Annual Review of Applied Linguistics (2009) **29**, 118–131. Printed in the USA.
Copyright © 2009 Cambridge University Press 0267-1905/09 $16.00
doi:10.1017/S0267190509090102

9. THE INTERSECTION OF TEST IMPACT, VALIDATION, AND EDUCATIONAL REFORM POLICY

Micheline Chalhoub-Deville

The article addresses the intersection of policy, validity, and impact within the context of educational reform in U.S. schools, looking in particular at the No Child Left Behind (NCLB) Act (2001). The discussion makes a case that it is important to reconsider the established views regarding the responsibility of test developers and users in investigating impact given the conflated roles of developers and users under NCLB. The article also introduces the concept of social impact analysis (SIA) to argue for an expansion of the traditional conceptualization of impact research. SIA promotes a proactive rather than a reactive approach to impact, in order to inform policy formulation upfront.

Introduction

The present article addresses the intersection of policy, validity, and impact within the context of educational reform in U.S. schools, looking in particular at the No Child Left Behind Act 2001 (NCLB, Public Law 107–110). The discussion focuses on the relationship between validity and impact and, in turn, on the responsibility of test developers and users in investigating impact. It is argued that the position articulated in the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education [AERA, APA, and NCME], 1999) regarding validation deserves reexamination given the merged roles of developers and users under NCLB. The article also introduces the concept of social impact analysis (SIA) from fields such as anthropology, sociology, and environmental science to argue for an expansion of the traditional conceptualization of impact research. SIA promotes a proactive rather than a reactive approach to impact, in order to inform policy formulation upfront.

NCLB Policies

The driving force for the discussion in this article is the policy mandates of NCLB for English language learners (ELLs) in U.S. schools. Therefore, a brief

https://doi.org/10.1017/S0267190509090102 Published online by Cambridge University Press

overview of NCLB is called for. (For a more detailed treatment of the topic, see Chalhoub-Deville and Deville [2008] and Menken [2008] and also Menken in this volume.) NCLB policies are formulated to shape and supposedly improve educational practices and outcomes in public schools. NCLB mandates the development of content and academic language standards accompanied by performance levels, and calls for standards-referenced assessments (SRAs) to measure student performance in grades 3–8 and in one high school grade in the areas of math and reading. Science is also tested but under different conditions. All students, including those previously excluded from testing, such as the ELL student group, sit for these exams. With regard to ELLs, NCLB states that students whose limited English language proficiency precludes them from being administered content area tests in English must be given academic English language proficiency tests. The regulations that stipulate the specifics for the academic language assessment of ELLs are presented in NCLB Title III. Title III mandates testing of ELLs' academic language abilities in the four modalities of listening, speaking, reading, writing, and comprehension. Comprehension has been interpreted to mean some combination of listening and reading scores. ELLs are expected to show progress on Title III SRAs in order to quickly be able to take the state's standardized achievement tests in English.

Public schools are held accountable for the performance of all students, including ELLs. NCLB specifies sanctions against schools who do not exhibit consistent progress from one year to the next, called adequate yearly progress (AYP). Failure to meet AYP goals over a period of time leads to that school being designated as "in need of improvement," which could lead to the restructuring or reconstitution of the school. One of the criteria for meeting AYP requirements is the demonstrated progress of ELLs on either (or both) the English language proficiency test and the state's content area achievement tests.

Impact

In discussing the influence of testing practices on individuals, groups, institutions, and society, the language testing field has employed terms such as *washback, backwash, impact,* and *consequences.* Although many researchers in the field (see the special issue of *Language Testing,* 1996; Cheng, 2008) have viewed these terms as interchangeable, some have tried to draw distinctions among their uses (Turner, 2001). Hamp-Lyons (1997), for example, contended that the term *washback* is too narrow, so she promulgated the use of *impact* as a broader and more appropriate term. In this article, following Hamp-Lyons's recommendation, the term *impact* is preferred and used. Additionally, given that this article draws heavily on arguments from the measurement literature where the term *consequences* is commonly used, the present article uses *impact* and *consequences* interchangeably.

