CHAPTER 8.

THE POLITICS OF RESEARCH AND ASSESSMENT IN WRITING

Peggy O'Neill, Sandy Murphy, and Linda Adler-Kassner
Loyola University, Maryland, University of California, Davis, and University of California, Santa Barbara

With the rise of social science research and the professionalization of education during the late nineteenth century, educational research and practice have been tightly entwined (Bender, 1993; Labaree, 2007). Since that time, researchers studying education —especially K-12 education—have investigated a series of related questions: What should students learn, and why those things? Through what methods? To what extent are students learning what they should? How can learning be improved? A consistent definition of research has informed work undertaken to investigate these questions: It is a systematic gathering and analysis of information. In academic contexts, research is considered a discipline-defining activity; definitions of research are informed by specific fields of study. Members of academic disciplines determine appropriate questions to explore, employ appropriate methods for addressing those questions and interpreting results, and identify means for disseminating the information (Smart, Feldman & Ethington, 2000, pp. 6-7). Moreover, in academic disciplines, research is traditionally understood to be context and content neutral. But this positioning elides the reality that the act of research—the construction of research methods, the shaping of research results—is influenced by social and political factors that extend from the individual (what a person is inclined to see or not see) to the social and contextual (such as what research is funded, what type is valued, and what role it plays in policy decisions) (e.g., West 1989).

Educational research is particularly controversial because education is a complex, highly contested, politicized activity. This reality is evident in contemporary education in the United States. Increasingly, this research comes in the form of multiple assessments—of students' learning in particular subject areas; of teacher performance; of schools' achievement of particular goals. As Barbara Walvoord (2004) notes, assessment is "action research" intended to "inform local practice" (pp. 2-3) a "systematic collection of information about student learning" (pp. 2-3). At the K-12 and, increasingly, the postsecondary level, a

DOI: https://doi.org/10.37514/PER-B.2012.0452.2.08

number of stakeholders and interested others—testing companies, policy think tanks, classroom teachers, university researchers—are engaged in this kind of research, which is often linked to the day-to-day work of teaching: classroom activities, curricula, school structure and design. Student performances on tests or assessments are frequently used as the primary means to determine the success of change or the new programs aimed at creating change. The results of assessment research, then, have become a significant component of educational research and reform. In this chapter, we examine several studies in which what is included in and excluded from research has, or has the potential to have, considerable consequences for the students and teachers whose learning experiences will be affected by the activities being investigated. We also consider these efforts within the broader context of educational policy and the push for evidence of success.

WHY NOW?

While discussions about literacy crises are ubiquitous throughout the history of literacy in the United States (Graff, 1987, p. 16), they seem particularly consequential in the early twenty-first century because they are intertwined with considerable economic and political turmoil. Educational historian Diane Ravitch (2010) points to the 2001 passage of No Child Left Behind (NCLB), which provides funding for US public schools, as a primary culprit.[1] Under this policy, schools are required to demonstrate proof of annual yearly progress (AYP); ultimately, this demonstration is linked to the school's continued eligibility for particular kinds of federal funding. It is the responsibility of individual states to create (or adopt) measures and methods by which students demonstrate AYP. But the passage of NCLB has coincided with a dramatic reduction in federal and state funding for education; as a result, states have moved toward developing standardized assessments. These tests are administered to students yearly; students' "progress" is marked by the improvement of scores year to year. While individual students are expected to improve, so are schools' overall scores. The problems with these kinds of assessments are multiple (see, for example, Bracey 2006; Kohn 2000; Ravitch 2010); yet, because of the high stakes associated with them, they have come to drive instructional practices in many K-12 schools. In keeping with the traditional academic view of research as content and context neutral, proponents assumed that the assessment regime mandated by NLCB would produce context and content neutral results.

