The Alberta Journal of Educational Research *Vol. 54, No. 3, Fall 2008, 327-341*

John A. Ross

and

Peter Gray

Ontario Institute for Studies of Education of the University of Toronto

Alignment of Scores on Large-Scale Assessments and Report-Card Grades

We examined how much agreement there was between scores from large-scale mandated assessments and report-card grades for 14,776 students in grades 3, 6, and 9 of a district in which conditions were conducive to alignment of assessments. We found significant mean differences between internal and external assessments: effect sizes were .29 to .63 in grades 3-6 and .10 to.30 in grade 9. Spearman correlations were in the .32-.59 range. Chance-adjusted agreement was low. Report-card grades were consistently higher than external assessments for grades 3 and 6 students and consistently lower for grade 9 students.

Les auteurs ont étudié la concordance entre les notes obtenues sur des évaluations prescrites à grande échelle et celles sur les bulletins pour 14 776 élèves en 3e, 6e et 9e année dans un district présentant des conditions favorables à une comparaison des évaluations. Les auteurs ont trouvé des différences moyennes significatives entre les évaluations internes et les évaluations externes : les valeurs de l'effet étaient de 0,29 à 0,63 en 3e et 6e année et de 0,10 à 0,30 en 9e année. Les corrélations de Spearman se situaient entre 0,32 et 0,59. Le taux de concordance dûe au hasard était bas. Les notes des bulletins étaient plus élevées de façon uniforme que les notes des évaluations externes pour les élèves en 3e et en 6e, alors que pour les élèves en 9e, ces notes étaient plus basses de façon constante.

Introduction

In standards-based reform, policymakers develop measurable standards of student learning and use mandated assessments to hold students and teachers accountable for standards attainment (Hamilton, 2003; Resnick, Rothman, & Slattery, 2004). Such assessments are expected to increase achievement by providing focus in the form of targets for teachers and students that are more specific than the standards themselves. Mandated assessments provide motivation in the form of rewards and sanctions attached to assessment results. They also contribute to equity because all students have access to the same learning opportunities.

Whether assessment-driven reform contributes to higher achievement of curriculum standards is hotly contested. In this article we focus on a central assumption of the approach: We examine how much agreement there is be-

John Ross is a professor of curriculum, teaching, and learning and Head of the field centre in Peterborough. His research interests are school change, student assessment, and program evaluation.

Peter Gray was a doctoral student at the time this research was conducted. His research interests are in student assessment and models of school change.

tween scores on large-scale assessments and report-card grades in a setting conducive to alignment.

Literature Review

Definitions of Assessment Alignment

In the United States, assessment alignment has come to mean the extent to which external assessments (i.e., those not developed by classroom teachers for their own use) address the same content as state curriculum standards. This narrowing of the construct occurred because of extensive research on procedures for establishing agreement between standards and assessments (Porter, 2002; Resnick et al., 2004; Roach, Elliott, & Webb, 2005; Smithson & Porter, 2004), and because the *No Child Left Behind Act*, 2001, required that there be third-party verification that large-scale state assessments match state standards (Case, Jorgensen, & Zucker, 2004). In addition, a policy brief of the American Federation of Teachers presented data to show that alignment of state tests with state standards varies extensively across the US. The American Federation of Teachers (2006) report found that only 11 states had testing policies that could be considered highly aligned with curriculum standards.

However, researchers have addressed other dimensions of assessment alignment such as agreement of large-scale assessments with content delivered to students (Borko & Elliott, 1999; Datnow, 2001; Firestone, Winter, & Fitz, 2000), with learning theories (Pellegrino, Baxter, & Glaser, 1999), with teacher inservice (Borko, 1997), and with classroom assessments. In this study we define *assessment alignment* as the degree to which internal and external assessments provide equivalent information about student performance. By equivalent information we mean that the assessments correlate (i.e., that students are ranked in the same order by both assessments) and produce similar means (i.e., that one assessment does not produce scores that are consistently higher than the other).

Alignment of Classroom Tests and External Assessments

Researchers investigating the relationship between classroom tests and large-scale assessments cluster around two poles. At one pole are researchers who have shown that classroom tests deviate from principles of standardized assessment (Frey, Schmitt, Petersen, & Peyton, 2004; Marso & Pigge, 1991). From this perspective, classroom and large-scale assessments should be brought closer together by removing the flaws of teacher-made tests.

