Exceptionality Education International

Volume 18 | Issue 2 Article 4

5-1-2008

Issues in Including Students with Disabilities in Large-scale Assessment Programs

Christopher DeLuca
Queen's University, 2cd16@queensu.ca

Abstract

Large-scale assessment programs are becoming increasingly common throughout Canada and the United States. Given the emphasis on inclusive education in North America, special education students are largely expected to participate in these programs. However, several challenges exist for educators, policymakers, and psychometricians with respect to including students with disabilities in large-scale assessments. This article is a critical interpretive review of the academic lit-erature in this area intended to identify and examine issues pertinent to inclusive practice. In particular, attention is given to consequences (both positive and neg-ative) of including students with disabilities in large-scale assessments, validity of assessment results, provisions for accommodations, and research limitations. Areas for continued research are also considered.

ISSN 1918-5227

Pages 38- 50

Special thanks to Dr. Nancy L. Hutchinson for her support in reviewing this article.

Issues in Including Students with Disabilities in Large-scale Assessment Programs

Cover Page Footnote

Special thanks to Dr. Nancy L. Hutchinson for her support in reviewing this article.

Exceptionality Education Canada
2008, Vol. 18, No. 2, pp. 38-50

Issues in Including Students with Disabilities in Large-scale Assessment Programs

Christopher DeLuca
Queen's University

Abstract

Large-scale assessment programs are becoming increasingly common throughout Canada and the United States. Given the emphasis on inclusive education in North America, special education students are largely expected to participate in these programs. However, several challenges exist for educators, policymakers, and psychometricians with respect to including students with disabilities in large-scale assessments. This article is a critical interpretive review of the academic literature in this area intended to identify and examine issues pertinent to inclusive practice. In particular, attention is given to consequences (both positive and negative) of including students with disabilities in large-scale assessments, validity of assessment results, provisions for accommodations, and research limitations. Areas for continued research are also considered.

It was my first week as a practicing teacher in a new school. The school was a *special* school including only students identified as disabled. In order to attend, students needed to demonstrate that their disability was more severe than others. The school was full to capacity with most students having multiple difficulties from physical to cognitive to emotional. Given the significant range of abilities, students were working on accommodated provincial curriculum in classes of 15-20 students. My arrival in early fall was accompanied by a week of large-scale assessment mandated by the provincial government. The school had too many exclusion appeals to file so all students from Grade 10 upward were expected to write the assessment. In contrast to most schools in the board where the test was administered only to Grade 10 students—with the exception of a few students who did not pass—this school's 99% fail rate meant these students would need to write the test each year until Grade 12.

In anticipation of their impending failure, some students did not bother to attend school during test week; others would come in, write their name on the test, skim through a few pages, and hand it back blank. Several hopefuls would try the test, sitting for the extended double-time slots writing brief responses between long daydreams. During assessment week, student morale

was lower than usual; some students were angry at the test, others were discouraged–saddened by their lack of ability. I recall one student's comment: "This test is stupid. None of us are going to pass anyway. We're not smart here like kids in other schools." While I admit I was only a student teacher, I could not rationalize this assessment process. To me, this task seemed demoralizing for students and an exercise in wasted learning time. I brought this assessment dilemma back to my professors and colleagues at the Faculty of Education. Other teacher candidates had similar experiences with one or two of their exceptional students. After much deliberation, we construed the experience as a mechanism of system accountability, providing data for public reports on system effectiveness. But was this sufficient justification for the assessment consequences I observed in my first school?

Accountability has been a central construct in recent educational reforms throughout North America (McEwen, 1995). As a result of this movement towards standards-based education, accountability has become synonymous with testing (Froese-Germain, 2001). In addition, the accountability reform has paralleled and been coupled with the shift from dual programming to a unified curriculum for general and special education students. In Canada, inclusive education initiatives are supported nationally through the Canadian Charter of Rights and Freedoms and the Canadian Human Rights Act of 1977, which stipulates equality of rights regardless of mental or physical disability. Further, each province and territory has specific legislation on the educational rights of individuals with disabilities. By and large, this legislation also encourages inclusive policies and practices. When combined, inclusion and accountability have considerable academic and social consequences for students with disabilities, perhaps more so than any other student population. This paper begins to examine the consequences of these practices through a review of existing research literature.

This article uses a critical interpretive review methodology to consider inherent issues to inclusive large-scale assessment practices. Employing an interpretivist approach situates theoretical assumptions within complex and subjective social and pragmatic frameworks, thereby linking the spirit of a theory to the context of its use (Eisenhart, 1998). An interpretive review not only organizes and synthesizes research—which is onto itself a useful activity—but also problematizes and contextualizes research to encourage further thinking in a given area (Schwandt, 1998). In particular, this interpretive review is intended to raise questions for practitioners, policymakers, and researchers, and promote deeper consideration about the practice of including students with disabilities in provincial testing programs. This review focuses primarily on the Canadian assessment context; however, given that large-scale assessment programs are relatively new in most provinces, there is little research published strictly from a Canadian perspective. As such, relevant literature from the United States is also included.

