness, what was the impact on student learning of the interventions implemented in the EBIC's instructional regime? What evidence can be presented to support the conclusions of the EBIC regarding program impact and what is the degree of confidence surrounding those conclusions? The highest degree of confidence concerning causal-effectiveness of educational programs would result from field experiments with randomized assignment of students to treatment and control classrooms or settings, in which case the program impact would be quantified by an effect size and a confidence interval. However, field experiments have been relatively rare in higher education and few EBICs would have the organizational and financial capacity to conduct them on a large scale. Evidence of causal-effectiveness would also result from well executed quasi-experimental analysis.

In addition, as argued by advocates of design-based research, rich narrative reports describing the iterative modifications and results of "engineered" experiments in the improvement of educational practice can provide evidence of the effectiveness of specific practices in a highly contextualized manner appropriate to informing practitioner knowledge about similar practices in other settings (Bannan-Ritland, 2003; Cobb et al., 2003; Design-Based Research Collective, 2003). Although this type of finding is not viewed as scientifically "warranted" knowledge by proponents of experimental field research (Shavelson et al., 2003), it does provide descriptive examinations of cause and effect that can

be judged by practitioners based on the authenticity and trustworthiness of the knowledge claims. In producing narrative reports, in particular, the external evaluation team would need to work closely with the practitioner-researchers in the EBIC to produce accurate reports of program effectiveness for the summative evaluation stage.

Second, regarding the efficiency of the instructional practices tested in the regime, the inquiry council would be asked to document the cost of implementing the instructional regime across the multiple institutional settings of the EBIC member colleges. What variations were observed in implementation and how did the variation in resources expended affect student outcomes? Do the results of the process benchmarking across the EBIC colleges provide evidence of optimal use of resources given more widespread adoption of the particular set of instructional practices tested in the regime?

Third, regarding the expanded productivity of higher education, is it feasible to "scale up" the instructional practices identified in the regime as effective across numerous colleges nationwide or were the circumstances of successful implementation restricted to colleges of a particular type serving a homogeneous student body? What would be the expected impact on student learning of adoption of practices recommended by the council, particularly on the number of additional students who would successfully complete an associate's or bachelor's degree?

Finally, regarding the effectiveness of the EBIC itself under an accountability agenda of increasing the number and quality of educated college graduates ready and able to contribute to national economic vitality, how did the EBIC contribute to the dissemination of innovative educational practices and the creation of a culture of inquiry on campus? To what extent did individual EBIC members adopt or test out new instructional practices and engage in self-reflective learning? Did practitioners engaged in EBIC activities develop a greater sense of self-efficacy in terms of their capacity to help students succeed?

It is also important to consider the potential cost-effectiveness of the EBICs as an accountability strategy. Although the meetings of the entire cross-institutional EBIC would incur direct costs for travel and materials, many campus-level activities could be conducted within existing academic and administrative structures, such as curriculum committees, accreditation self-study teams, faculty senates and subcommittees, and institutional research reports. The duration of an EBIC would vary, but the four phases are structured such that both the formative and summative evaluation results would be available within

the five- and ten-year periods of accreditation and mid-point review as evidence of institutional effectiveness in the focal area of the EBIC and of commitment to institutional assessment.

While ETS has proposed wide scale testing in higher education in general education and discipline-specific content areas (Dwyer et al., 2006), testing within EBIC instructional regimes may offer a more cost-effective strategy, by allowing focused evaluation of specific instructional practices considered by higher education faculty and administrators as those with the most promise to raise student achievement. As proposed by ETS, the universal collection of assessment data for all colleges may provide a means of continually tracking college student progression to a degree and thereby monitoring individual college performance. However, even after correlational analysis controlling for initial student characteristics, the significant investment in the assessment test data would not reveal information about the specific institutional practices and contexts that contributed to higher or lower performance among particular colleges. Even if educators were motivated to investigate causes of lower than expected performance at their college, they might not have sufficient knowledge outside a community of practice, such as that created by the EBICs, to interpret the causes of that poor performance. The integration of assessment and accountability reporting functions within the EBIC offers the potential to capitalize on existing expenditures on these activities and to increase utilization of the results.