At the other pole are researchers who view classroom assessments positively, arguing that teacher-made assessments capture elements of student performance that are not measured in large-scale tests. For example, LeMahieu and Reilly (2004) identified differences between classroom and large-scale assessments in their conceptual frameworks, their form and substance, technical characteristics, and administration procedures. LeMahieu and Reilly argued that although distinct, the two assessment approaches could and should be more compatible. The key theme in these initiatives is the construction of developmental profiles and rubrics at an intermediate level of generality (more detailed than standards, less detailed than lesson objectives) as a blueprint that would tighten the connections among standards, large-scale assessments, and classroom testing (Forster & Masters, 2004; Frederickson & White, 2004; Wilson

& Draney, 2004). Although viewing classroom assessments more positively than researchers at the other pole, these researchers share the assumption that classroom tests and large-scale assessments should be highly correlated.

Existing research on the relationship between classroom tests and large-scale assessments can be criticized at three levels. First, many researchers (Forster & Masters, 2004; Frederickson & White, 2004; Frey et al., 2004; Marso & Pigge, 1991; Wilson & Draney, 2004) assume that tighter alignment of classroom and large-scale assessments is desirable. They share the assumption of standards-based reformers that student achievement will improve if the educational system becomes more tightly coupled. In fact Moss (2004) argued that aligning internal and external assessments would reduce the role of teachers to followers of templates designed by others. Moss argued that tighter alignment would increase external control over classrooms and reduce diversity and teachers' attention to the social context of learning. The second critique is that earlier researchers have not examined how much agreement there is between classroom assessments and large-scale assessments. These researchers assume that because there are substantive differences in purposes, psychometric characteristics, and procedures, there will be differences in the results so that consumers of the two assessment systems such as parents and guardians will receive discrepant information. The third criticism is that by focusing on individual classroom items and tests, these researchers ignore the important difference between classroom test scores and report-card grades. For example, in Ontario (but not necessarily in other provinces), high-stakes decisions are based on report-card grades, not on individual teacher-made tests. It may be that the deficiencies of individual tests are diluted when the results of tests are combined with other forms of assessment and aggregated to the report-card level.

Scores on Large-Scale Assessments and Report-card Grades

Comparing large-scale assessment scores and report-card grades requires that both measure the same curriculum standards and use the same metric. In such conditions, we would expect students to receive scores on mandated assessments administered by independent agencies to be similar to grades awarded by their teachers. Exact agreement is unlikely because the assessments differ in important ways. For example, report-card grades are based on a large number of independent appraisals across a whole year and typically include a wide variety of techniques and a broad sample of learning objectives derived from the standards. External assessments are usually administered in a single time block, using a narrow range of methods (e.g., Stecher, 1998, found that portfolios are rarely included in external assessments), and address a smaller sample of objectives.

Few studies have examined the relationship between large-scale assessments and report-card grades. Brennan, Kim, Wenz-Gross, and Siperstien (2002) compared scores on the Massachusetts Comprehensive Assessment System (MCAS) with teacher-assigned grades in mathematics, English, and science for 736 grade 8 students. They found that correlations within subjects were much higher than correlations between subjects. Within-subject correlations of MCAS scores and teacher grades ranged from .54 to .60. Brennan et al. also found that classroom grades and state tests were not equally well corre-

lated for all students. In mathematics and science, girls scored higher than boys on classroom grades, but girls received lower MCAS scores than boys.

Willingham, Polack, and Lewis (2002) compared National Education Longitudinal Study (NELS) tests and teacher-assigned grades for 8,454 high school seniors. They reported a correlation of grades and standardized tests of .62. The most important influences on the degree of assessment alignment were similarity of subjects tested, scholastic engagement of students, and teachers' ratings of student behavior. Reliability of tests/grades and the grading metric also had an effect. When all five factors were controlled, the correlation of grades and external tests increased to .90. The effect of the five factors was similar with each student subgroup.

A recent study from British Columbia included large-scale assessment scores and report-card grades. Although the focus of the study addressed issues unrelated to our research, Lloyd, Walsh, and Yailagh (2005; J. Walsh, personal communication, February 12, 2006) found that the two assessments correlated positively: r=.69 for 161 grades 4 and 7 students.

These studies provide a helpful starting point. However, each study was limited to a single grade (two for Lloyd et al., 2005) and a single year of data; none examined teacher characteristics or school policies as possible moderators of assessment alignment; and each limited its definition of alignment to correlations. In our study we examined large-scale assessment scores and report-card grades for students in grades 3, 6, and 9 in reading, writing, and mathematics over two years; and we defined alignment in two ways: in terms of difficulty levels and in terms of correlations. Our research question was: How well aligned are large-scale assessments and report-card grades? That is, do students receive similar proficiency ratings and do scores and grades correlate significantly?