The Canadian Large-scale Assessment Context

Given the move towards standards-based education throughout much of Canada, large-scale assessments are occupying a more prominent role within educational systems. Large-scale assessment programs in Canada serve multiple purposes including accountability, monitoring student achievement, gatekeeping, and instructional diagnosis (Klinger, DeLuca, & Miller, 2008). Depending on the programs' purposes, the results from these assessments can have consequences for individual students and for the educational system at large. At a student level, assessment results may lead to placement decisions, remediation and programming modifications, privileges such as graduation or postsecondary acceptance, as well as a range of emotional

and social fallouts. At a systems level, data obtained from large-scale assessment programs may be used to inform curricular reforms, allocate resources, and support reports for public accountability. Currently, every province and territory with the exception of Prince Edward Island has at least one, and more commonly, multiple provincially administered assessment programs (Klinger et al., 2008). These programs are in addition to national testing initiatives such as the Student Achievement Indicator Program in Canada and international assessments such as the Programme for International Student Assessment. Provincial assessment programs are often characterized by a series of standardized tests administered at various grade levels and in various subject areas. While some exclusion policies exist for these assessments, the emphasis on inclusive education in North America has placed pressure for all students to participate in provincial/territorial testing (e.g., Council for Exceptional Children, 2004).

Large-scale assessments are relatively recent additions to provincial systems of education, with the exception of Alberta and British Columbia which have substantial testing histories (Klinger et al., in press). As such, administration and planning of these assessment programs must consider special student populations, especially in light of the dominant legislation supporting inclusive education. Earl (1995) stated that providing appropriate assessment for diverse student groups is a challenge for psychometricians, educational policymakers, and educators. Further she attested that inclusive assessment practices raise significant questions about assessment validity. As a case-in-point, Ontario provides a recent example of the implementation of a provincial assessment program. The central tenet regarding inclusion on the province's grades 3, 6, 9, and 10 assessments was that "educators must make every effort to enable exceptional students to participate in all aspects of the assessment to demonstrate their learning" (Hutchinson, 2007, p. 234). Bennett, Dworet, and Daigle (2001) argued that the swift implementation of Ontario's assessment program, coupled with changes in special education programming and funding, has left special education in a prohibitive state at least for the present moment. As the number of students identified as exceptional steadily increases in Ontario, equity remains a central issue in educational reforms (Weber & Bennett, 1999) with continued implications for the province's assessment program. This trend is paralleled in other regions of Canada where exceptional students are required to participate in large-scale assessments.

As part of inclusive assessment practices, students with disabilities may be given testing accommodations and in some instances be granted exclusion. Student participation and decisions surrounding eligibility for accommodations are typically under the control of individual provinces/territories with many provinces favouring inclusion. As an example, Alberta's policy indicates that all students in English as a Second Language and special education programs are expected to partake in provincially administered assessments unless under special circumstances (Lupart, 2001). The decision of student inclusion and provision for accommodation is largely dependent on the local special education team (e.g., Identification, Placement, and Review Committee in the province of Ontario) and based on recommendations in the students' Individual Education Plans (IEP; Destefano, Shriner, & Lloyd, 2001). The argument in support of this strategy is that students in special education typically have unique learning needs resulting in a case-by-case decision making process. In many provinces/territories, it is typical protocol for local special education teams to make an appeal to the governing testing body (i.e., the provincial/territorial ministry of education or affiliate) for individual student exclusion, deferral, or provision for accommodation (Klinger et al., 2008).

All testing programs in Canada maintain policy for accommodation provisions. Accommodations are generally understood to be changes to the test format or the allocation of assisting

resources that serve to diminish the effects of the disability in the measurement of student knowledge and skills. Testing accommodations are typically consistent with those given to students during instructional periods. The underlying theory supporting accommodations in assessment contexts is that the accommodation is assumed not to privilege students with exceptionalities, but rather to enable equitable testing conditions compared with general education students (Shaftel, 2005). On the other hand, few assessment programs offer modified testing. Modifications serve to change the content of what is assessed, resulting in lower levels of comparability. Thus, students who work from a modified curriculum would require a modified assessment. Under most provincial policies, these students would typically be considered for exclusion from the assessment program.