IV. SUMMARY

To return to the statement by Feuer, Towne, and Shavelson (2002) cited earlier, we agree that decisionmakers at all levels are thirsting for rational knowledge to inform their decisions, if the concept of "rationality" is not too narrowly framed. There is a form of rationality that is critical to improved student learning that differs from the technical rationality underlying the concept of scientifically-based research or the "hard evidence" of test scores (Dwyer et al., 2006). It has been widely recognized by many scholars and is referred to as ordinary knowledge by Lindblom and Cohen (1979), practical rationality or deliberation by Lather (2004, citing work by Flyvbjerg), or reflective understanding by Polkinghorne (2004). Practitioners with this form of knowledge are responsive to the needs of individual clients (or students) and engage in problem-solving in context-sensitive ways.

Polkinghorne (2004) argues the critical necessity of drawing on practitioners' judgments to achieve excellence in the human sciences. He contrasts Plato's views of "techne" knowing, which aims to transcend the limitations of human experiential understanding through mathematical and calculative reasoning, with Aristotle's views of "phronetic reasoning," which "produces a perceptive understanding or insight about what is called for in a particular situation" (p. 106). He notes that "People's actions take place in situations of complexity and conflict. For an action to be appropriate to the occasion, it cannot simply be deduced from general knowledge or codified into a metric" (p. 107). Aristotle termed the "practical wisdom" that comes from this type of perceptive understanding "phronesis" (cited in Polkinghorne, p. 106).

Polkinghorne (2004) convincingly makes the case that our society must recapture Aristotle's sense of "phronesis" as an antidote to the dominance of "techne" knowing, which is inadequate on its own for problem-solving in the complex realms of practice in care-giving fields such as education and psychotherapy. Phronetic knowledge is responsive to the particularity of situations and the contexts in which they are embedded. The practical wisdom that is associated with phronetic reasoning enables practitioners to determine an appropriate course of action when dealing with contingent and changing situations and the uniqueness of the individuals involved in them.

Despite its uncertain impact, accountability continues to be politically prominent at the state and federal levels (Field, 2005a; Leveille, 2005; Zumeta, 2001). The fluctuations in accountability policy reflect the difficulty of mandating indicators of campus performance and of selecting appropriate measures of the complex ways in which colleges serve different students and communities. These limitations are present even when administrators and faculty accept or endorse the premise of public accountability. They are exacerbated when accountability plans take a punitive approach.

Numerous proposals have been made to modify and improve the design of accountability policies. Three emphasize developing a particular culture of higher education to promote educational effectiveness. The Spellings Commission relied heavily on the creation of data bases of collegiate performance indicators and consumer choice to establish what they described as a "culture of accountability." ETS proposed a "culture of evidence" in which colleges would extensively employ standardized assessments of student learning to benchmark their performance over time and against peer institutions. Our proposal

revolves around the concept of a "culture of inquiry," in which performance indicators and test scores are only two forms of evidence influencing educational change, which, we believe will come about only through changes in the beliefs, attitudes, and knowledge of practitioners. Quantitative data and outcome indicators are essential as part of a comprehensive accountability agenda that also emphasizes professional development and learning. Assessment processes are also an essential complement to results-based accountability. Much more empirical work needs to be done to understand the levers of effective accountability policy, including how a culture of inquiry can be developed to improve student learning. However, multiple strands of the theoretical social science literature support our argument that a focus on practitioner knowledge is critical to achieve the accountability goals of greater levels of higher education among the increasingly diverse population of young adult and adult learners in the United States. It is also supported by the uneven results of accountability policies to date in achieving their expressed goal of increasing educational effectiveness.

References

2005 Lumina Foundation annual report. (2006). Indianapolis, IN: Lumina Foundation for Education.

Abelmann, C., & Elmore, R. (1999). *When accountability knocks, will anyone answer?* Philadelphia: Consortium for Policy Research in Education, University of Pennsylvania.