DEFINING THE "GOLD STANDARD": THE NATIONAL READING PANEL AND THE PRIVILEGING OF CERTAIN KINDS OF RESEARCH

Some of the tangled roots of NCLB extend from the National Reading Panel (NRP), whose work in the late 1990s revealed just how significant the impact of research definitions could be. Convened by the US Department of Education in 1997, the Panel was charged with "assess[ing] the status of research based knowledge about reading acquisition in young children" (NRP, 2000). In reviewing and evaluating the research to determine the most effective methods for teaching reading, the NRP only considered research that met their "gold standard"—that is, research using experimental or quasi-experimental designs. This decision "completely eliminated correlational and other observational research, two other branches of scientific study long accepted by the educational research community as valid and productive" (Yatvin, Weaver, & Garan, 1998, n.p.). The NRP definition of "gold standard" research had direct effects on education policy, which in turn affected classroom practice. It was used for the Reading Excellence Act of 1998 and the Reading First Initiative, both of which explicitly connected the results of research with teaching by providing funding for schools to implement curriculum shown to be effective by experimental and quasi-experimental research, but did not fund curriculum that had been proven effective through other research methods. Elements of the definition also found their way into NCLB. The definition of "gold standard" work extending from the NRP study continues to be used (almost exclusively) by the US Department of Education. The What Works Clearinghouse (WWC) and the Investing in Innovation (i3) fund, both programs sponsored by the Department of Education under the auspices of the Institute of Educational Sciences that provide funding for educational innovation, privilege experimental and quasi-experimental studies for determining program effectiveness (Investing in Innovation, 2010). These criteria are also used to assess research included in the WWC, a "central and trusted source of scientific evidence for what works in education" (United States Department of Education Institute of Education Science What Works Clearinghouse, 2010). According to the WWC evidence standards, "only well-designed and well-implemented randomized controlled trials (RCTs) are considered strong evidence, while quasi-experimental designs (QEDs) with equating may only meet standards with reservations; evidence standards for regression discontinuity and single-case designs are under development" (United States Department of Education Institute of Education Science, 2008, n.p.). Thus,

researchers looking for evidence of effective practice will find only "gold standard" studies in this Education Department site.

The ubiquity of experimental and quasi-experimental research in US Department of Education policy and practice might suggest that it has gone unchallenged since the late 1990s. But in fact, as soon as the NRP findings were published in 1998, the educational research community began to provide alternative definitions of "the best" and "appropriate" research that would enable inclusion of a greater range of research methodologies and evidence and thus allow for a wider range of educational practices extending from research conducted within those definitions. The National Research Council Committee on Scientific Principles for Education Research published a monograph suggesting that scientific research must pose significant questions that can be investigated empirically, link research to relevant theory, use methods that permit direct investigation of the question, provide a coherent and explicit chain of reasoning, replicate and generalize across studies, and disclose research to encourage professional scrutiny and critique (Shavelson & Towne, 2002, p. vii). The committee also supported the use of multiple types of research methods (Shavelson & Towne, 2002, p. 25). Other professional organizations also argued for multiple methods in response to the narrowly defined "gold standard" that made its way from the NRP to Reading First, and from NCLB to K-12 classrooms across the country. The American Evaluation Association noted that "[a]ctual practice and many published examples demonstrate that alternative and mixed methods are rigorous and scientific" (AEA, 2003). The American Educational Research Association also actively supported a more inclusive definition of scientifically-based research: "the term 'principles of scientific research' means the use of rigorous, systematic, and objective methodologies to obtain reliable and valid knowledge" (AERA, 2008).

Additionally, researchers examined the problems extending from narrow definitions of what research is appropriate that arise when research is used as the basis for policy decisions that, in turn, affect classroom teaching. The use of research to make such decisions is exceedingly complicated. Luke, Green and Kelly (2010) argue that teachers, students and schools do not function in neutral, universal, generalizable contexts (p. xiii). Furthermore, they contend that educational research cannot be transposed into policy that then becomes unexamined practice; instead, teachers must adapt policy research and policy so that they are appropriate for their specific classroom contexts.

Together, these researchers point to the issues associated with treating experimental and quasi-experimental research as the "gold standard." Limiting research to only experimental and quasi-experimental methods narrows the amount and kind of data that is collected, which in turn narrows the possi-

bilities for interpreting those data and creating a variety of teaching practices appropriate for different classrooms and learners. Far from functioning as a neutral definition of what research is appropriate, this standard has marginalized researchers and narrowed research-based perspectives. Additionally, it has extended beyond the boundaries of classroom or institutional study to profoundly affect educational policy in the United States. In other words, "gold standard" research may not provide the kind of data that would lead to information needed to make effective decisions about teaching and learning in real contexts.

BROADENING PERSPECTIVES THROUGH RESEARCH

While the "gold standard" holds sway at the federal level, many educators and researchers have attempted to assert that rigorous evaluation of an educational program requires more than test scores or other metrics related to experimental and quasi-experimental research (e.g., Davies, 2009; Luke, Green & Kelly, 2010; Wiseman, 2010). Additionally, researchers have gone on to make the case that including teachers and others who are involved in teaching and learning (such as administrators, students, and or parents) as partners in assessment research contributes to the development of robust tools and capacities to enhance students' learning.

Two recent national efforts involving writing scholars and teachers illustrate how much can be accomplished through alternative conceptualizations of research that enable the application of different questions and methods and allow for engagement by a broader range of participants. One is a multi-state, multi-year research project focused on K-12 writing instruction orchestrated by the National Writing Project; the other is a multi-state research project focused on first-year composition supported by the Fund for the Improvement of Postsecondary Education. To illustrate the potential of research conducted beyond the rigid confines of standardized measures and randomized control groups, we examine five elements of these efforts: the purpose of research; how research is defined; who was involved in the development of the research; what role was played by instructors as part of the research; and what kind of evaluation instrument emerged from or was linked to the effort.