While the overall procedure for making an appeal for testing exceptions is consistent throughout regions, the criteria for determining what exceptions students receive is largely inconsistent. A major dilemma facing inclusive initiatives is a lack of common criteria at national, provincial, and local levels with decision processes varying among schools and districts. In their review of the identification and accommodation practices of three school districts, Shriner and Destefano (2003) found that although formal training led to increased consistency over identification and accommodation provisions written on IEPs, there was still wide variation amongst accommodation use in testing scenarios. This suggests that accommodations are inconsistently applied amongst students with disabilities which reduces comparability and (to a large extent) utility of assessment scores. Thus, inferences drawn from these assessments may have reduced validity. Depending on the purpose of the assessment, scores may lead to misdirected instruction, program modifications, and resource allocation (Froese-Germain, 2001).

The large-scale assessment context in Canada is still evolving. The bridging of assessment policy and practice with inclusive education continues to be a challenge for many policymakers, educators, and psychometricians. As provinces continue to refine their assessment programs, attention to the consequences of inclusive practice, validity of assessment results, and the effects of accommodation provisions is required. The following sections of this article interpret research pertaining to each of these issues.

Consequences of Inclusive Practice

Inclusive practice in large-scale assessment programs holds both intended and unintended consequences for students with disabilities, educators, and policymakers. These consequences may be positive or negative with short and long-term effects. Social and academic consequences can impact students well beyond their formal school years (Thurlow & Johnson, 2000). Given the level of complexity associated with consequential aspects of assessments, there is little research given to this area, especially from a Canadian perspective. As such, much of the literature reviewed has been conducted in the United States.

Large-scale assessments are often classified by the consequences they have on students. Typically, assessments are discerned as high-stakes or low-stakes. Madaus (1988) characterized high-stakes assessments as having perceived or real consequences for students, staff, or schools. For example, since the British Columbia Graduation Provincial Examinations are a requirement for students in order to obtain their secondary school diploma and proceed to university, these assessments have direct consequences on educational decisions and thus can be considered high-stakes. However, Madaus's definition also carries with it the assumption that high-stakes assessments have both intended and unintended consequences, where the importance of the test is

dependent on the perceptions of those directly involved with the test (i.e., test takers and teachers). Therefore, even large-scale assessments that do not explicitly impact educational decisions, such as assessment used for cohort analysis, may still be perceived as high-stakes to the students who write the tests.

Langenfeld, Thurlow, and Scott (1997) reviewed academic literature pertaining to the effects of high-stakes testing on students with disabilities focusing on effects towards curriculum access and student learning, attitudes and school climate, and additional costs and benefits of inclusive practice. The implicit and explicit consequences for curriculum and instruction have been documented in several studies. Shepard and Dougherty (1991) reported that high-stakes assessments served to narrow the enacted curriculum, while Berger and Elson (1996) found that these assessments added focus, coherence, and clarity to instruction. Given the discrepancy in results, Langenfeld et al. (1997) suggested that further research is required into the consequences of large-scale assessments on curriculum and instruction. In addition, little research has been conducted with respect to how high-stakes assessments affect the educational opportunities for students with exceptionalities. One hypothesis, suggested by Langenfeld et al., is that since graduation requirements are increasing throughout many states, the emphasis and resource allocation for fewer academic subjects has decreased, which has a negative impact on students with exceptionalities. With respect to the effects of large-scale assessment on student learning, Langenfeld et al. stated that "to date, there is virtually no evidence that adequately addresses the question of how high-stakes testing affects student learning" (p. 7). Examining how well high-stakes tests measure students' learning is a challenging task because multiple factors contribute to final scores on such assessments. These factors create error in measurement and, therefore, an increase in test score may not necessarily indicate an increase in student learning.

Furthermore, concerns exist that large-scale assessment programs are mechanisms of systemic bias against students with disabilities. Langenfeld et al. (1997) suggested that "studies point to a frightening but very real possibility that children will be systematically and deliberately labeled, excluded, and pushed out of the system altogether in order to improve test scores" (p. 11). In addition, Darling-Hammond (1994) indicated that extensive research supports the argument that large-scale testing actually reinforces and increases social inequalities in educational opportunities. In particular, Darling-Hammond commented that test results are used for social advancements including granting and distributing educational and employment benefits rather than informing and supporting teaching and learning.