Accountability for better results: A national imperative for higher education. (Report of the National Commission on Accountability in Higher Education)(2005). Report of the National Commission on Accountability in Higher Education). Denver, CO: State Higher Education Executive Officers.

Alexander, F. K. (2000). The changing face of accountability. *Journal of Higher Education, 71*(4), 411.

ALFI Assessment Toolkit. (n.d.). Retrieved May 14, 2005, from http://www.cael.org/alfi/isas.html

Alon, S. (2005). Model mis-specification in assessing the impact of financial aid on academic outcomes. *Research in Higher Education, 46*(1), 109–125.

Ayers, D. F. (2005). Neoliberal ideology in community college mission statements: A critical discourse analysis. *Review of Higher Education, 28*(4), 527–549.

Bailey, T., Calcagno, J. C., Jenkins, D., Leinbach, T., & Kienzl, G. (2006). Is student right to know all you should know? An analysis of community college graduation rates. *Research in Higher Education, 47*(5), 491–519.

Bailey, T., Jenkins, D., & Leinbach, T. (2005). *Graduation rates, student goals, and measuring community college effectiveness* (CCRC Brief No. 28). New York City: Community College Research Center, Teachers College, Columbia University.

Bannan-Ritland, B. (2003). The role of design in research: The integrative learning design framework. *Educational Researcher, 32*(1), 21–24.

Banta, T. W. (Ed.). (2004). *Community college assessment*. San Francisco: Jossey-Bass.

Barak, R. J., & Kniker, C. R. (2002). Benchmarking by state higher education boards. In B. E. Bender & J. H. Schuh (Eds.), *Using benchmarking to inform practices in higher education* (Summer ed., Vol. 118). San Francisco: Jossey Bass.

Bauman, G. L. (2005). Promoting organizational learning in higher education to achieve equity in educational outcomes In A. Kezar (Ed.), *Higher education as a learning organization: Promising concepts and approaches* (Vol. 131). San Francisco: Jossey Bass.

Bender, B. E., & Schuh, J. H. (Eds.). (2002). *Using benchmarking to inform practice in higher education* (Summer ed., Vol. 118). San Francisco: Jossey Bass.

Bensimon, E. M. (2004, January/February). The diversity scorecard: A learning approach to institutional change. *Change*, 45–52.

Bensimon, E. M. (2005). Closing the achievement gap in higher education: An organizational learning perspective. In A. Kezar (Ed.), *Higher education as a learning organization: Promising concepts and approaches* (Vol. 131). San Francisco: Jossey-Bass.

Bensimon, E. M., Polkinghorne, D. E., Bauman, G. L., & Vallejo, E. (2004). Doing research that makes a difference. *Journal of Higher Education, 75*(1), 104–126.

Berliner, D. C. (2002). Educational research: The hardest science of all. *Educational Researcher, 31*(8), 18–20.

Bogue, F. G. (1998). Changing leadership metaphors: From a culture of blame to a culture of responsibility. *Conference on Values in Higher Education*. Knoxville: University of Tennessee.

Bollag, B. (2004, July 16). Opening the door on accreditation. *Chronicle of Higher Education*, p. 22.

Borden, V. M. H. (2004, March/April). Accommodating student swirl: When traditional students are no longer the tradition. *Change*, 10–17.

Boulding, K. (1996). In praise of inefficiency. In D. W. Breneman, L. L. Leslie, & R. E. Anderson (Eds.), *Finance in higher education* (pp. 415–418). Boston, MA: Pearson Custom Publishing.

Bowen, H. (1996). What determines the costs of higher education? In D. W. Breneman, L. L. Leslie, & R. E. Anderson (Eds.), *Finance in higher education* (pp. 113–128). Boston, MA: Pearson Custom Publishing.