The National Writing Project: Research and Engagement

The National Writing Project (NWP) is a network of professional development sites anchored at colleges and universities that serve teachers across disci-

plines and grade levels. The core principles at the foundation of NWP's national program model stress the centrality of writing for students, and the expertise and agency of teachers to act as researchers and "agents of reform" for writing education. Through its local sites, NWP teacher consultants provide professional development, create resources, conduct research, and act on knowledge to improve the teaching of writing and student learning. As part of its work to improve the teaching of writing, NWP has conducted research projects at local sites to "examine professional development, teacher practices, and student writing achievement" (NWP, 2010, p. 1). The broad purpose of the research has been to learn about the effectiveness of particular approaches to writing instruction in *specific settings*.

While these studies purposefully used experimental or quasi-experimental designs, the methods employed in each study depended on the local participants. However, all collected samples of student writing and employed pre- and post- measures to compare the performance of students whose teachers had participated in NWP programs to that of students whose teachers had not. The samples of student writing were independently scored at a national scoring conference using the Analytic Writing Continuum (NWP, 2006, 2008), an instrument developed and tested over a period of years by a group of writing assessment specialists and teachers of writing affiliated with the NWP.

All told, the sixteen research projects included in the studies ranged across seven states with an average contribution of 42 hours per teacher. One hundred forty-one schools, 409 teachers, and 5,208 students from large and small schools, urban and rural, with learners from diverse language backgrounds were involved (NWP, 2010, p. 4).

The local teachers and researchers who participated were not objective, neutral outsiders, but well-informed participants who understood the contexts for writing. However, this insider view was balanced by the national component of the research that brought participants from various sites together with writing assessment experts to score writing samples collected through the local research studies with a standardized rubric. The research design thus brought a number of voices involved and invested in education into the projects. One study in California, for instance, examined a program designed to improve students' academic writing that included sustained partnerships with teams of teachers from low-performing schools in both urban and rural areas. Another examined the impact of a program focused on the teaching of writing in grades 3 through 8 on teachers' classroom practice and on students' performance and attitudes. A study in Mississippi examined the impact of Writing Project partnerships on the achievement of ninth-graders in two high schools with predominately African American populations.

One important difference between these studies and others that have employed experimental or quasi-experimental methods was the ability, as part of the overall study design itself, to consider relationships between context and achievement. Another difference concerned the individuals and groups involved in the projects and the collaborative nature of the project itself. Instead of research conducted by disinterested outsiders—neutral researchers—these studies were developed and carried out by teachers and a range of others interested in the results of the studies and knowledgeable about the classrooms and the contexts: parents, other teachers, students, and school administrators. Teachers at the scoring conference were positioned as co-researchers in a form of action research, an approach where teachers and researchers work together and data are used for continuous, extended program improvement (Gilmore, Krantz & Ramirez, 1986; O'Brien, 2001). Findings from a study of the scoring conference showed that participant/scorers gained skills and knowledge about writing, and writing assessment, instruction, and development, and they took what they learned into their professional roles (Swain et al., 2010).

The collaborative, participatory nature of this research also led to assessment instruments that were employed across a variety of local sites to assess writing. Because the instrument had been developed by and with teachers, cultivating additional "buy-in," use of the instrument to develop yet more data that could be used to improve education, was not difficult. Shared use led to the development of shared language for the evaluation of writing among the participants in the studies. Equally important, it enabled assessment that was locally contextualized yet linked to common standards of performance shared across multiple sites. The results of the research, including the assessments of student writing and investigations of the effects on teachers of participating in the scoring sessions, indicated that both teaching and learning improved through local research initiatives and the scoring sessions (NWP, 2010; Swain et al., 2010).

NWP's work provides an example of experimental and quasi-experimental research that was sensitive to local context and included contributions from interested parties. The work was "based upon the premise that writing assessment and writing instruction exert an influence on one another" and that they are "situated within the larger contextual dynamic of district, school, classroom, and other professional policies and practices" (Swain et al., 2010, p. 5). Researchers associated with NWP claimed that "teachers thinking together with writing assessment experts helped to create a technically sound and rigorous assessment, one that is useful in the classroom as well as in research" (Swain & LeMahieu, in press, p. 22). An important assumption guiding this research and assessment project was that teachers bring an important perspective about what is happening in their classrooms, schools and districts to both research and as-

sessment. This approach, then, honored the local contexts while also meeting national standards.