The effects of testing on student learning may also be dependent on whether or not the test is norm or criterion referenced. Typically, students with exceptionalities are excluded or denied accommodations when the test is explicitly norm-referenced because these tests tend not to be standardized on special student populations (Langenfeld et al., 1997). Further, in norm-referenced testing the underlying assumption is that a preset proportion of the population will score below the mean. Research suggests that it is often students with disabilities that fall within the lowest percentiles for these assessments. In theory, criterion-referenced assessment allows for greater leniency for the inclusion of students with exceptionalities because criterion-referenced tests do not preset pass rates, and standards or criteria can be established so that students with exceptionalities are able to achieve success. The dilemma with the use of criterion-referenced testing in large-scale educational scenarios is that standards are typically linked to grade-level expectations. Depending on the method for developing curriculum, these standards may reflect norm ability. Therefore, while a province, territory, or state may espouse that they are administering a criterion-referenced test, it is most likely rooted in normative criteria. From

the few studies that investigated the performance of students with exceptionalities on large-scale assessments, it has been demonstrated that these students continuously score below the norm group regardless of whether the assessment is criterion or norm referenced (Darling-Hammond, 1994). Even when provisions for accommodations are made for these students, the overall scores still remain lower than those for general education students.

When considering the effects of high-stakes assessments on school climate and attitude, research suggests that these assessments have both positive and negative consequences. These assessments cause stress and frustration in teachers and students (Langenfeld et al., 1997). In addition, teachers have reported both decreased autonomy and a decreased ability to rely on their professional judgments. For students, assessments have been reported to increase anxiety levels and some data support increased drop-out rates (Darling-Hammond, 1994). As predicted by Berger and Elson (1996), one positive effect on school climate was an increased focus on the collective school mission. However, it is difficult to conclusively state the effects of this benefit on students with exceptionalities. It can be speculated that the effects for these students are the same, if not increased, compared with the effects on general education students.

As mentioned above, the inclusion of students with disabilities in assessment programs may have both positive intended and positive unintended consequences. Ysseldyke, Dennison, and Nelson (2004) conducted a study examining the benefits of student participation in large-scale, high-stakes testing. Multiple methods were used to gather data for this national study including a survey of State Special Education Directors, a media-based questionnaire, focus groups, and a national environmental scan of state assessment policies and practices. Four positive themes were observed: (a) higher participation rates of students with exceptionalities in state-wide assessment programs, (b) overall higher expectations and standards for special education students, (c) improved teaching, and (d) improved student performance (i.e., greater pass rate) and increased access to the general education curriculum. In addition, secondary positive consequences (i.e., not observed across all data sources) suggest overall improved assessments (i.e., multiple measures and greater alignment with state curriculum standards); increased diploma options; decreased drop-out rates; and greater communication between parents, special education teachers, and general education teachers.

Assessing the consequences, both positive and negative, of high-stakes assessments for students with exceptionalities is a complex task. If the majority of these students are performing below norm levels on these assessments, an increase in remediation and specialized programming is required in order to improve test scores (Langenfeld et al., 1997). The costs of such resources are high and may not be warranted against the benefits derived from having these students participate. Langenfeld et al. suggested that alternative assessments, such as portfolios and authentic tasks, may yield a more profitable outcome that more accurately demonstrates the knowledge, skills, and ability of these students. However, incorporating alternative assessments within large-scale assessment programs maintains its own set of challenges, especially with respect to issues of reliability.

The consequential aspect of large-scale assessments is an area of much concern for all students including those with disabilities. Evidence from the United States suggests that both positive and negative consequences exist for students and teachers. As evident from this review, research is required from a Canadian perspective if educators are to better understand the effects large-scale assessment programs have on students and teachers in Canada. This research should seek to link specific testing programs with descriptions of explicit and implicit consequences on students, teachers, curriculum, resource allocation/administrative decisions, and policymaking.

Academic and social consequences must be considered in relation to the multiple purposes espoused by assessment programs because certain assessment programs may call for inclusion while others may not require students with disabilities to participate. Further, consideration of consequences must be made in relation to the validity of assessment inferences given non-standardized (i.e., accommodated) assessment conditions.

Validity of Assessment Results

Validity, as a concept and activity, may pose as one of the greatest threats to the inclusion of students with disabilities in large-scale assessment programs. Broadly defined in relation to assessment, validity represents the value implications of scores and is linked to consequence and action (Messick, 1989). Understanding and improving validity judgments within these contexts is critical because assessment scores often hold high-stakes consequences. Central to the issue of validity is the consideration of fully standardized test administrations (i.e., no accommodations and thus a more homogeneous group of fewer students) versus drawing inferences based on the full range of ability groups when given accommodated testing conditions. Thus, a cost-benefit dilemma exists between including special education student populations and maintaining a homogenous testing group.