Somewhat simply, impact analysis calls for documentation of the influence of test content, results, and practices on learning, instruction, and the curriculum. Impact could be examined at the micro or macro level. Investigations could be set up to study impact in terms of direct or indirect influences on test takers, the educational community, or society at large.

https://doi.org/10.1017/S0267190509090102 Published online by Cambridge University Press

Impact is a well-established area of research in language testing. Over the years, the language testing field has accumulated a sizable body of research—articles, chapters, and books—that deal with this topic (e.g., Alderson & Wall, 1993; Bachman, 1990; Chalhoub-Deville & Deville, 2006; Cheng, 2005, 2008; Cheng, Watanabe, & Curtis, 2004; McNamara, 2008; Shohamy, 1993, 1996, 2001; Spolsky, 1997; Wall, 1996, 2005; Wall & Alderson, 1993; special issue of *Language Testing,* 1996). On the whole, language testers have conscientiously engaged in impact research and have argued that impact is integral to any validation work. Many researchers argue that validity must include the sociopolitical consequences of tests and their uses (Bachman, 1990; McNamara, 2008; Shohamy, 2001).

Impact and Validity

The relationship between impact or consequences and validity has long been a contentious topic in the measurement field. Historically, major measurement leaders such as Cronbach (1971, 1988) and Messick (1989a, 1989b, 1996) have agreed on the importance of studying the consequences of tests, but have differed in terms of their representation of impact as an organizing principle in a definition of validity. Although Cronbach acknowledged that test developers and researchers have the obligation to investigate impact, he preferred to maintain the separability of consequences from validity. Messick, on the other hand, proposed that consequences are an integral aspect of validity. In his unified model of validity, Messick argued that scores are always associated with value implications, which are a basis for score meaning and action (interpretation and use), and which connect construct validity, consequences, and policy decisions.

> Attention must . . . be directed to important facets of validity, especially those of value as well as of meaning. This is not a trivial issue, because the value implications of score interpretation are not only part of score meaning, but a socially relevant part that often triggers score-based actions and serves to link the construct measured to questions of applied practice and social policy. (Messick, 1989b, p. 9)

Measurement professionals seem to be divided about whether to include a study of consequences as part of validation. Although some such as Linn (1997) and Shepard (1997) have argued in favor of a broader definition that encompasses impact, Reckase (1998) and Borsboom, Mellenbergh, and van Heerden (2004) contended that test developers should not and cannot bear the burden of investigations of consequences. They proposed a more limited definition of validity that excludes consequences. An important document on this topic is the *Standards for Educational and Psychological Testing* (AERA, APA, and NCME, 1999), which basically contain the official guidelines from the measurement profession. In the section titled "Evidence Based on Consequences of Testing," we find the following statements:

> Evidence about consequences can inform validity decisions. Here, however, it is important to distinguish between evidence that is

https://doi.org/10.1017/S0267190509090102 Published online by Cambridge University Press

> directly relevant to validity and evidence that may inform decisions about social policy but falls outside the realm of validity. . . . Thus evidence about consequences may be directly relevant to validity when it can be traced to a source of invalidity such as construct underrepresentation or construct-irrelevant components. Evidence about consequences that cannot be so traced . . . is crucial in informing policy decision but falls outside the technical purview of validity. (AERA, APA, and NCME, 1999, p. 16)

This excerpt declares that consequences or impact investigations are the responsibility of test developers, but only to the extent that they are related to issues of validity. Additionally, unlike Messick (1989a, 1989b), the *Standards* (AERA, APA, and NCME, 1999) endorse the separation of validity and policy research.

The *Standards* (AERA, APA, and NCME, 1999) argue that an investigation of consequences falls under the purview of test developers when the consequences are the result of construct underrepresentation or construct irrelevant features. Construct underrepresentation is likely to occur, for example, if a test purports to measure academic language use but does not include a measure for classroom speaking. An example of construct irrelevant variance would be the performance of ELLs on NCLB content area assessments in science when they still lack the needed academic language proficiency in English to be able to show their true level of skills and abilities in the subject matter. In this latter case, scores from content area tests are documenting ELLs' poor command of academic language proficiency, which many would argue is irrelevant to the construct being measured.