POSTSECONDARY INTER-INSTITUTIONAL WRITING ASSESSMENT

While the NWP's work has largely focused on education prior to postsecondary study, American colleges and universities are beginning to face some of the same pressures for "accountability" that have led to the test-driven processes associated with No Child Left Behind (NCLB). This is a relatively new phenomenon, however, because the structures through which postsecondary education has developed in the United States vary from those surrounding K-12 education. The federal government has overseen K-12 education through a department (or part of a department) dedicated to education since the early twentieth century. Historically, there has been variation in the curriculum among schools, and local and state governments have had substantial influence. Addressing inequities perpetuated by some of this variation, in fact, is one of the motivations for legislative action such as NCLB.

An important difference between K-12 and postsecondary education in the US is that colleges and universities have intentionally differentiated themselves from one another, based on their missions. Particularly following the end of World War II, the United States has endorsed access to higher education for all citizens. As a result, a variety of different kinds of institutions have developed (two-year colleges focusing on vocational training and/or preparing students to transfer to four-year institutions; four year institutions of various types such as liberal arts colleges and technical institutes as well as comprehensive and research universities), each driven by its own individual mission (Bastedo & Gumport, 2003, p. 341). A second important difference is that as the American academy developed in the late nineteenth and early twentieth centuries, its professoriate relied heavily on peer review for everything from vetting research to determining standards. Thus, accreditation for postsecondary institutions, whose missions are specific to the institution, comes from private organizations (grounded in peer review), not the government.

Although the accreditation system has required postsecondary institutions to undergo program reviews and evaluations, until recently neither policymakers nor the public had questioned the autonomy or results of this system. However, in the last 10-15 years, calls for postsecondary educators to be "accountable" to public audiences and provide *comparable* data about their institutions have become ever-louder. As a result of the increasing emphasis on student

achievement (and in an attempt to ward off the kind of top down, legislated assessments associated with K-12), the higher education community has intensified efforts to document student. But because US colleges and universities tend to be independent, mission driven institutions, serving different populations in different ways, most assessment programs operate at the level of the institution with little history of networking or collaboration among institutions (with some notable exceptions linked to basic competency testing at the state level, such as programs legislated in Georgia, Florida and Texas). Thus, recent interest in accountability that draws in part on comparability across institutions, and sometimes missions, means that building networks and partnerships such as the NWP are in the nascent stages.[2]

The largest of these cross-institutional postsecondary assessment efforts is the Voluntary System of Accountability, a collaboration of two postsecondary organizations that has been adopted by "over 520 public institutions that enroll 7.5 million students and award 70 percent of bachelor's degrees in [the] US each year" (VSA).[3] While the VSA does not explicitly mention "gold standard" research, it draws on similar conceptualizations of research as earlier projects mentioned here, and does not engage faculty in the process of assessment of learning that is presumed to be occurring in their classes and programs.

Through the VSA, institutions create "College Portraits," online pages that purport to present unbiased, neutral information about colleges and universities for comparison purposes (VSA College Portrait, 2008, n.p.). While writing is not the exclusive focus of assessment used for these portraits, institutions participating in the VSA are required to administer (yearly) one of three standardized exams that are "designed to measure student learning gains in critical thinking (including analytic reasoning) and written communication." These tests are said to "measure these broad cognitive skills … at the institution level across all disciplines and are intended to be comparable across institution types" (VSA Background and Overview, 2008, n.p.). But this claim and the exams developed for it, like the claims underscoring the gold standard of experimental and quasi-experimental research extending from the NRP, reflect a particular perspective on the methods that should be used in research and assessment. Institutions participating in the VSA can choose from among three exams:

1. the Collegiate Assessment of Academic Proficiency (CAAP), developed by ACT, creators of one of two standardized exams taken by most American students who want to attend college or university;
2. the ETS Proficiency Profile, developed by ETS, creators of the SAT, the other standardized exam taken by most college-bound American students, as well as other tests taken by students wishing to enter postsecondary or graduate study; or

3. the Collegiate Learning Assessment, a product of the Council for Aid to Education.

The CAAP includes multiple choice questions intended to measure writing skills (broken down into "usage and mechanics" and "rhetorical skills") and a written portion that requires students to produce two, 20 minute responses to a prompt. The ETS Proficiency Profile includes multiple choice questions and an optional essay that is scored by eRater, a computer program that scores writing. The CLA asks students to produce written responses to case studies and has been scored with Pearson's Intelligent Essay Scorer since fall 2010, with some responses scored by human raters (Council for Aid to Education, n.d., p. 5).