In order for valid inferences to be drawn from test scores, content standards, instruction, and test constructs must be aligned (Haladyna & Downing, 2004). Haladyna and Downing identified several threats to test validity: construct under-representation, faulty logic of the causal inferences regarding test scores, misdirected use of test score interpretations, lack of reproducibility of test scores, and construct-irrelevant variance (CIV). In particular, Haladyna and Downing investigated how CIV functions in high-stakes assessments. They argued that in determining test validity, it is first necessary to identify the construct being assessed. Two constructs are possible in high-stakes assessment scenarios. Test questions can either be drawn from a larger domain of knowledge or can seek to evaluate a cognitive ability; however, neither construct is immune to CIV. Haladyna and Downing defined CIV as error variance arising from systematic error. Such error can be either group or person specific, but in either case, it is not random. Because variance in test scores is not a function of the test construct, the result is construct-irrelevant. In addition, systematic error can be directed to all members of a particular examinee group (constant error) or can affect examinees differentially. Special education student populations have been identified as one group contribution to CIV because differences in test scores between general and special education students are largely due to students' cognitive, physical, and behavioural abilities and not the test construct. While accommodations are provided to students with disabilities in an attempt to diminish CIV, given the inconsistent granting of accommodations and lack of research on accommodation effect, the impact of these provisions on CIV is still largely undetermined. At minimum, it could be said that accommodations are not consistently targeting and reducing CIV.

The perception of accommodation provisions and the impact on score comparability has significant implications for the utility and validity of assessment results. Ultimately, if results from large-scale assessments are perceived as invalid, the consequences associated with having exceptional students take these assessments may not be merited. Further, if results are not valid indicators of student achievement, then education reform, program modifications, and resource allocation will inevitably be misdirected.

Based on their 3-year meta-analysis of accommodation-related research across 50 states, Cox, Herner, Demczyk, and Nieberding (2006) found that "educators have tended to think of accommodations narrowly, as adjustments to the assessment process rather than as specific teaching strategies designed to minimize the effects of a student's disability and to maximize a student's ability to learn" (p. 350). In Canada, Brackenreed (2004) specifically examined teachers' perceptions of testing accommodations on the Ontario Secondary School Literacy Test (OSSLT) in an effort to solicit their views toward the validity of accommodations in large-scale assessments. The OSSLT is a province-wide Grade 10 literacy assessment that is intended as a requirement for attaining a secondary school diploma. The requirement policy for the OSSLT stipulates that students who have failed the test or who have been exempt a minimum of two times can enroll in and must pass a literacy course in lieu of passing the OSSLT in order to receive their diploma. Brackenreed surveyed 250 Grade 9 and 10 English teachers from nine school districts in Ontario. The results indicated that teachers do not view all accommodations as being equitable and enabling valid comparisons across students. Teachers considered accommodations as invalidating test results when they perceived the accommodation as modifying the nature of what was being assessed; this was the case for accommodations that altered the test format (e.g., extended time limits, reading a test aloud, reducing the number of items on a page, rewording questions, and teaching test-taking skills). However, accommodations for students with sensory impairments and those that addressed adjustments in response format were perceived as equitable, thereby enabling valid comparisons among students.

Because validity claims are a matter of fit between assessment score and program context, validity must be considered in relation to the assessment's purpose and the student's educational context. Certain assessment programs may require the participation of students with exceptionalities for claims to be valid. For instance, assessments for the purpose of system accountability may seek to include all students so that aggregate scores report on the functionality and effectiveness of province-wide curriculum as it pertains to all students learning from that curriculum. In such cases, additional measures may be considered to reduce negative consequences for individual students. For example, it may not be necessary to publicly report individual or school level results to limit identification of students and teachers and reduce social consequences. Assessment programs that serve more individualistic purposes, such as instructional diagnosis, may not require the inclusion of students with disabilities if the consequences outweigh the benefits of their participation. When considering the validity of assessment claims, it is critical to contextualize practices with respect to the purposes of the assessment program. As such, policymakers and educators must remain mindful that assessment scores are only meaningful when considered within program context.

Provision for Accommodations

Given the vast cognitive, behavioural, and physical differences among exceptional students, it is a challenge to conduct research applicable in large-scale contexts. As a result, few studies exist that examine the effectiveness of accommodations on the overall performance of special education students on standardized tests. Given what little is known about the impact of accommodations on the reliability and validity of assessment processes, provisions for accommodations in provincial and state-wide programs remains a controversial area (Shaftel, 2005). The six studies reviewed in this section (all of which are from the United States) have been selected because they provide examples of strong methodology for research into the effectiveness

of accommodations. Given the relatively recent history of large-scale assessments in Canada, research into this area is comparatively sparse.