In short, the position in the *Standards* (AERA, APA, and NCME, 1999) essentially limits test developers' responsibility to pursue impact beyond the confines of construct representation. The *Standards* do acknowledge the importance of investigating the impact of test scores, but that falls outside the realm of validation and is the responsibility of those engaged in social policy.

In the most recent edition of *Educational Measurement,* which typically represents the state of the art knowledge on major topics in the field, Kane (2006, p. 8) questioned the extent to which "*all* consequences of test use" should fall under the heading of validity. More importantly, Kane argued that the measurement field seems to have been preoccupied with validity theory and the theoretical aspects of consequences at the expense of validation practices. Focusing on validation practices, he prompted researchers to rethink the relationship between consequences and validity in terms of positive and negative, and intended and unintended consequences. Within this four-way classification, Kane suggested that positive consequences, especially the intended ones, are not likely to be contested aspects of validity. Kane stated that unintended negative consequences are at the heart of the debate regarding validity, consequences, and social policy, and are the most problematic and contentious.

Negative and unintended consequences are typically attributed to misuses of scores by test users. In reflecting on the negative and unintended consequences, Kane

https://doi.org/10.1017/S0267190509090102 Published online by Cambridge University Press

(2006, p. 8) wrote: "Test developers are understandably incensed when others use scores on their test for a purpose unintended by them (the developers) that has negative consequences for some examinees or stakeholders. Tension rises considerably when users are unwilling to accept responsibility for their role in such misuse" Kane contended that measurement professionals have tried to redress this issue by holding users accountable for their actions. He refers to the *Standards* (AERA, APA, and NCME, 1999, p. 112), which state that while test developers are obliged to provide content and technical documentation about their test and its uses, "the ultimate responsibility for appropriate test use and interpretation lies predominantly with the test user." But as Kane pointed out, this is not a workable solution because test users are rarely interested in or capable of performing validation research.

To summarize, and as Kane (2006, p. 8) aptly posited, "One can surely postulate scenarios in which unintended negative consequences are so egregious that they demand special attention. It is quite another matter to specify who should [or could] bear the burden for preventing or remedying such consequences." Kane sought to move discussion away from theoretical considerations of validity and social consequences to the delineation of more concrete responsibilities in terms of investigating consequences. This issue is especially relevant in today's NCLB policy-driven testing where the roles of developers and users are confounded.

Impact, Validation, and Responsibility Under NCLB

A critical question in recent considerations of validation is who is responsible for addressing social consequences. Traditionally, this question has not received appropriate attention in the language testing or the measurement field at large (Kane, 2006; McNamara, 2008; Nichols & Williams, 2008; Shohamy, 2001). With the advent of NCLB policies, the issue has been rendered quite complex because NCLB blurs the line between test developers and users. The issue, therefore, demands immediate and serious consideration. A useful framework for investigating the responsibilities of test developers and users is advanced by Nichols and Williams (2008). The framework posits three dimensions to flesh out test developers' and users' responsibilities. The dimensions include *breadth of construct, test use,* and *time.* The framework shows the section where test developers are held responsible, that is, upper right corner, and that where test users are responsible, that is, the lower left section. The framework also depicts an area called *zone of negotiated responsibility* (ZNR), which delineates the responsibility for addressing impact needs (Figure 1).

Breadth of construct refers to the extent to which the construct is narrowly or broadly defined. The responsibility of test developers depends on the advertised breadth of the construct. The broader the construct representation, the more extensive are test developers' responsibility for documenting the impact of score use. With broad constructs, such as academic language proficiency, required by NCLB Title III, test developers are responsible for broader documentation of impact (e.g., potential construct irrelevance and underrepresentation). Various stakeholders tend to agree that this is unquestionably a test developers' responsibility (AERA, APA, and NCME, 1999).

https://doi.org/10.1017/S0267190509090102 Published online by Cambridge University Press