Bowen, W. G., Kurzweil, M. A., & Tobin, E. M. (2005). *Equity and excellence in American higher education*. Charlottesville: University of Virginia Press.

Burke, J. C. (2005a). The many faces of accountability. In J. C. Burke et al. (Eds.), *Achieving accountability in higher education* (pp. 1–24). San Francisco: Jossey-Bass.

Burke, J. C. (2005b). Reinventing accountability: From bureaucratic rules to performance results. In J. C. Burke et al. (Eds.), *Achieving accountability in higher education* (pp. 216–245). San Francisco: Jossey-Bass.

Burke, J. C. (2005c). The three corners of the accountability triangle: Serving all, submitting to none. In J. C. Burke et al. (Eds.), *Achieving accountability in higher education* (pp. 296–324). San Francisco: Jossey-Bass.

Burke, J. C. et al. (2002). *Funding public colleges and universities for performance: Popularity, problems, and prospects*. Albany: Rockefeller Institute Press.

Burke, J. C., & Minassians, H. (2003). *Performance reporting: "Real" accountability or accountability "lite"* (Seventh Annual Survey). Albany: Nelson A. Rockefeller Institute of Government, State University of New York.

Callan, P. M., & Finney, J. E. (2005). State-by-state report cards: Public purposes and accountability for a new century. In J. C. Burke et al. (Eds.), *Achieving accountability in higher education* (pp. 189–215). San Francisco: Jossey-Bass.

CCSSE (2004). *About CCSSE*. Retrieved September 14, 2004, from www.ccsse.org

Chatterji, M. (2005). Evidence on "what works": An argument for extended-term mixed method (ETMM) evaluation designs. *Educational Researcher, 34*(5), 14–24.

Clark, B. R. (1983). *The higher education system: Academic organization in cross-national perspective*. Berkeley: University of California Press.

Cobb, P., Confrey, J., diSessa, A., Lehrer, R., & Schauble, L. (2003). Design experiments in educational research. *Educational Researcher, 32*(1), 9–13.

Cohen, D. K., Raudenbush, S. W., & Ball, D. L. (2003). Resources, instruction, and research. *Educational Evaluation and Policy Analysis, 25*(2), 119–142.

Dellow, D. A., & Romano, R. M. (2002). Measuring outcomes: Is the first-time, full-time cohort appropriate for the community college? *Community College Review, 30*(2), 42–54.

Design-Based Research Collective. (2003). Design-based research: An emerging paradigm for educational inquiry. *Educational Researcher, 32*(1), 5–8.

DesJardins, S. L. (2002). Understanding and using efficiency and equity criteria in the study of higher education policy. In J. C. Smart & W. G. Tierney (Eds.), *Higher education: Handbook of theory and research* (pp. 173–219). New York: Agathon Press.

DesJardins, S. L., Ahlburg, D. A., & McCall, B. P. (2006). An integrated model of application, admission, enrollment, and financial aid. *Journal of Higher Education*, 77(3), 381–429.

Dill, D. D. (2003). Allowing the market to rule: the case of the United States. *Higher Education Quarterly, 57*(2), 136–157.

DiMaggio, P. J., & Powell, W. W. (1983). The iron cage revisited: Institutional isomorphism and collective rationality in organizational fields. *American Sociological Review, 48*, 147–160.

Doerfel, M. L., & Ruben, B. D. (2002). Developing more adaptive, innovative, and interactive organizations. In B. E. Bender & J. H. Schuh (Eds.), *Using benchmarking to inform practices in higher education* (Summer ed., Vol. 118). San Francisco: Jossey Bass.

Dougherty, K. J. (2002, April 1–5). *Performance accountability and community colleges: Forms, impacts, and problems*. Paper presented at the American Educational Research Association, New Orleans, LA.

Dougherty, K. J., & Hong, E. (2005). *State systems of performance accountability for community colleges: Impacts and lessons for policymakers* (Achieving the Dream Policy Brief). Boston, MA: Jobs for the Future.