One study that addressed provisions for accommodation was conducted by Tindal, Heath, Hollenbeck, Almond, and Harniss (1998). In this study, the researchers examined two types of accommodations for Grade 4 students with exceptionalities who wrote a Florida state-wide language and mathematics standardized test. The accommodations examined included a response accommodation (i.e., bubbling in responses versus writing out responses) and a presentation accommodation (i.e., reading the mathematics component aloud to students rather than having students read the test to themselves). Of the 481 participants in this study, 78 were identified as exceptional with 44 having an IEP relating to reading difficulties and an additional 20 having an IEP relating to mathematics ability. Tindal et al. argued that there was no significant advantage for the response accommodation. However, a significant difference was observed in students with reading and mathematics disabilities and low achieving general education students for the presentation accommodation on the mathematics assessment. Thus, the study supported the use of read aloud accommodation for students with disabilities in reading and mathematics.

Johnson, Kimball, Brown, and Anderson (2001) conducted a review of Washington's use of accommodations in their large-scale, high-stakes assessments. Generally, this assessment program attempts to include students with disabilities as much as possible by making provisions for accommodations. In particular, Johnson et al. investigated the performance of special student populations in grades 4 and 7 on the Washington Assessment for Student Learning (WASL). Four broad categories of accommodations are permitted for identified students on the WASL: (a) aids, such as native language dictionaries, physical supports, isolate portions of the test, and clarify direction; (b) scribe, such as answer orally, use voice recognition technology, and sign an answer; (c) large print or Braille format; and (d) oral presentation. Students were categorized by accommodation provision and identified as part of a special population category. Mean scores and standard deviations were calculated for each subgroup. In addition, scores for students who were read the math test questions and those who were provided a scribe were compared to an equivalent number of general education students. The data from this study supported five significant results. First, more accommodations were provided for fourth grade students than for seventh grade students. Second, special education students received the majority of accommodations provided to all students within the category of special populations (i.e., special education and English as a Second Language students). Third, the percentage of students provided accommodations is consistent with national expectations that 85% of special education students should not be given accommodations (data published by the National Center on Educational Outcomes). Fourth, special education students who received accommodations on average outperformed special education students who did not receive accommodations. Last, when compared with general education students, the effects of accommodations did not appear to provide an unfair advantage.

In a study of 115 fourth grade students, Johnson (2000) examined the effects on student performance of reading the mathematics items on the WASL to special education students. Students were categorized into three separate groups: (a) the control group (group A), (b) students without disabilities (group B), and (c) students receiving special education services for reading disabilities (group C). The students in group A and half of the students from group B were administered the test under standard conditions while the remainder of students in group B and those in group C were provided the accommodation in accordance with their IEPs and state guidelines. Data from groups A and B were analyzed through a repeated-measures analysis to determine if reading the mathematics items affected the validity of the test for general education students. Us-

ing the same procedure, the control group was compared with group C. In order to determine that the accommodation was in fact working to diminish a reading disability, and not just aiding poor readers, students in groups B and C were identified as "good readers" or "poor readers" based on their scores from the reading component of the WASL. A repeated-measures analysis of these two groups determined that neither "good" nor "poor" readers significantly benefited from the accommodation. This finding differs from the results of the Tindal et al. (1998) study previously described. Due to the lack of an interaction in the results of groups A and B, it appears that the accommodation did not have a significant effect for students without learning disabilities. The results also indicated that mean scores for students with learning disabilities benefited from the accommodation. This study supports the continued use of this accommodation for students with reading disabilities.

Research on accommodations in large-scale testing is challenging. Thus far, the research in this area presents rigorous effort; however, further research is required specifically from a Canadian perspective if accommodations are to be implemented with confidence in provincial assessment programs. Once again, research in this area should be conducted in reference to the assessment program's purpose because psychometric properties of the assessment will differ by program, possibly resulting in differing accommodation effects. While there has been much research on classroom-based accommodations, researchers who examine large-scale assessments are faced with the challenge of generalizing what has typically been seen as individualized levels of accommodation across a highly diverse population. Several limitations exist in this form of research, some of which are discussed in the subsequent section.

Limitations of Research

Research in the area of students with disabilities and large-scale assessments maintains several limitations as evident through the studies reviewed in this article. Three central limitations for research in this area are identified relating to issues of assessment reliability, inference validity, and characteristics of the special education field. From an assessment perspective, standardized testing for exceptional student populations poses an inherent dilemma for both the reliability of the testing scenario and the validity of the resulting knowledge claims. It is difficult and problematic to generalize knowledge claims regarding the appropriateness of accommodations because provisions for accommodations are highly specific to individual students' needs and may not be the same for students diagnosed with the same disability. For example, not all students diagnosed with dyslexia will require (or benefit from) a scribe in large-scale testing situations. In addition, the effectiveness of an accommodation is highly dependent on the context of its use and on the specifics of the assessment including its purpose, presentation format, response format, and content. As a result, researchers must be cautious when stating the effectiveness of an accommodation with respect to a particular disability or test construct.