Method

Setting

We selected one school district in the province of Ontario because that province provides many of the conditions conducive to alignment of report-card grades and large-scale assessments. The large-scale assessments were conducted by the Education Quality and Accountability Office, an agency charged with the measurement of student outcomes within strict requirements established by the Ministry of Education and Training. The Ministry also specified strict conditions for the report card, stating, "there should be no changes of any kind made to the three pages of the Provincial Report Card" (Ontario Ministry of Education and Training, 2000a, p. 4).

Both assessments were based on specific expectations spelled out in provincial curriculum standards. The electronic version of the report card contained a curriculum browser that enabled elementary teachers to view curriculum expectations by subject, strand, and grade while filling out the report card; the browser also contained the achievement level descriptors for all subjects. Secondary teachers were expected to consult hard copies of the standards for their grade and subject. The Educational Quality and Accountability Office (EQAO, 2006a, 2006b) produced detailed tables showing how its items mapped onto specific curriculum standards—the items themselves were not published. Provincial curriculum standards specified levels of proficiency (several catego-

ries denoting performance below level one and levels one through four). Curriculum documents defined what each level meant in general and in the context of specific grades, subjects, and category of learning objective. The EQAO reported performance in terms of these levels. Report cards provided letter grades (for grades 3 and 6) and percentages (for grade 9). There was a formula (shown on page 12) for converting from one metric to the other.

In many settings, report-card grades, unlike large-scale assessments, mix together academic achievement with judgments of student work habits, contributing to error variance (Willingham et al., 2002). Ontario report cards provided separate sections for learning skills such as class participation, cooperation with others, and completing homework (the lists varied slightly for elementary and secondary students) and for punctuality and attendance. Ministry policy stated that achievement judgments should be completely separate from measurement of other attributes. "The assessment and evaluation of learning skills is distinct from and should not influence the determination of percentage grades" (Ontario Ministry of Education and Training, 2000b; compare 2000a).

The Ministry and the district published rubrics, scoring guides, and exemplars used by EQAO markers (EQAO, n.d.) and provided training to teachers on their use. These materials were presented as guides for teacher marking. Training sessions frequently included moderated marking; that is, teachers used EQAO criteria to assess the work of their own students and received feedback on their assessments. Such training increases the reliability of assessment (Schafer, Swanson, Bené, & Newberry, 2001; Shepard et al., 1996).

The Ministry, the EQAO, and the district made considerable efforts to align report-card grades with large-scale assessments, especially at the level of content coverage. However, some curriculum standards were measured only by the classroom teacher (e.g., speaking skills: EQAO, 2006a) and report cards sampled a much larger proportion of the curriculum. On balance, the setting of this study supported the expectation that the assessments would be aligned.

Sample

We examined large-scale assessments and teacher-assigned grades for students in 80 elementary and 15 secondary schools in one school district in Ontario. The external assessments, administered in May, were compared with the nearest preceding report card period, March.

We were able to match EQAO and report-card data for 74% of the students in the district. Because of student mobility, our sample sizes were lower for 2002 (grade 3=2,066; grade 6=2,212; grade 9=2,450) than for 2003 (grade 3=2,633; grade 6=3,011; grade 9=2,404). These numbers are for students for whom we had at least one EQAO score and one report-card grade, not complete datasets. The sample sizes of the individual analyses were smaller (shown in the Results section) because of missing values, especially for mathematics (district policy did not require that teachers report a score for all strands in the March reporting period). We did not replace missing values because the sample size is sensitive to small differences (i.e., we were able to detect a difference between the assessments as small as *ES*=.08, with 80% power and 5% type I error; see Dennis, 1994).

Instruments

Grades 3 and 6 Mathematics. The EQAO reported 10 scores based on 28 multiple-choice and six open-ended items (the latter were performance assessments) for each grade. We used six scores: five mathematical strands (Number sense & Numeration, Measurement, Geometry & Spatial Sense, Patterning & Algebra, Data Management & Probability) and a global score. We did not use the other four EQAO scores (four dimensions of problem-solving) because there were no comparable report-card grades. The EQAO reported performance in terms of proficiency levels; that is, a five-point scale consisting of levels 0-4 (we converted several infrequently used categories that indicate less than level 1 to level 0). Wolfe, Childs, and Elgie (2004) reported that EQAO's procedures for equating raw scores to proficiency levels varied by test and over time. For grades 3 and 6 classical test theory was used (80% constructed-response items + 20% multiple-choice); for grade 9 a three-parameter IRT procedure was used.