Dowd, A. C. (2003, March). From access to outcome equity: Revitalizing the democratic mission of the community college. *Annals of the American Academy of Political and Social Science, 586*, 92–119.

Dowd, A. C. (2004, May 12). *Income and financial aid effects on persistence and degree attainment in public colleges*. Retrieved May 12, 2004, from http://epaa.asu.edu/epaa/v12n21/

Dowd, A. C. (2005). *Data don't drive: Building a practitioner-driven culture of inquiry to assess community college performance* (Research Report). Indianapolis, IN: Lumina Foundation for Education.

Dowd, A. C., Bensimon, E. M., & Gabbard, G. (2006). Transfer Access Self-Assessment Inventory [assessment instrument]. Los Angeles: Center for Urban Education, Rossier School of Education, University of Southern California.

Dowd, A. C., Bensimon, E. M., Gabbard, G., Singleton, S., Macias, E., Dee, J., et al. (2006). *Transfer access to elite colleges and universities in the United States: Threading the needle of the American dream*. Boston and Los Angeles: University of Massachusetts Boston and the University of Southern California.

Dowd, A. C., & Cheslock, J. J. (2006). *An estimate of the two-year transfer population at elite institutions and of the effects of institutional characteristics on transfer access*. Boston, MA and Tucson, AZ: University of Massachusetts Boston and University of Arizona.

Dwyer, C. A., Millett, C. M., & Payne, D. G. (2006). *A culture of evidence: Postsecondary assessment and learning outcomes*. Princeton, NJ: Educational Testing Service.

Dynarski, S. (2002a). The behavioral and distributional implications of aid for college. *American Economic Review, 92*(2), 279–285.

Dynarski, S. (2002b). *Loans, liquidity, and schooling decisions*. Boston, MA: Harvard University, Kennedy School of Government.

Erickson, F., & Gutierrez, K. (2002). Culture, rigor, and science in educational research. *Educational Researcher, 31*(8), 21–24.

Erisman, W., & Gao, L. (2006). *Making accountabiilty work: Community colleges and statewide higher education accountability systems*. Washington, DC: Institute for Higher Education Policy.

Ewell, P. T. (1991). To capture the ineffable: New forms of assessment in higher education. *Review of Research in Education, 17*, 75–125.

Ewell, P. T. (2002). An emerging scholarship: A brief history of assessment. In T. W. Banta (Ed.), *Building a scholarship of assessment*. San Francisco, CA: Jossey Bass.

Feuer, M. J., Towne, L., & Shavelson, R. J. (2002). Scientific culture and educational research. *Educational Researcher, 31*(8), 4–14.

Field, K. (2005a, October 14). Educators cast a wary eye at U.S. panel: Some fear federal intrusion into academe; others are pleased by the attention. *Chronicle of Higher Education*, p. A1.

Field, K. (2005b). *"Political rigidity" in academe undermines federal support for higher education, senator tells commission*. Retrieved December 13, 2005, from http://chronicle.com

Field, K. (2006). Draft report from federal panel takes aim at academe. *Chronicle of Higher Education*, p. A21.

Fleming, B. (2004, July 14). *House republicans defend proposals for holding colleges more accountable*. Retrieved July 14, 2004, from http://chronicle.com/

Gabbard, G., Singleton, S., Macias, E., Dee, J., Bensimon, E. M., Dowd, A. C., et al. (2006). *Practices supporting transfer of low-income community college transfer students to selective institutions: Case study findings*. Boston, MA and Los Angeles, CA: University of Massachusetts Boston and University of Southern California.

Greater expectations: A new vision for learning as a nation goes to college. (2002). Washington, D.C: Association of American Colleges and Universities.

Greene, J. C. (2000). Understanding social programs through evaluation. In N. K. Denzin & Y. S. Lincoln (Eds.), *Handbook of qualitative research* (2nd ed., pp. 981–999). Thousand Oaks: Sage Publications.

Gumport, P. J., & Pusser, B. (1995). A case of bureaucratic accretion: A case of context and consequences. *Journal of Higher Education, 66*(5), 493–520.