The problem with these exams, as writing researcher Patricia Lynne (2004) has noted, and Chris Gallagher (2010) has reinforced, is that they do not assess writing in context, done for genuine audiences and purposes—three principles of effective assessment that have been reiterated time and again (e.g., CCCC 2009; NCTE-WPA 2008). Additionally, *institutions*—not faculty members—choose to participate in the VSA. The extent to which faculty are involved in any aspect of this decision depends on the institution; increasingly, writing researchers and instructors share stories about their exclusion from such decisions. This large effort to conduct cross-institutional assessment at the postsecondary level, then, reflects many of the issues associated with experimental and quasi-experimental research. It is a top-down mandate that does not engage participants; relies on artifacts created outside of the day-to-day contexts for student learning; and does not bring instructors into decisions about development, implementation, or interpretation of results.

A second approach to recent demands to create cross-institutional postsecondary assessments is the Valid Assessment of Learning in Undergraduate Education (VALUE) project from the Association of American Colleges and Universities (AACU).[4] While the VSA relies on standardized assessment results to generate information purported to attest to the development of students' abilities, institutions participating in the VALUE project use rubrics created by faculty from across different institutions and institutional types to assess portfolios of students' work from actual courses. The VALUE project also rejects the premises underscoring the "gold standard" of experimental and quasi-experimental research, stating that "that there are no standardized tests for many of the essential outcomes of an undergraduate education." Instead, it has "developed ways for students and institutions to collect convincing evidence of student learning" through the use of common rubrics ("Project Description"). The rubrics, according to the Project Outcomes, "reflect broadly shared criteria and performance levels for assessing student learning;" however, faculty are encouraged to "translate" the criteria "into the language of individual cam-

puses." As with the VSA, the VALUE project includes written communication as one of several competencies students should develop across the curriculum and throughout their education. However, it differs from the VSA in significant ways: it does not use standardized exams, it encourages institutions and their faculty to accommodate their individual contexts, it uses authentic class work, and it involves local faculty in the scoring. Yet, it still allows for cross-institutional comparisons.

While both the VSA and VALUE projects include writing, they are not focused on writing exclusively or on writing programs. Writing assessments more narrowly focused on writing programs have remained, for the most part, concentrated on local issues and curriculum. A notable exception is an interinstitutional assessment effort developed by writing faculty members at six different institutions of higher education, each with its own mission and institutional identity. This partnership reflects a unique response to requests for data about student learning at the college level (Pagano, Bernhardt, Reynolds, Williams, & McCurrie , 2008). Like the NWP's ongoing work, it is sensitive to concerns about assessment of student learning across institutions and within the context of public concerns; at the same time, it is driven by and dependent upon faculty's engagement with student learning and their own teaching and subject matter expertise. The collaboration also arose out of discussions about accountability in higher education, taking into consideration the rapid adoption across institutions of the VSA and the standardized exams it specifies (Pagano et al., 2008). But rather than rely on assessment perspectives reflected in experimental or quasi-experimental research and standardized tests, here a group of post-secondary writing faculty came together to create an alternative assessment to speak to demands to "assess individual change and establish effectiveness relative to national norms" (Pagano et. al., 2008, p. 287). The researchers sought to create a process for "jointly assessing authentic, classroom-produced samples of student writing … [and] create a public argument for the multiplicity of forces that shape an individual's writing and an institution's writing program" (Pagano et al., 2008, p. 287). Both this process and the assessment that resulted, then, were developed by and with the educators who would be affected by the assessment and, in turn, any effects resulting from it.

To undertake the investigation, each participating institution appointed a representative with expertise in composition studies to the project team. Team members worked together to develop the study and the mechanism used to evaluate data collected as a part of the research; at the same time, the "autonomy of individual programs" and "the goals of writing as taught within an institutional setting" were understood to be of primary importance (Pagano et al., 2008, pp. 290-291). This point highlights the productive tension between local

missions and purposes and the desire for cross-institutional comparison and consistency. Ultimately, each institution in the study decided to collect writing that involved students' "response to a text," a frequent requirement of academic writing (e.g., Greene & Orr, 2007, p. 138; Thaiss & Zawacki, 2006). But while the parameters of the prompt were shared ("response to a text"), what "respond to a text" meant for the specific campus was shaped by individual programs in the context of their institution. Team members met, scored project, and revised the rubric used for scoring; as a result of repeated scoring meetings, the team also created a more thorough set of descriptors for each criterion and increased the rating scale from five points to six (Pagano et al., 2008, pp. 295; 315-317). In this research project, then, the teacher-researchers used their expertise as both writing instructors and researchers to develop the rubric and use it.