The comparable report-card grades were the five mathematical strands and a global mathematics score consisting of the mean of these five strands. Letter grades used in report cards were transformed into 0-4 levels, using the provincial formula (Ontario Ministry of Education and Training, 1999, 2000a, 2000b).

- R [need for remediation] = < 50% = < Level 1 (Level 0)
- D– to D+ = 50-59% = Level 1
- C– to C+ = 60-69% = Level 2
- B– to B+ = 70-79% = Level 3
- A– to A+ = 80-100% = Level 4

Grades 3 and 6 Language. The EQAO reported four scores for reading (based on 32 multiple-choice and 12 open-ended items), four scores for writing (based on eight multiple-choice and three open-ended items), and a global score for each subject—we used only the global scores. The report card provided reading, writing, and oral communication grades. We used only reading and writing, after transforming report-card letter grades into levels 0-4 using the conversion formula.

Grade 9 mathematics. The EQAO tested only mathematics in grade 9. The EQAO reported a global score for two courses, applied and academic mathematics based on 24 multiple-choice and three open-ended items for each. The EQAO also reported seven separate scores for mathematical strands and problem-solving dimensions that we did not use because the report card provided only one score for each course. We compared global scores for applied and academic courses after converting report-card grades to levels 0-4.

Reliability of Educational Quality and Accountability Office Scores and Report Cards. The EQAO used three procedures to ensure reliability: (a) group marking: all markers scored the same student, and the results were photocopied for discussion; (b) reinsertion, that is, a sample of papers was scored by two or more markers; (c) if a marker was in the top or bottom 5% for levels awarded on a given day, that person's output was re-marked to guard against lenience/severity differences. The EQAO does not report reliability coefficients. However, in the 2003 administration, a reliability study found that markers did not substantially influence scores: generalizability coefficients averaged .76 (Dunn, Childs, Cleland, Pang, & Saunders, 2004).

Analysis

We tested the representativeness of the sample by using chi-square tests to determine if the proportion of students who achieved the provincial standard (level 3 or 4) was the same in the sample as in the district. We made 14 sample-population comparisons (elementary=2 years x 2 grades x 3 subjects + secondary=2 courses).

We tested agreement between large-scale assessments and report-card grades by comparing means and correlations. To compare means, we conducted a multivariate, repeated-measures General Linear Model (GLM) with assessment type (report card vs. EQAO) as the independent variable and reading, writing, and overall mathematics scores as the dependent variables, beginning with the 2002 grade 3 data. We repeated the analysis using the grade 6 data for 2002, grade 3 for 2003, and grade 6 for 2003. For the grade 9 data, we conducted two paired-sample *t*-tests (one for students in applied mathematics and one for students in academic mathematics) on the EQAO mathematics scores and report-card grades, testing the 2002 data and repeating the analysis for the 2003 data. This analysis enabled us to determine whether the mean-level placements that students received were the same or significantly different for the two assessments.

The second analysis used Spearman's correlation (because the data were nonparametric) to determine whether students were ranked in the same order by the two assessments. We calculated the percentage of perfect agreement between the assessments to determine whether students were placed in the same proficiency level by both. We used Kappa coefficients to represent chance-adjusted agreement between the assessments. We began the analysis with the 2002 grade 3 data for reading, writing, mathematics, and the five mathematical strands. We repeated the analysis for the other datasets: 2002 grades 6 and 9, 2003 grades 3, 6, and 9.

Results

Representativeness of Sample

We used chi-square tests to determine if the proportion of students who achieved the provincial standard (level 3 or 4) in the district was the same proportion as in our sample. Comparisons were made for each subject in each grade for each year. Of the 14 sample-population comparisons, only one resulted in significant differences between the sample and the population (2002 grade 3 writing: χ^2=4.151, *df*=1, *p*=.042). This result provided support for the claim that the sample represented the district population.

Research Question: How well aligned are large-scale assessments and report-card grades?