Higher education revenues and expenditures: A study of institutional costs. (1998). Arlington, VA: Research Associates of Washington.

Hirschhorn, L. (1990). *The workplace within: Psychodynamics of organizational life*. Cambridge, MA: MIT Press.

Hurley, R. G. (2002). Identification and assessment of community college peer institution selection systems. *Community College Review, 29*(4), 1–27.

Identifying and implementing educational practices supported by rigorous evidence: A user friendly guide. (n.d). Retrieved November 1, 2004, from http://www.ed.gov/print/rschstat/research/pubs/rigorousevid/guide/html

Indicators of Engagement Project. (n.d.). Retrieved September 14, 2004, from http://www.compact.org/indicators/indicators.html

Jaquette, O. (2006). *Making performance accountability work: English lessons for U.S. community colleges* (Achieving the Dream Policy Brief). Boston, MA: Jobs for the Future.

Kaplan, R. S., & Norton, D. P. (1998). The balanced scorecard: Measures that drive performance. In H. B. Review (Ed.), *On measuring corporate performance* (pp. 123–145). Boston: Harvard Business School Press.

Koestenbaum, P., & Block, P. (2001). *Freedom and accountability at work: Applying philosophic insight to the real world*. San Francisco: Jossey-Bass/Pfeiffer.

Kuh, G. D. (2001). Assessing what really matters to student learning. *Change, 33*, 10–29.

Kuh, G. D. (2005). Imagine asking the client: Using student and alumini surveys for accountability in higher education. In J. C. Burke et al. (Eds.), *Achieving accountability in higher education* (pp. 148–172). San Francisco: Jossey-Bass.

Lagemann, E. C. (2002). *Useable knowledge in education*. Retrieved September 13, 2005, from http://www.spencer.org/publications/

Lather, P. (2004). This IS your father's paradigm: Government intrusion and the case of qualitative research in education. *Qualitative Inquiry, 10*(1), 15–34.

Lazerson, M. (n.d.). *Discontent in the field of dreams: American higher education, 1945–1990*. Stanford, CA: National Center for Postsecondary Improvement, Stanford University.

Lederman, D. (2006). *Commission report, take 2*. Retrieved July 29, 2006, from http://insidehighered.com

Leveille, D. E. (2005). *An emerging view on accountability in American higher education* (Center for Studies in Higher Education Research and Occasional Paper Series No. CSHE.8.05). Berkeley: University of California, Berkeley.

Levin, H. L., & McEwan, P. J. (2001). *Cost-effectiveness analysis*. Thousand Oaks: SagePublications.

Levin, H. L., & McEwan, P. J. (2002). Cost-effectiveness and educational policy. In H. L. Levin & P. J. McEwan (Eds.), *Cost-effectiveness and educational policy* (pp. 1–17). Larchmont, NY: Eye on Education.

Levin, J. S. (2001). *Globalizing the community college: Strategies for change in the twenty-first century*. New York: Palgrave.

Lindblom, C. E., & Cohen, D. K. (1979). *Usable knowledge: Social science and social problem solving*. New Haven, CT: Yale University Press.

Linsenmeier, D. M., Rosen, H. S., & Rouse, C. E. (2001). *Financial aid packages and college enrollment decisions: An econometric case study* (Working Paper No. 459). Princeton, NJ: Princeton University, Industrial Relations Section.

Lipman-Blumen, J., & Leavitt, H. J. (1999). *Hot groups*. New York: Oxford University Press.

Maki, P. L. (2004). *Assessing for learning*. Herndon, VA: Stylus.

Marti, C. N. (2004). *Overview of the CCSSE instrument and psychometric properties*. Retrieved September 14, 2004, from http://www.ccsse.org/aboutsurvey/psychometrics.pdf

Massy, W. F. (2005). Academic audit for accountability and improvement. In J. C. Burke et al. (Eds.), *Achieving accountability in higher education* (pp. 173–197). San Francisco: Jossey-Bass.