Ultimately, the inter-institutional study resulted in information that each of the participating programs used to contribute to the development of student learning and enhance the "value added" in their institutions—certainly, a desired outcome of any assessment. Because researchers were engaged in the process of creating the design and conducting the study, they also were able to raise important questions about their process, as well as their results. This degree of reflection on the very process used for the assessment is only occasionally included by researchers engaged in experimental and quasi-experimental work.[5] Two elements of this inter-institutional study, then, provide important models for postsecondary writing research moving forward. First, like the NWP's writing assessment research discussed above, it attempted to address national concerns about learning development across a broad range of institutional concepts, by using *locally determined* questions and the means for addressing those questions. Second, it turned a lens back on itself, continually examining not just the subject of its study (writing development among college students), but the *methods used for that study.* That is, it worked from the presumption that these methods are not neutral, not unbiased, and not distinct from the very process of investigation itself.

Like the NWP research, the inter-institutional assessment demonstrates the extent to which quality writing instruction must be responsive to the institution where the instruction is taking place, and the benefits of assessment grounded in the actual work of classroom instruction for student and faculty development. It also highlights the complexity in collaborating across postsecondary institutions that have very different missions, students, instructional personnel, and curricula. Balancing the commitment to the individual context with the desire for comparability is difficult as demonstrated by Pagano and his research partners and by the critiques of research into student learning that relies on standardized exams and is conducted only by outsiders.

LESSONS LEARNED

Both the inter-institutional college writing assessment and the NWP assessment were developed and led by teachers to determine effectiveness of particular writing programs and practices. Both involved low-stakes writing assessments. Both relied on voluntary participation and collaboration across institutions and states. Both honored local conditions, expertise, and curricula; they were responsive as the situations demanded. Both produced research results that were useful for the specific teachers and writing programs involved as well as for determining effectiveness, including cross-institutional comparative information. Yet, neither conformed strictly to the "gold standard" definition of research. In fact, participants in both initiatives identified engagement in the research projects—not just the results produced—as a key benefit. Thus, the projects included more than an assessment of student work. They encompassed professional and curricular development with teachers positioned as co-researchers and professionals with requisite knowledge and expertise, not as technicians delivering a program and curriculum.

These studies also illustrate challenges facing writing researchers who aim to develop evidenced-based research studies exploring program effectiveness. In this kind of research, tests should be just one piece of evidence used to determine program effectiveness, teacher quality or comparability. Unfortunately, in the current research and assessment climate, student test results are considered the primary—or only—evidence of success. Researchers need to use multiple methods, as professional disciplinary organizations and scholars advocate, if we are really concerned with promoting learning and teaching.

The two research projects we highlight here also demonstrate the complexity of developing evaluation systems that balance local context with the need for some degree of standardization. Because both the NWP and the inter-institutional projects relied on voluntary participation, translating the approach to a top-down, mandated evaluation system may be difficult. These projects also demonstrate the wealth of resources needed—especially in terms of teacher time—to carry out the projects. However, the needs of policymakers for cost effective assessment information must not outweigh the potential benefits to the educational system as a whole. Although the assessments and research studies described here may be time consuming and costly, they offer important benefits. Healthy educational systems create situations in which teachers profit from their experience with research and assessment development, and which promote the professionalization of teaching. Further, they accommodate diversity in the programs that teachers offer and in the ways that students, local districts and states can demonstrate accomplishment. The rubrics developed by

NWP and the inter-institutional group enable diversity at the local level, but comparable standards across diverse sites.

As these projects also illustrate, discussions about what research counts, how research will be used, and how program effectiveness is determined are not academic, abstract, or carried out only in scholarly journals and conferences. Everyday, K-12 researchers, teachers, and students experience the repercussions extending from the privileging of experimental research as evidenced through the NRP and ensuing policies. As educational reform continues to be championed through federal programs such as Race to the Top and the "voluntary" Common Core State Standards Initiative that it endorses, policymakers could look to research like that conducted through the NWP, VALUE and inter-institutional assessments to learn more about how to use research and assessment in ways that position teachers as professionals who take responsibility for student learning and who care about what students are learning and to what degree. Approaching research and assessment in this way recognizes teachers' expertise and promotes research and assessment as means of professional development. It values the knowledge and experience that teachers have, yet it still enforces research standards and allows comparability, providing information that helps educators and the public understand how students are performing. As US policymakers push to make the transition from K-12 and college education more seamless for students, encouraging research and assessment that goes beyond experimental and quasi-experimental methods will provide a richer and more complete understanding of both teaching and learning. Relying on a narrowly defined, top down approach will misrepresent not only what students know and can do but also what it means to write and to teach writing.

NOTES

1. NCLB was the name given to the reauthorization of the Elementary and Secondary Authorization Act of 2001. The original Elementary and Secondary Authorization Act was passed in 1965 during the administration of President Lyndon B. Johnson. (http://www.aect.org/about/history/esea.htm)

2. A significant exception is the collaboration among prestigious Northeastern colleges during the early to mid twentieth century that resulted in the College Board and the SAT (see Traschel, 1992; Lemann, 1999).