We began with the 2002 grade 3 data. A multivariate, repeated measures GLM was conducted with assessment type (report card versus EQAO) as the independent variable; reading, writing, and mathematics scores were the dependent variables. An effect for assessment type was found ($F(3,1500)=177.215$, $p<.001$): the two assessments produced significantly different mean scores. Individual repeated-measures GLMs were then performed for each of the three overall scores and the five mathematical strands. The results, displayed in Table 1, revealed statistically significant differences for each comparison except

Table 1
Comparison of Assessment Means for 2002 Grade 3 Students by Subject

Subject	EQAO Mean	EQAO SD	Report Card Mean	Report Card SD	N	GLM	ES
Reading	2.59	0.74	2.94	0.76	1,558	F (1,1557)=339.49, p<.001	0.46
Writing	2.69	0.66	2.70	0.73	1,663	F (1,1662)=0.15, p=.695	ns
Math	2.72	0.73	3.06	0.67	1,647	F (1,1646)=375.39, p<.001	0.49
Number Sense and Numeration	2.59	0.96	2.97	0.79	1,714	F (1,1713)=278.20, p<.001	0.43
Measurement	2.45	0.93	2.99	0.77	1,405	F (1,1404)=454.40, p<.001	0.63
Geometry and Spatial Sense	2.73	0.94	3.04	0.68	1,547	F (1,1546)=156.13, p<.001	0.37
Patterning and Algebra	2.66	0.84	2.90	0.72	1,399	F (1,1398)= 88.85, p<.001	0.30
Data Management and Probability	2.64	0.96	2.89	0.71	968	F (1,967) = 64.60, p<.001	0.29

writing. In all comparisons except writing, the mean report-card scores were higher than EQAO scores.

The same analysis was conducted on the 2002 grade 6 data. As with the grade 3 results, there was a statistically significant multivariate effect for assessment type (F(3,1738)=189.836, p<.001). The individual repeated-measures GLMs, displayed in Table 2, indicated that the means were significantly higher for report cards than EQAO on all subjects.

The analyses were repeated using the 2003 data for grades 3 and 6. The results were the same as for the 2002 analysis. The multivariate GLM showed statistically significant differences between the means of the report-card grades and EQAO scores for grade 3 (F(3,2324)=221.34, p<.001) and grade 6 (F(3,2715)=281.20, p<.001). The univariate analyses (not shown) indicated that in every case, report-card grades were significantly higher than scores on the large-scale assessments. The effect sizes were ES=.14–.64 (grade 3) and .31–.56 (grade 6).

In summary, when we compared the two assessments in 36 elementary school contexts (2 grades x 9 subjects / topics x 2 years), report-card grades were higher than EQAO assessments in every case. The differences were statistically significant in 35 of the 36 comparisons.

For the 2002 grade 9 results, we conducted two paired-sample t-tests (one for students in applied mathematics and one for students in academic mathematics) to compare EQAO mathematics scores with report-card marks. There was a significant difference for each course: applied: t(888)=8.653, p<.001, ES=.30; academic: t(1560)=4.185, p<.001, ES=.10. For both grade 9 courses, the report-card mean was lower than the EQAO mean (the opposite of the findings in grades 3 and 6). We repeated the analysis for the 2003 grade 9 data. The results were virtually identical to 2002: applied: t(805)=6.519, p<.001, ES=.23; academic: t(1597)=5.941, p<.001, ES=.14, with report-card means lower than EQAO means.

These comparisons showed that mean proficiency levels assigned to students were not the same on large-scale assessments and report-card grades.

Table 2
Comparison of Assessment Means for 2002 Grade 6 Students by Subject

Subject	EQAO Mean	EQAO SD	Report Card Mean	Report Card SD	N	GLM	ES
Reading	2.59	0.76	2.90	0.79	1,814	$F(1,1813)=296.73$, $p<.001$	0.40
Writing	2.57	0.78	2.78	0.77	1,873	$F(1,1872)=149.02$, $p<.001$	0.27
Math	2.53	0.85	2.96	0.82	1,831	$F(1,1830)=509.96$, $p<.001$	0.52
Number Sense and Numeration	2.60	0.91	2.89	0.96	1,910	$F(1,1909)=160.92$, $p<.001$	0.31
Measurement	2.45	0.89	2.79	1.01	1,494	$F(1,1493)=166.48$, $p<.001$	0.35
Geometry and Spatial Sense	2.65	0.83	2.94	0.91	1,475	$F(1,1474)=124.25$, $p<.001$	0.33
Patterning and Algebra	2.55	0.78	2.91	0.90	1,475	$F(1,1474)=211.33$, $p<.001$	0.43
Data Management and Probability	2.52	0.88	2.86	0.90	1,347	$F(1,1346)=160.47$, $p<.001$	0.39

Teachers of grades 3 and 6 gave significantly higher scores than those awarded by EQAO whereas teachers in grade 9 did the opposite: students received lower scores on their report cards than on external assessments. We explored this finding further by examining correlations and levels of agreement between the two assessments.