McEwan, E. K., & McEwan, P. J. (Eds.). (2003). *Making sense of research: What's good, what's not, and how to tell the difference*. Thousand Oaks, CA: Corwin Press.

Miller, D. L. (1998). Nothing to teach! No way to teach it! Together with the obligation to teach! Dilemmas in the rhetoric of assessment and accountability, *Conference on values in higher education*. University of Tennessee, Knoxville, TN: University of Tennessee.

Moore, W. S. (2002). Accountability is more than "accounting": Promoting and sustaining institutional assessment-as-learning. *Journal of Applied Research in the Community College, 9*(2), 89–96.

Our students best work: A framework of accountability worthy of our mission. (2004). Washington, D.C: Association of American Colleges and Universities.

Palmer, P. (1998). *The courage to teach: Exploring the inner landscape of a teacher's life*. San Francisco: Jossey-Bass.

Pena, E. V., Bensimon, E. M., & Colyar, J. (2006). Contextual problem defining: Learning to think and act. *Liberal Education, 92*(2), 48–55.

Peterson, M. W., & Einarson, M. K. (2001). What are colleges doing about student assessment? Does it make a difference? *Journal of Higher Education, 72*(6), 629–669.

Polkinghorne, D. E. (2004). *Practice and the human sciences: The case for a judgment-based practice of care*. Albany: State University of New York Press.

Pusser, B., & Turner, J. K. (2004, March/April). Student mobility: Changing patterns challenging policymakers. *Change*, 36–43.

Raudenbush, S. W. (2005). Learning from attempts to improve schooling: The contribution of methodological diversity. *Educational Researcher*, 25–31.

Richardson, R. C., & Ramirez, A. (2005). *Policies and performance: The U.S. cross-case analysis*. New York The Alliance for International Higher Education Policies, New York University.

Richardson, R. C., & Smalling, T. R. (2005). Accountability and governance. In J. C. Burke et al. (Eds.), *Achieving accountability in higher education* (pp. 55–77). San Francisco: Jossey-Bass.

Rossi, P. H., Lipsey, M. W., & Freeman, H. E. (2004). *Evaluation: A systematic approach* (7th ed.). Thousand Oaks: Sage Publications.

Rouse, C. E. (1998). Do two-year colleges increase overall educational attainment? *Policy Analysis and Management, 17*(4), 595–620.

Rueda, R. (2006). *A sociocultural perspective on individual and institutional change: The equity for all project*. Los Angeles: Center for Urban Education, University of Southern California.

Scientifically-based research. (n.d.). Retrieved March 15, 2004, from http://www.ecs.org/html/educationIssues/Research/primer/researchsays.asp

Selingo, J. (2006). *To avoid federal mandate, colleges should devise own gauges of student learning, land-grant group says*. Retrieved April 7, 2006, from http://chronicle.com

Sergiovanni, T. J. (1992). *Moral leadership: Getting to the heart of school improvement*. San Francisco: Jossey-Bass Publishers.

Serving adult learners in higher education: Principles of effectiveness. (2000). Retrieved May 14, 2005, from http://www.cael.org/

Shavelson, R. J., Phillips, D. C., Towne, L., & Feuer, M. J. (2003). On the science of education design studies. *Educational Researcher, 32*(1), 25–28.

Simon, H. A. (1997). *Administrative behavior: A study of decision-making processes in administrative organization*. New York: The Free Press.

St. Pierre, E. A. (2002). "Science" rejects postmodernism. *Educational Researcher, 31*(8), 25–27.

Stake, R. E. (1995). *The art of case study research*. Thousand Oaks: Sage Publications.

Strout, E. (2004, July 13). *Colleges should become more accountable, Smith's leader says, or face the legislative consequences*. Retrieved July 13, 2004, from http://chronicle.com/daily/2004/07/20040713n.htm

Tharp, R. G. (1993). Institutional and social context of educational practice and reform. In E. A. Forman, N. Minick, & C. A. Stone (Eds.), *Contexts for learning: Sociocultural dynamics in children's development*. New York: Oxford University Press.