As inclusion of students with disabilities is a relatively recent development in our educational system, the field of special education (as it relates to mainstream education) continues to evolve. This shift has coincided with increased use of large-scale assessments in many provinces and states. As such, both fields (i.e., special education and assessment) are in a response period characterized by efforts to bring together these two initiatives in education. Thus, the second limitation of research in this area involves consistency of terminology. This includes definitions and disability classifications as well as terminology relating to assessment. Within Canada there is a high degree of discrepancy from province to province with respect to disability classification

methods and related terms (Dworet & Bennet, 2002). As such, researchers may be working under different assumptions about student ability thereby compromising the applicability of findings across contexts. This notion also holds true for the terminology related to assessment and measurement. Researchers classify assessments based on various properties, and in some cases (e.g., high-stakes) definitions can have varying implications.

Last, accommodations are typically only provided to students identified as *disabled* and researchers tend to use this identification when establishing their subject groupings. The difficulty with this classification method is that it does not account for those students in the general education stream that remain unidentified. Having these students within the control or general education group will lower the mean score for these groups and diminish the effect size of the accommodation. This will result in an underestimate of accommodation effectiveness with significant differences potentially going unnoticed.

Conclusion

As inclusive policies and practices become dominant across Canada, educational systems are in the process of merging general and special education programming. This process is a challenging one especially when coupled with the move toward standards-based education. Issues inherent to this challenge are highlighted when including students with exceptionalities in large-scale assessment programs. Given what little is known about the effectiveness of accommodations and the consequences of participation in these programs, the concern that "the experience of the majority of special needs students is far from the ideal" (Lupart, 2001, p. 64) appears to be valid.

As large-scale assessment programs are increasing throughout Canada, issues of inclusion have started to occupy a more prominent role within the research agenda; however, research in this area is still exceedingly sparse. Research from the United States points to potentially high social, emotional, and academic consequences for students with disabilities included in large-scale assessment programs. This suggests that Canadian-based research is both critical and urgent if such negative consequences are to be reduced. As such, I outline the following program for research into the inclusion of students with disabilities in large-scale assessment programs. First, research is required that responds to issues of assessment reliability as it relates to nonstandardized assessment conditions (i.e., accommodation provisions). This research needs to evaluate the effectiveness of various accommodation strategies and consider their utility across exceptionality groups in large-scale contexts. Second, research is required that considers the multiple aspects of validity inherent to these assessment programs. Validity research in this context would seek to consider the positive and negative social, emotional, and academic consequences (i.e., consequential validity) of assessment practices (e.g., Ysseldyke et al., 2004) as well as the perception of score validity by educators and policymakers. Finally, research is needed that examines the practical processes of these assessment programs. Topics for research may include identification strategies of exceptional learners, teacher education models on accommodations, procedures for granting accommodations and exclusions, test preparation, and dissemination and use of test results.

Given that this is a relatively new area of research in Canada, policymakers and educators must remain cautious about the claims they generate from large-scale assessment processes. It is critical that scores are interpreted with respect to program purpose and educational context. Scores generated from large-scale assessments only offer one piece of the puzzle—students

demonstrate their ability in multiple ways which may not be represented on large-scale instruments. Large-scale assessments' results should be considered alongside other indicators of student achievement for valid inferences to be drawn about student ability. Only when inferences are validated, can instruction, curriculum, and educational policy reform move in a positive direction.

When considering the participation of students with disabilities in large-scale assessment programs, it is necessary to align decisions with the purpose of the assessment. Including students with disabilities may have significant and worthwhile gains for low-stakes assessment programs that are intended to evaluate the effectiveness of subject-specific programming and provincial curriculum—so long as these students are learning from the general education curriculum. On the other hand, standardized assessments that are linked to educational privileges and that publicly report student scores may have severe negative effects for students with disabilities. Several issues must be considered when deciding whether or not to include students with disabilities in large-scale assessments: (a) the value added (in relation to assessment purposes) from student inclusion, (b) the effects of accommodation provisions on inference validity and testing reliability, and (c) the intended and unintended consequences on test takers. Policymakers, educators, and assessment specialists must balance the educative value of including students with disabilities in large-scale testing with the potential negative consequence for students.

References

Bennett, S., Dworet, D., & Daigle, R. (2001). Educational provisions for exceptional students in the province of Ontario. *Exceptionality Education Canada, 11*(2 & 3), 55-70.

Berger, N., & Elson, H. H. (1996, April). *What happens when MCT's are used as an accountability device: Effects on teacher autonomy, cooperation and school mission.* Paper presented at the annual meeting of the American Educational Research Association, New York.

Brackenreed, D. (2004). Teachers' perceptions of the effects of testing accommodations. *Exceptionality Education Canada, 14*(1), 5-22.