Table 3 shows the correlations between the two assessment types for grade 3 students in 2002. We used Spearman's correlation because the data were nonparametric. All the correlations were significant ($p<.001$), although lower than the $r=.64$ to .69 reported in earlier studies (Brennan et al., 2002; Lloyd et al., 2005; Willingham et al., 2002).

Table 3 also reports the agreement between the assessment in terms of Cohen's (1988) Kappa, a chance-adjusted coefficient of comparison. There was statistically significant agreement between the two assessments for all subjects and topics. The coefficients were low. For example, in the first row of the table, K=.15 means that once chance agreements had been taken into account, the assessments matched on only 15% of the comparisons. None of the Kappas reached even the "fair" level of agreement defined by Bakeman and Gottman (1997) as .40.

Table 3 indicates that there was perfect agreement between the two assessments for a substantial proportion of students (38%-58% depending on the subject). Most disagreements involved a single level—90-95% of the comparisons were within one level of each other. However, 52% of all comparisons on a five-point scale will be within one level through chance alone. When there was a discrepancy in the assessments, report-card grades were more likely to be one level higher than lower than EQAO scores.

The results for grade 6 in 2002 and for grades 3 and 6 in 2003 (not displayed) were virtually identical. All comparisons showed significant but low Kappa results and significant Spearman correlations that ranged from .39 to .55. These results indicate that in the elementary panel, in both years, EQAO and classroom assessments were in moderate agreement.

Table 3
Correlations of Assessment Types for 2002 Grade 3 Students by Subject

				Perfect Level of Agreement in Percentages EQAO Score—Report Card Score								
Subject	*N*	r_s	*K*	*–4*	*–3*	*–2*	*–1*	*agree*	*+1*	*+2*	*+3*	*+4*
Reading	1,558	0.49	0.15	0.0	0.0	4.3	38.1	46.5	10.5	0.6	0.0	0.0
Writing	1,663	0.52	0.29	0.0	0.0	0.8	20.7	57.7	19.8	0.9	0.0	0.0
Math	1,647	0.47	0.17	0.0	0.3	3.3	36.7	49.7	9.8	0.1	0.0	0.0
Number Sense and Numeration	1,714	0.44	0.15	0.4	1.7	8.1	32.8	41.3	14.5	1.3	0.0	0.0
Measurement	1,405	0.40	0.10	0.4	1.7	10.5	38.1	37.5	11.2	0.6	0.1	0.0
Geometry and Spatial Sense	1,547	0.32	0.10	0.6	1.7	6.2	30.4	42.0	18.4	0.7	0.0	0.0
Patterning and Algebra	1,399	0.32	0.11	0.3	1.5	5.1	28.6	44.8	18.2	1.5	0.0	0.0
Data Management and Probability	968	0.37	0.15	0.2	1.3	6.6	27.5	44.7	18.1	1.4	0.1	0.0

Note. All r_s (Spearman rank correlation): $p<.001$; all Kappa: $p<.001$.

Table 4 shows the results for the grade 9 mathematics scores. The correlations between report cards and EQAO assessments were large and statistically significant for both courses. Kappa coefficients were low, but there was statistically significant agreement between the two assessments. The proportions of exact agreement were lower than was the case for grades 3 and 6. Agreement within one level was also lower, but still in the 82-88% range. Agreement of report cards and EQAO was higher for academic than applied courses. Discrepancies between assessments for students in grade 9 applied courses typically consisted of EQAO scores being one or even two levels higher than report-card grades. The pattern for the grade 9 academic courses was similar: students were likely to score higher on EQAO than on report cards, although the size of the effect was smaller.

Discussion

In this study conditions in the school district were conducive to the alignment of report-card grades and external assessments: provincial curriculum standards and proficiency levels drove both assessments; teachers were trained to use the same rubrics as EQAO markers; an electronic browser encouraged teachers to link their assessments to the standards; EQAO produced maps linking its items to the standards; report cards provided separate sections for student behaviors (e.g., homework completion) that might distort grades assigned for academic work. Yet the results indicated that EQAO assessments and report cards produced scores that were significantly different. The finding was robust across grades and subjects. The effect size of the differences (.29 to .63 in grades 3-6; .10 to .30 in grade 9) could be described as small to medium based on Cohen's (1988) rules or on Lipsey and Wilson's (1993) meta-analysis of 300 studies of psychological, behavioral, and educational interventions that found a median effect size of .50. By another standard, Borman, Hewes, Over-