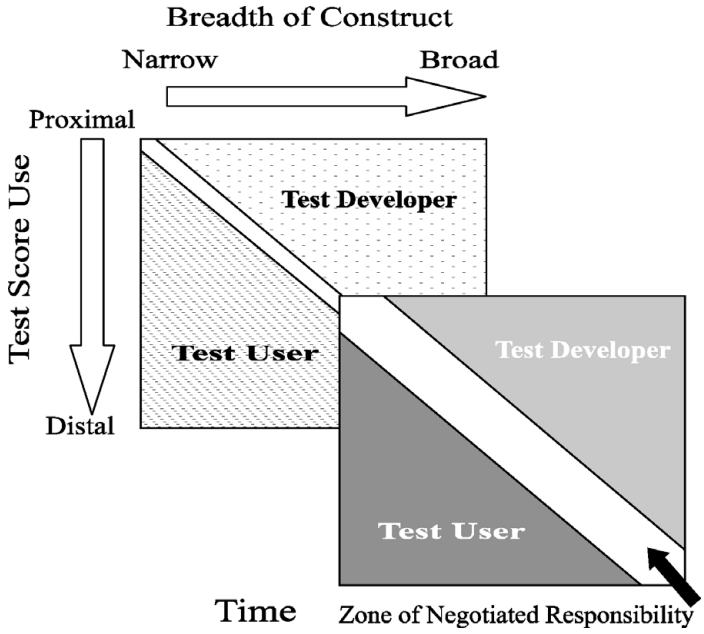


Figure 1. Delineating the responsibilities of test developers and test users

The *test use* dimension refers to the extent to which the test is employed as originally intended by the test developer. In principle, the greater the distance from the original, intended test score use, that is, toward the distal end of the continuum, the more the test user must be held responsible for documenting the evidence to support the impact of score use. An example of the distal use of scores pertains to sanctions against schools and teachers based on students' performance on NCLB tests. One could argue that developers were aware of the sanctions at the design stage and hence they are implicated and need to collect evidence to support the use of student performance data to punish schools and teachers. Clearly, however, test users are equally responsible. The situation necessitates that developers and users enter into the ZNR to negotiate their respective roles and responsibilities in determining the impact of the policy on those affected and any claims regarding educational progress.

The *time* dimension refers to the amount of time since the test has been published. Nichols and Williams (2008) stated that "test developer's responsibilities to collect evidence of test score use may expand (or shrink) over time as experience with test score use increases" (p. 19). In other words, a test may be used in an unintended manner and the test developer may not feel responsible for examining impact. With the passage of time, however, test developers cannot continue to look at the unintended use as unforeseen by them. As any use becomes common practice, test developers and users move into the ZNR, where they have to discuss and agree on a plan to address the impact of what has become intended use. Some might argue that test developers are more likely to get involved to prevent actions that may be injurious

https://doi.org/10.1017/S0267190509090102 Published online by Cambridge University Press

to their sales but less likely to intervene when uses are increasing their revenues. Nevertheless, professional practice dictates that developers attend to how the market has employed the test and work with test users to ensure appropriate validation support of expanded uses.

The framework by Nichols and Williams (2008) moves us away from a static and into a more fluid representation of validation, impact, and responsibilities. This fluidity and its incorporation of the notion of the ZNR is especially helpful in educational reform policies like NCLB where government agencies are explicitly dictating testing requirements and are increasingly shaping how scores should be interpreted and used. In other words, government agencies seem to be controlling and overseeing the roles of the test users and the test developers.

If we examine the groups of test developers and users under NCLB, we note that for the most part, they are the same. Federal and state agencies are, of course, test users: They undertake curriculum changes or instructional intervention based on test scores and issue rewards or sanctions. Yet, how are these agencies considered test developers? The following quote from Nichols and Williams (2008, p. 12) helps explicate this role: "the federal government has passed federal legislation specifying features of the test, the state legislature has passed state legislation specifying additional features of the test and the state board of education may have approved the test design." Additionally, the various state educational agencies are responsible for identifying the content standards and achievement levels, that is, interpreting student performance in terms of meeting proficiency requirements on NCLB tests.