3. The Association of Public and Land Grant Universities and the American Association of State Colleges and Universities

4. In fact, the VALUE project and the VSA were created simultaneously as part of a grant shared by AAC&U and AASCU, and APLU to develop two different pilot frameworks for assessing student learning across institutions.

5. Education researchers have, of course, voiced multiple concerns about the methodologies associated with experimental and quasi-experimental work—however, these are published separately from the studies themselves (e.g., Lather, 2004; Altwerger, 2005.)

REFERENCES

Adler-Kassner, L., & O'Neill, P. (2010). *Reframing writing assessment to improve teaching and learning. L*ogan, UT: Utah State University Press.

Altwerger, B. (2005). *Reading for profit: How the bottom line leaves kids behind.* Portsmouth, NH: Heinemann.

American Educational Research Association. (2008). *Definition of scientifically based research.* Retrieved from http://www.aera.net/Default.aspx?id=6790

American Evaluation Association. (2003). *Response to U. S. Department of Education notice of proposed priority, Federal Register RIN 1890-ZA00, November 4, 2003 "Scientifically Based Evaluation Methods."* Retrieved from http://www.eval.org/doestatement.htm

Association of American Colleges and Universities. (n.d.). *VALUE: Valid assessment of learning in undergraduate education.* Retrieved from http://www.aacu.org/value/index.cfm

Bastedo, M., & P. Gumport. (2003.) "Access to what? Mission differentiation and academic stratification in US public higher education. *Higher Education 46*, 341-359.

Bender, T. (1993.) *Intellect and public life.* Baltimore: Johns Hopkins University Press.

Bracey, G. (2006). *Reading educational research: How to avoid getting statistically snookered.* Portsmouth, NH: Heinemann.

Conference on College Composition and Communication. (2009). *Writing assessment: A position statement.* Retrieved from http://www.ncte.org/cccc/resources/positions/writingassessment123784.htm

Council for Aid to Education. (n.d.). *Architecture of the CLA tasks.* Retrieved from http://www.collegiatelearningassessment.org/files/Architecture_of_the_CLA_Tasks.pdf

Davies, R. (2009, November). Evaluating what works in education: Causation or context. Paper presented at the American Evaluation Association's annual conference, Orlando, FL.

Educational Testing Service. (2010). *ETS proficiency profile: User's guide.* Retrieved from http://www.ets.org/s/proficiencyprofile/pdf/Users_Guide.pdf

Gallagher, C. W. (2010). Opinion: At the precipice of speech: English studies, science, and policy (ir)relevancy. *College English 73,* 73-90.

Gilmore, T., Krantz, J., & Ramirez, R. (1986). Action based modes of inquiry and the host-researcher relationship. *Consultation, 5*(3), 160-176.

Graff, H. (1987). *Labyrinths of literacy: Reflections on literacy past and present.* Sussex: Falmer Press.

Greene, S., & Orr, A. J. (2007). First-year college students writing across the disciplines. In P. O'Neill (Ed.), *Blurring boundaries: Developing writers, researchers, and teachers* (pp. 123-156). Cresskill, NJ: Hampton,.

Hillocks, G., Jr. (2002). *Testing trap: How states' writing assessments control learning.* New York: Teachers College Press.

Johnson, T. S., Smagorinsky, P., Thompson, L., & Fry, P. G. (2003). Learning to teach the five-paragraph theme. *Research in the Teaching of English, 38,* 136–176.

Ketter, J., & Pool, J. (2001). Exploring the impact of a high-stakes direct writing assessment in two high school classrooms. *Research in the Teaching of English 35, 3*44-393.

Kohn, A. (2000). *The case against standardized testing.* Portsmouth, NH: Heinemann.

Labaree, D. (2007.) *Education, markets, and the public good.* London: Routledge.

Lather, P. (2004.) Scientific research in education: A critical perspective. *British Educational Research Journal 30,* 759-772.

Lemann, N. (1999). *The big test: The secret history of the American meritocracy.* New York: Farrar, Straus and Giroux.

Loofbourrow, P. (1994). Composition in the context of the CAP: A case study of the interplay between composition assessment and classrooms. *Educational Assessment, 2*(1), 7–49.

Luke, A., Green, J., & Kelly, G. J. (2010) Introduction: What counts as evidence and equity? *Review of Research in Education 34,* vii-xvi.

Miller, C. et al. (2006). *A test of leadership: Charting the future of US higher education.* Washington, DC: US Department of Education.

National Council of Teachers of English. (2002). *Resolution on the reading first initiative.* Retrieved from http://www.ncte.org/positions/statements/readingfirst.