Council for Exceptional Children. (2004). Policy on assessment and accountability. *Teaching Exceptional Children, 36*(4), 70-71.

Cox, M. L., Herner, J. G., Demczyk, M. J., & Nieberding, J. J. (2006). Provision of testing accommodations for students with disabilities on statewide assessments. *Remedial and Special Education, 27,* 346-354.

Darling-Hammond, L. (1994). Performance-based assessment and educational equity. *Harvard Educational Review, 64*(1), 5-30.

Destefano, L., Shriner, J. G., & Lloyd, C. A. (2001). Teacher decision making in participation of students with disabilities in large-scale assessment. *Exceptional Children, 68,* 7-22.

Dworet, D., & Bennet, S. (2002). A view from the north: Special education in Canada. *Teaching Exceptional Children, 34*(5), 22-27.

Earl, L. (1995). Assessment and accountability in education in Ontario. *Canadian Journal of Education, 20*(1), 45-55.

Eisenhart, M. (1998). On the subject of interpretive reviews. *Review of Educational Research, 68,* 391-399.

Froese-Germain, B. (2001). Standardized testing + high-stakes decisions = educational inequity. *Interchange, 32,* 111-130.

Haladyna, T. M., & Downing, S. M. (2004). Construct-irrelevant variance in high-stakes testing. *Educational Measurement: Issues and Practice, 23,* 17-27.

Hutchinson, N. L. (2007). *Inclusion of exceptional learners in Canadian schools: A practical handbook for teachers.* Toronto, ON: Prentice Hall.

Johnson, E. S. (2000). The effects of accommodations on performance assessments. *Remedial and Special Education, 21*, 261-267.

Johnson, E. S., Kimball, K., Brown, S. O., & Anderson, D. (2001). A statewide review of the use of accommodations in large-scale, high-stakes assessments. *Exceptional Children, 67*, 251-264.

Klinger, D., DeLuca, C., & Miller, T. (2008). The evolving culture of large-scale assessments in Canadian education. *Canadian Journal of Educational Administration and Policy, 76*, 1-34..

Langenfeld, K., Thurlow, M. L., & Scott, D. L. (1997). *High stakes testing for students: Unanswered questions and implications for students with disabilities.* Washington, DC: National Center on Educational Outcomes.

Lupart, J. L. (2001). Meeting the educational needs of exceptional learners in Alberta. *Exceptionality Education Canada, 11*(2 & 3), 55-70.

Madaus, G. (1988). The influence of testing on the curriculum. In L. Tanner (Ed.), *Critical issues in curriculum: 87th yearbook of the NSSE part 1* (pp. 83-121). Chicago, IL: University of Chicago Press.

McEwen, N. (1995). Accountability in education in Canada. *Canadian Journal of Education, 20*(1), 1-17.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). New York: Macmillan.

Schwandt, T. A. (1998). The interpretive review of educational matters: Is there any other kind? *Review of Educational Research, 68*, 409-412.

Shaftel, J. (2005). Improving assessment validity for students with disabilities in large-scale assessment programs. *Educational Assessment, 10*, 357-375.

Shepard, L. A., & Dougherty, K. (1991). *Will national tests improve student learning?* LA: University of California, Center for Research on Evaluation, Standards, and Student Testing.

Shriner, J. G., & Destefano, L. (2003). Participation and accommodation in state assessment: The role of individualized education programs. *Exceptional Children, 69*(2), 147-161.

Thurlow, M. L., & Johnson, D. R. (2000). High-stakes testing of students with disabilities. *Journal of Teacher Education, 51*(4), 305-314.

Tindal, G., Heath, B., Hollenbeck, K., Almond, P., & Harniss, M. (1998). Accommodating students with disabilities on large-scale tests: An experimental study. *Exceptional Children, 64*(4), 439-450.

Weber, K., & Bennett, S. (2004). Special education in Ontario schools (5th ed.). Palgrave, ON: Highland Press.

Ysseldyke, J., Dennison, A., & Nelson, R. (2004). *Large-scale assessment and accountability systems: Positive consequences for students with disabilities* (Synthesis Report 51). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Ysseldyke, J., Nelson, J. R., Christenson, S., Johnson, D. R., Dennison, A., Triezenber, H., et al. (2004). What we know and need to know about the consequences of high-stakes testing for students with disabilities. *Exceptional Children, 71*(1), 75-95.

Author's Note

Correspondence concerning this article should be addressed to Christopher DeLuca, Faculty of Education, Queen's University, Room A106, Kingston, ON, K7L 3N6.

E-mail: 2cd16@queensu.ca

Special thanks to Dr. Nancy L. Hutchinson for her support in reviewing this article.