Under NCLB, states and educational groups or agencies are not simply implementing the assessment tools test developers provide them, but are dictating, to a large extent, all critical specifications of the entire testing program. Nevertheless, while these agencies are assuming a new role as test developers, they are not engaged in the traditional responsibilities of test developers such as conducting validation research. The concern, as raised above, is that educational agencies are not as well equipped as conventional test developers to perform needed and responsible investigations. The excuse, however, cannot continue to hold water, given the growing and more aggressive control of government agencies of the work of test developers and users. At the very least, agencies can commission such work as needed. What is significant and important to note here is that these issues need to be addressed explicitly and upfront. In conclusion, with this conflated role of test developer and test user, it becomes all the more important for individuals and groups to confer and address the shared responsibility of dealing with impact, whether that impact evidence relates directly to the construct or more broadly to the educational profession and society at large.

The discussion thus far has focused primarily on after-the-fact impact analysis where the policy is in place and the policy-driven tests are operational. This is a reactive approach to impact investigation, and while it serves a critical function, it is not sufficient. In the next section of the article, I propose expanding the reactive

https://doi.org/10.1017/S0267190509090102 Published online by Cambridge University Press

conceptualization of impact research to include a proactive approach to policy formulation and implementation.

Social Impact Analysis

Discussion of impact in this section deals with the need to expand the framework for examining test impact. Given the ever increasing purview of educational reform policies such as NCLB, it is all the more critical for test developers and researchers to become active partners in the formulation of policies that better serve education and have a more favorable chance of accomplishing policy goals. A useful concept for the present argument is social impact assessment (SIA), which is commonly employed in fields such as anthropology, sociology, tourism, and environmental science. According to Barrow (2000), "modern SIA is a field that draws on over three decades of theoretical and methodological development to improve foresight of future change and understanding of past developments" (p. 1). SIA practices are intended "to help individuals, communities, as well as government and private sector organizations understand and be able to *anticipate* the possible social consequences on human populations and communities of proposed project development or policy changes" (Burdge, 2007, emphasis added). This approach to impact analysis differs from prevalent practices in language testing and the measurement community at large by its focus on impact analysis even before a policy is put in place. SIA emphasizes anticipatory impact and a proactive approach to studying potential consequences. SIA proposes systematic analysis of the foreseen implications of policies such as NCLB. The analysis calls for the evaluation of intended and potentially unintended impact of policy mandates and the need to formulate mitigation plans to address anticipated negative impact.

The following are examples of adverse NCLB mandates that could have been avoided had proactive impact analysis been undertaken. These examples illustrate how, in one instance, negative impact could have been easily predicted and remedied, and in a second situation, why a proactive analysis would have necessitated plans for mitigation of adverse impact. In the first instance, NCLB mandates that schools report annually on the performance of ELLs as a separate group. Additionally, NCLB states that once students are designated as proficient, they are moved out of the ELL category for reporting purposes. Yet school officials filed complaints that these proficient ELLs represent the better-performing students that schools worked hard to get to the proficient level. By taking the proficient students out of the ELL category, schools can no longer take credit for these students in terms of AYP reporting for the ELL category. In response to these complaints, the U.S. Department of Education now allows schools to include the performance of proficient ELL students in the ELL category for up to 2 years after they have been redesignated as proficient. This aspect of the policy could have been predicted and the problem avoided had SIA been performed at the time of policy formulation.

The second example where negative impact could have been minimized had SIA been carried out is the mandating of SRAs when standards have not yet been developed. SIA could have documented the impoverished state of available standards,

https://doi.org/10.1017/S0267190509090102 Published online by Cambridge University Press