National Council of Teachers of English & Council of Writing Program Administrators. (2008) *NCTE-WPA White paper on writing assessment in colleges and universities.* Retrieved from http://wpacouncil.org/whitepaper

National Reading Panel. (2000). *Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction.* National Institute of Child Health and Human Development/US Department of Education. Retrieved from http://www.nationalreadingpanel.org

National Writing Project. (2006, 2008). *Analytic Writing Continuum*. Berkeley, CA: National Writing Project.

National Writing Project. (2010). *Writing project professional development continues to yield gains in student writing achievement.* (Research Brief.) Retrieved from http://www.nwp.org/cs/public/download/nwp_file/14004/FINAL_2010_Research_Brief.pdf?x-r=pcfile_d

O'Brien, R. (2001). An overview of the methodological approach of action research. In R. Richardson (Ed.), *Theory and practice of action research.* Retrieved from http://www.web.net/~robrien/papers/arfinal.html#_Toc26184650

Pagano, N., Bernhardt, S. A., Reynolds, D., Williams, M., & McCurrie, M. K. (2008). An interinstitutional model for college writing assessment. *College Composition and Communication 60,* 285-320.

Ravitch, D. (2010). *The life and death of the great American school system.* New York: Basic Books.

Rudd, A., & Johnson, R. B. (2008). Lessons learned from the use of randomized and quasi-experimental field designs for the evaluation of educational programs. *Studies in Educational Evaluation 34,* 180-188.

Scherff, L., & Piazza, C. (2005.) The more things change, the more they stay the same: A survey of high school students' writing experiences. *Research in the Teaching of English, 39(3),* 271–304.

Shavelson, R. J., & Towne, L. (Eds.). (2002). *Scientific research in education.* Washington, DC: National Academy.

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

Swain, S. S., & LeMahieu, P. (in press). Assessment in a culture of inquiry: The story of the National Writing Project's analytic writing contiuum. In N. Elliot & L. Perelman (Eds.), *Writing assessment in the 21st century: Essays in honor of Edward M. White.* Cresskill, NJ: Hampton.

Swain, S. S., LeMahieu, P., Sperling, M., Murphy, S., Fessahai, & Smith, M. (2010). Writing assessment and its impact on scorers. Paper presented and the Annual Conference of the American Educational Research Association, Denver, CO.

Thaiss, C., & T. Myers Zawacki. (2006). *Engaged writers and dynamic disciplines.* Portsmouth, NH: Heinemann.

Traschel, M. (1992). *Institutionalizing literacy: The historical role of the college entrance examinations in English.* Carbondale, IL: Southern Illinois University Press.

United States Department of Education. (1998). *Reading First funding guidelines.* Retrieved from http://www2.ed.gov/programs/readingfirst/index.html

United States Department of Education Institute of Education Science. What Works Clearinghouse. (2008). Procedures and Standards Handbook Version 2.0. Retrieved November 1, 2010, from http://ies.ed.gov/ncee/wwc/references/idocviewer/Doc.aspx?docId=19&tocId=4#design

United States Department of Education Institute of Education Science. (2010). *Investment in Innovation Fund.* Retrieved from http://www2.ed.gov/programs/innovation/index.html

United States Department of Education Institute of Education Science What Works Clearinghouse. (2011). *Welcome to WWC.* Retrieved from http://ies.ed.gov/ncee/wwc/

United States Department of Education Office of the Inspector General. (2006). *The Reading First Program's grant application process: Final inspection report.* Retrieved from www.ed.gov/about/offices/list/oig/aireports/i13f0017.pdf.

Voluntary System of Accountability. (n.d.). *VSA Online.* Retrieved from http://www.voluntarysystem.org/index.cfm?page=homePage.

Voluntary System of Accountability College Portraits. (2009). *College portraits of undergraduate education.* Retrieved from http://www.collegeportraits.org

Wallace, V. L. (2002). Administrative direction in schools of contrasting status: Two cases. In G. Hillocks Jr. (Ed.), *The testing trap: How state writing assessment control learning* (pp. 93–102). New York: Teachers College Press.

Walvoord, B., & Banta, T. (2004). *Assessment clear and simple: A practical guide.* San Francisco: Jossey-Bass.

West, C. (2009.) *The American evasion of philosophy: A geneology of pragmatism.* Madison, WI: University of Wisconsin Press.

Wiseman, A. W. (2010). Uses of evidence for educational policymaking: Global contexts and international trends. *Review of Research in Education 34,* 1-24.

Yatvin, J., Weaver, C., & Garan, E. (1998). *Reading First cautions and recommendations.* Retrieved from http://www.edresearch.info/reading_fir