Table 4
Correlations of Assessment Types for Grade 9 Mathematics Students by Course and by Year

				Perfect Level of Agreement in Percentages EQAO Score—Report Card Score								
Subject	*N*	r_s	*K*	*–4*	*–3*	*–2*	*–1*	*agree*	*+1*	*+2*	*+3*	*+4*
Applied 2002	889	0.50	0.14	0.2	0.3	4.6	16.0	33.6	32.4	12.3	0.6	0.0
Academic 2002	1,561	0.58	0.20	0.2	0.3	1.7	27.4	38.8	20.9	9.4	1.3	0.0
Applied 2003	806	0.52	0.15	0.4	0.7	3.8	17.2	35.2	32.1	10.2	0.2	0.0
Academic 2003	1,598	0.59	0.15	0.1	0.1	1.4	29.0	33.8	25.0	9.1	1.5	0.0

Note. All r_s (Spearman rank correlation): $p<.001$; all Kappa: $p<.001$.

man, and Brown (2003) in a meta-analysis of over 232 studies found a mean effect size of .15 for whole-school reform initiatives. Given that EQAO scores represent the main source of data about school achievement in Ontario (Johnson, 2005) and report cards are the primary source of parents' information about their children's success, the differences between the assessments found in this study should be viewed as meaningful.

One likely explanation for the difference between report cards and large-scale assessments is that teachers consider many sources of evidence about achievement that are not reflected in achievement tests. For example, many teachers draw on portfolio assessment as an important component of report-card grades, even though provincial assessments do not include it. In addition, report cards include greater curriculum coverage. For example, EQAO assessments do not address speaking skills, an important part of the grade 3 literacy curriculum.

The correlations between EQAO and classroom assessments were slightly lower (.32-.59) than the .50s and .60s reported in earlier studies (Brennan et al., 2002; Lloyd et al., 2005; Willingham et al., 2002). One explanation might be that our data required the use of a nonparametric statistic (which produces lower correlations than Pearson correlations).

We found that report-card grades tended to be higher than EQAO scores for grades 3 and 6 and lower for grade 9. A plausible explanation for the reversal in direction of the EQAO-report-card score differences might be the motivational strategies prevailing in the two panels. We speculate that in the elementary panel, the key concern of teachers is to develop students' confidence, to create willing readers, energetic writers, and fearless mathematicians. Students' belief in their ability to be successful in academic settings contributes to enhanced goal-setting, effort, persistence, and achievement (Pajares, 1996). Teachers committed to the enhancement of student confidence create tasks that can be successfully completed by willing students of average ability. Such teachers also provide safety nets for students, balancing difficult tasks requiring deep understanding with easier items requiring only procedural application.

Secondary mathematics teachers recognize the importance of student confidence, but they are also committed to defending the discipline, for example,

by ensuring that students who are not ready for the next level of mathematics do not pass on to the next grade. The latter concern might result in demanding classroom assessments. We suspect that even when a classroom assessment proves to be overly difficult, its result is still included in the calculation of the final grade, if this assessment measures an important disciplinary objective.

An important limitation of our study was our inability to link scores on performance tasks in mathematics (i.e., the five dimensions of problem-solving reported by EQAO) to specific report-card indicators. However, performance tasks contribute to report-card grades according to provincial assessment policy. The problem-solving scores were embedded in the global EQAO scores and the total report-card grades.

Conclusion

The purpose of external assessment is to persuade teachers to implement curriculum standards in their classrooms (i.e., teach to the test in a positive sense) and to regulate the implementation of educational policy (McDonnell, 1994). Persuasion, regulation, and fairness are heightened when curriculum and assessment practices are tied to the same student expectations (American Educational Research Association, 2000; English & Steffy, 2001; Wolf, Borko, Elliott, & McIver, 2000), conditions that prevailed in the setting of our research.