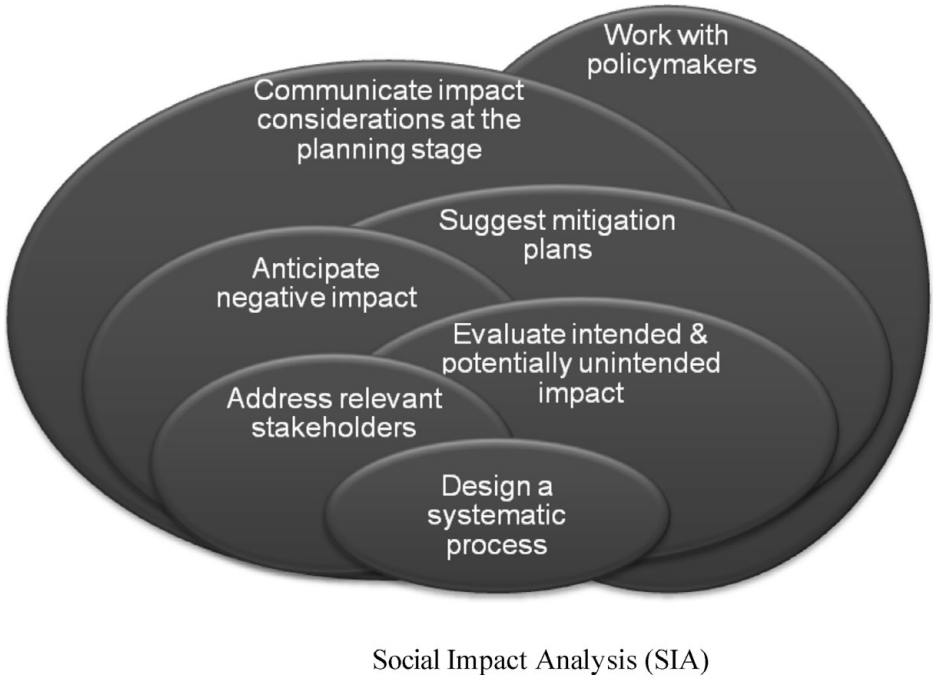


Figure 2. Outline of social impact analysis actions

especially in terms of academic language ability, and made the case for more lead time to allow states to develop rigorous standards before SRAs were mandated to become operational. (For a more detailed discussion on the state of content and language standards in the United States, see Chalhoub-Deville & Deville, 2008).

It behooves us, as responsible professionals, to demand and engage in impact research at the policy planning level. It is important that we not only communicate our SIA findings but also work with policymakers. As researchers, we typically confine ourselves to our offices and rationalize that it is not our responsibility to lobby the legislature and policymakers. But I argue that it should be our responsibility as responsible professionals. We need to work with policymakers to ensure not only better policy, but also to study and understand potential consequences. It is important to emphasize that *working with policymakers* does not mean merely debriefing them on the research findings. This is not likely to be effective. We need to construct a more involved and sustained relationship with them. Language testers need to become—if not as individuals then as groups—more engaged in policy formulation research. We, language testers, and our language testing organizations are remiss in this arena.

In summary, and as Figure 2 shows, SIA involves the following actions:

- Design a systematic process of investigating impact.
- Address relevant stakeholder individuals and organizations.

https://doi.org/10.1017/S0267190509090102 Published online by Cambridge University Press

- Evaluate intended and potentially unintended impact.
- Anticipate negative impact.
- Suggest mitigation plans.
- Communicate impact considerations at the policy planning stage.
- Address policymakers, and better yet work with policymakers.

Thus far, discussion has not ventured into the type of research methodologies deemed appropriate for carrying out SIA research. To address this issue it is instructive to consider what research methodology mandates have been put in place with policies such as NCLB. It is safe to say that within the education community this topic is very contentious. The U.S. Department of Education dictates the nature of methodological research for carrying out policy-related investigations and, consequently, for receiving research funds.

> In numerous passages, NCLB details that rigorous educational research on improving educational practices should equate with the use of experimental designs (i.e., randomized, controlled field trials). This preference for a particular type of design has also translated into proposed criteria for federal funding of future educational policy research. (Heck, 2004, p. 182)

As might be expected, researchers and educators have questioned the legitimacy of the government to dictate and limit the scope of educational research and have contended that positivism and quantitative research is a damaging reduction of the sociopolitical realities of educational reform (Heck, 2004).

Arguments have been presented that experimental and quasi-experimental research studies, which are successful tools in the natural science fields, are less useful in educational settings where we are dealing with complex social realities (McGroarty, 2002; Tollefson, 2002). Many reject what they call pseudoscientific neutrality and call for critical perspectives, which explore "the links between language policies and inequalities of class, region, and ethnicity/nationality" (Tollefson, p. 5). These critical perspectives seek to use research and policies to promote social justice. In conclusion, if current NCLB policy practices are an indication of what to expect for SIA investigations, then critical language testers who are likely to favor a broader research perspective need to be prepared to argue vigorously for diverse methodologies.