Tharp, R. G., & Gallimore, R. (1988). *Rousing minds to life: Teaching and learning in social context*. New York: Cambridge University Press.

Thomas, J. B., Clark, S. M., & Gioia, D. A. (1993). Strategic sensemaking and organizational performance: Linkages among scanning, interpretation, action, and outcomes. *Academy of Management Journal, 36*(2), 239–270.

Titus, M. A. (in Press). Detecting selection bias, using propensity score matching, and estimating treatment effects: An application to the private returns to a master's degree. *Research in Higher Education*.

Toutkoushian, R. K., & Danielson, C. (2002). Using performance indicators to evaluate decentralized budgeting systems and institutional performance. In D. Priest, W. Becker, D. Hossler, & E. P. St. John (Eds.), *Incentive-based budgeting systems in public universities*. Northampton, MA: Edward Elgar Publishing.

Trow, M. (1996). Trust, markets and accountability in higher education: A comparative perspective. *Higher Education Policy, 9*(4), 309–324.

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

Volkwein, J. F. (2003). Using and enhancing existing data to respond to campus challenges. In F. K. Stage & K. Manning (Eds.), *Research in the college context* (pp. 183–207). New York: Brunner-Routledge.

Volkwein, J. F., & Grunig, S. D. (2005). Resources and reputation in higher education: Double, double, toil and trouble. In J. C. Burke et al. (Eds.), *Achieving accountability in higher education* (pp. 246–274). San Francisco: Jossey-Bass.

Walleri, R. D. (2003). The role of institutional research in the comprehensive community college. *Journal of Applied Research in the Community College*, 11(1), 49–56.

Weiss, C. H. (1975). Evaluation research in the political context. In E. L. Struening & M. Guttentag (Eds.), *Handbook of Evaluation Research* (Vol. 1, pp. 13–26). Beverly Hills, CA: Sage Publications.

Wellman, J. V. (2002). Statewide higher education accountability: Issues, options, and strategies for success. In N. G. Association (Ed.), *Higher expectations: Second in a series of essays on the future of higher education* (pp. 7–15). Washington, DC

What we know about access and success in postsecondary education: Informing Lumina Foundation's strategic mission. (2006). Retrieved July 30, 2006, from www.luminafoundation.org/

What Works Clearinghouse: A trusted source of scientific evidence of what works in education. (2006). U.S. Department of Education Institute of Education Sciences.

What Works Clearinghouse: Study design classification. (2006). U.S. Department of Education Institute of Education Sciences.

What Works Clearinghouse: Study review standards. (2006). U.S. Department of Education Institute of Education Sciences.

Wolff, R. A. (2005). Accountability and accreditation: Can reforms match increasing demands? In J. C. Burke et al. (Eds.), *Achieving accountability in higher education* (pp. 78–103). San Francisco: Jossey-Bass.

Wyner, J. (2006, February 10). Educational equity and the transfer student. *Chronicle of Higher Education*, p. B6.

Zaritsky, R., Kelly, A. E., Flowers, W., Rogers, E., & O'Neill, P. (2003). Clinical design sciences: A view from sister design efforts. *Educational Researcher, 32*(1), 32–34.

Zemsky, R. M. (2005). The dog that doesn't bark: Why markets neither limit prices nor promote. In J. C. Burke et al. (Eds.), *Achieving accountability in higher education* (pp. 275–295). San Francisco: Jossey-Bass.

Zemsky, R. M., Massy, W. F., & Oedel, P. (1993, May/June). On reversing the ratchet. *Change, The Magazine of Higher Learning, 25*, 56–62.

Zumeta, W. (2001). Public policy and accountability in higher education: Lessons from the past and present for the new millenium. In D. E. Heller (Ed.), *The states and public higher education policy: Affordability, access, and accountability*. Baltimore: Johns Hopkins University Press.