Conclusion

At the heart of this article is the pervasive influence of educational reform policies on U.S. schools and the ELL students. These policies are increasingly relying on testing systems to affect change, which puts test developers and researchers at a critical cross road in terms of their roles and responsibilities. Given the conflated roles of developers and users under NCLB, the restricted view of impact research in validation, as conveyed in the *Standards* (AERA, APA, and NCME,1999), is no

https://doi.org/10.1017/S0267190509090102 Published online by Cambridge University Press

longer tenable. NCLB requires a more complex understanding of impact research and an engagement in negotiations to allocate and share the responsibilities for carrying out impact research. Additionally, given the comprehensive influence of NCLB policies on education in the schools, it is all the more critical for researchers to engage in proactive impact research, that is, SIA, to inform more reasoned educational reform policies. Finally, SIA investigations cannot be restricted to any one research paradigm. Methodologies from diverse paradigms are needed to investigate the complex contexts in which such policies are being implemented.

ANNOTATED REFERENCES

Chalhoub-Deville, M., & Deville, C. (2006). Old, borrowed, and new thoughts in second language testing. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 516–530). Washington, DC: National Council on Measurement in Education & American Council on Education.

The article provides a comparative analysis of European and North American large-scale, standardized testing practices. It points out that testing practices are typically driven by various pragmatic purposes. However, what distinguishes policy-driven testing systems is their imperviousness to professional criticisms and research findings. The article calls for the profession to be more engaged to curb the impact of such tests.

Cheng, L., (2008). Washback, impact and consequences. In E. Shohamy & N. H. Hornberger (Eds.), *Encyclopedia of language and education, Vol. 7: Language testing and assessment* (2nd ed., pp. 349–364). Dordrecht, The Netherlands: Springer.

The article provides a historical as well as an up-to-date review of the terms washback, impact, and consequences together with the associated arguments and research. In addition, the article outlines some of the shortcomings of research performed so far in this area and makes recommendations on how to move forward.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Washington, DC: National Council on Measurement in Education & American Council on Education.

The 2006 edition of Educational Measurement is the fourth in the series. A period of about 15 years tends to separate the publication of the various editions. Educational Measurement typically includes chapters that represent changes in the measurement field since the publication of the last edition. The chapter by Kane, titled "Validation," follows the 1989 "Validity" chapter by Messick. The chapter explores the practical aspects of validity investigations.

https://doi.org/10.1017/S0267190509090102 Published online by Cambridge University Press

Shohamy, E. (2001). *The power of tests: A critical perspective on the uses of language tests*. Essex, England: Longman.

Shohamy is a pioneer of critical language testing, which has been embraced by researchers such as Lynch and McNamara. In this book, and as part of critical language testing, Shohamy seeks to move the language testing field to tackle, in addition to psychometrics, the sociopolitical dimensions of testing systems. The book relies on arguments as well as research to advance its claims.

OTHER REFERENCES

Alderson, J. C., & Wall, D. (1993). Does washback exist? *Applied Linguistics, 14*, 115–129.

Alderson, C., & Wall, D. (Eds.). (1996). Special issue. *Language Testing, 13*, 239–354.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (AERA). (1999). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

Bachman, L. F. (1990). *Fundamental considerations in language testing.* Oxford, England: Oxford University Press.

Barrow, C. J. (2000). *Social impact assessment: An introduction.* Oxford, England: Oxford University Press.

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The concept of validity. *Psychological Review, 111*, 1061–1071.

Burdge, R. J. (2007). Retrieved March 9, 2009, from http://www.socialimpactassessment.net/

Chalhoub-Deville, M., & Deville, C. (2006). Old, borrowed, and new thoughts in second language testing. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 516–530). Washington, DC: National Council on Measurement in Education & American Council on Education.

Chalhoub-Deville, M., & Deville, C. (2008). National standardized English language assessments. In B. Spolsky & F. Hult (Eds.), *Handbook of educational linguistics* (pp. 510–522). Oxford, England: Blackwell.

Cheng, L. (2005). Changing language teaching through language testing: A washback study. Cambridge, England: University of Cambridge ESOL Examinations and Cambridge University Press.

Cheng, L. (2008). Washback, impact and consequences. In E. Shohamy & N. H. Hornberger (Eds.), *Encyclopedia of language and education, Vol. 7: Language testing and assessment* (2nd ed., pp. 349–64). Dordrecht, The Netherlands: Springer.

Cheng, L., Watanabe, Y., & Curtis, A. (Eds.). (2004). *Washback in language testing: Research contexts and methods*. Mahwah, NJ: Erlbaum.

Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 443–507). Washington, DC: American Council on Education.

https://doi.org/10.1017/S0267190509090102 Published online by Cambridge University Press

Cronbach, L. J. (1988). Five perspectives on validity argument. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 3–17). Hillsdale, NJ: Erlbaum.

Hamp-Lyons, L. (1997). Washback, impact and validity: Ethical concerns. *Language Testing, 14*, 295–303.

Heck, R. (2004). *Studying educational and social policy: Theoretical concepts and research methods*. Mahwah: NJ: Erlbaum.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Washington, DC: National Council on Measurement in Education & American Council on Education.

Linn, R. L. (1997). Evaluating the validity of assessments: The consequences of use. *Educational Measurement: Issues and Practice, 16*, 28–30.

McGroarty, M. (2002). Evolving influences on educational language policies. In J. W. Tollefson (Ed.), *Language policies in education: Critical issues* (pp. 17–36). Mahwah, NJ: Erlbaum.

McNamara, T. (2008) The social-political and power dimensions of tests. In E. Shohamy & N. H. Hornberger (Eds.), *Encyclopedia of language and education Vol. 7: Language testing and assessment* (2nd ed., pp. 415–27). Dordrecht, The Netherlands: Springer.

Messick, S. (1989a). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). Washington, DC: American Council on Education & National Council on Measurement in Education.

Messick, S. (1989b). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher, 18*, 5–11.

Messick, S. (1996). Validity and washback in language testing. *Language Testing, 13*, 241–256.

Nichols, P., & Williams, N. (2008). *Evidence of test score use in validity: Roles and responsibility*. Paper presented at the annual meeting of the National Council on Measurement in Education. New York.

No Child Left Behind. (2001). Act of 2001, Pub. L. No. 107–110, 115 Stat. 1425.

Reckase, M. (1998). Consequential validity from the test developer's perspective. *Educational Measurement: Issues and Practice, 17*, 13–16.

Shepard, L. A. (1997). The centrality of test use and consequences for test validity. *Educational Measurement: Issues and Practice, 16*, 5–8, 13, 24.

Shohamy, E. (1993). *The power of tests: The impact of language tests on teaching and learning*. Washington, DC: National Foreign Language Center Occasional Papers.

Shohamy, E. (1996). Testing methods, testing consequences: Are they ethical? Are they fair? *Language Testing, 13*, 340–349.

Shohamy, E. (2001). *The power of tests: A critical perspective on the uses of language tests*. Essex, England: Longman.

Spolsky, B. (1997). The ethics of gatekeeping tests: What have we learned in a hundred years? *Language Testing, 14*, 242–247.

Tollefson, J. W. (2002). Introduction: Critical issues in educational language policy. In J. W. Tollefson (Ed.), *Language policies in education: Critical issues* (pp. 3–16). Mahwah, NJ: Erlbaum.

Turner, C. (2001). The need for impact studies of L2 performance testing and rating: Identifying areas of potential consequences at all levels of the testing cycle.

https://doi.org/10.1017/S0267190509090102 Published online by Cambridge University Press

In M. Milanovic & C. J. Weir (Eds.), *Studies in language testing: Vol. 11: Experimenting with uncertainty: Essays in honour of Alan Davies*. (pp. 138–149). Cambridge, England: Cambridge University Press.

Wall, D. (1996). Introducing new tests into traditional systems: Insights from general education and from innovation theory. *Language Testing, 13*, 334–357.

Wall, D. (2005). The impact of high-stakes examinations on classroom teaching: A case study using insights from testing and innovation theory. Cambridge, England: University of Cambridge ESOL Examinations and Cambridge University Press.

Wall, D., & Alderson, J. C. (1993). Examining washback: The Sri Lankan impact study. *Language Testing, 10*, 41–69.

https://doi.org/10.1017/S0267190509090102 Published online by Cambridge University Press