Our contribution to this debate is twofold. First, we found that despite conducting the study in a district where conditions were conducive to alignment of internal and external assessments, agreement was at best moderate. There were statistically significant differences between report cards and mandated assessments in all but one of the 40 grade-subject-year comparisons of means. The correlations of the two assessments were modest (in the .40s and .50s), proportions of perfect agreement on a 0-4 scale were below 50%, and chance-adjusted agreement was low (Kappa coefficients in the .10-.20 range). A key factor accounting for moderate agreement might be the number of items required to cover the curriculum compared with the number of items included in EQAO assessments. Students worked on EQAO assessments for half a day for a five-day period. Because the tests were not speeded, students were permitted to take longer. But they completed a relatively small number of items, for example, 28 multiple-choice and six open-ended items for grades 3/6 mathematics. Most of the time was spent on the open-ended items, which in the case of grade 3/6 mathematics consisted of two investigations. Linn (1994) found that for the reliabilities of performance tasks reported in the literature, 10-23 instances of each task would be required to ensure generalizability from the test to the curriculum standard. Although the number of items could easily be increased by replacing investigations with multiple-choice items, the curriculum standards emphasize deep understanding that is best measured with complex tasks in which students are required to explain their reasoning. In addition, performance on one set of standards does not readily generalize to another. Agreement of report-card grades with large-scale assessments may be constrained by the inability of the external assessments to cover the curriculum to the degree that is possible in classroom assessments.

The second contribution of our study was the finding of the moderating effect of grade on assessment alignment, that is, we found that report-card grades were higher than EQAO assessments in grades 3 and 6 in all subjects

and in both years, but that the pattern reversed in grade 9. We attributed the reversal to differences in the motivational theories of teachers in the two panels.

We encourage future researchers to investigate issues of assessment alignment, focusing not only on the alignment of large-scale assessments with curriculum standards as required by NCLB, but on other forms of alignment that impinge on student achievement. We particularly recommend three directions. First, we need to find out whether the findings consistently found in our research environment will replicate to settings that are less conducive to the alignment of report-card grades and external assessments. Second, we need to know more about the characteristics of schools with high- and low-assessment alignment. No study has investigated or even formulated a theory of how school characteristics such as leadership practices, capacity beliefs, relationships with home communities, and SES status might be related to assessment alignment. Third, we need to test our speculation about the effect of elementary and secondary teachers' motivational beliefs on report-card grades. Pursuit of this research agenda will move the debate about the alignment of large-scale assessments and report cards to a consideration of the conditions under which high and low alignment occurs and its effect on student performance.

Acknowledgments

An earlier draft of this article was presented at the joint meeting of the American Evaluation Association and Canadian Evaluation Society, Toronto, October 2005. The research was funded by the Education Quality and Accountability Office of Ontario. The views expressed in the report do not necessarily reflect the views of EQAO. Cecil Knight compiled the EQAO and report-card databases used in the study.

References

American Educational Research Association. (2000). Position statement of the American Educational Research Association concerning high-stakes testing in preK-12 education. *Educational Researcher, 29*(8), 24-25.

American Federation of Teachers. (2006). *Smart testing: Let's get it right: How assessment-savvy have states become since NCLB?* Retrieved February 20, 2007, from: http://www.aft.org/presscenter/releases/2006/smarttesting/Testingbrief.pdf

Bakeman, R., & Gottman, J. (1997). *Observing interaction: An introduction to sequential analysis*. Cambridge, UK: Cambridge University Press.

Borko, H. (1997). New forms of classroom assessment: Implications for staff development. *Theory into Practice, 36*, 231-238.

Borko, H., & Elliott, R. (1999). Hands-on pedagogy versus hands-off accountability: Tensions between competing commitments for exemplary math teachers in Kentucky. *Phi Delta Kappan, 80*, 394-400.

Borman, G.D., Hewes, G.M., Overman, L.T., & Brown, S. (2003). Comprehensive school reform and achievement: A meta-analysis. *Review of Educational Research, 73*(2), 125-230.

Brennan, R., Kim, J., Wenz-Gross, M., & Siperstein, G. (2002). The relative equitability of high stakes testing versus teacher-assigned grades: An analysis of the Massachusetts Comprehensive Assessment System (MCAS). *Harvard Educational Review, 71*, 173-216.

Case, B.J., Jorgensen, M.A., & Zucker, S. (2004). *Alignment in educational assessment*. Retrieved February 20, 2006, from: http://harcourtassessment.com/hai/Images/pdf/assessmentReports/AlignEdAss.pdf

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Datnow, A. (2001, April). *The sustainability of externally developed school reforms in changing district and state contexts*. Paper presented at the annual meeting of the American Educational Research Association, Seattle.

Dennis, M. (1994). Ethical and practical randomized field experiments. In J. Wholey, H. Hatry, & K. Newcomer (Eds.), *Handbook of practical evaluation* (pp. 155-197). San Francisco, CA: Jossey-Bass.

Dunn, J., Childs, R., Cleland, P., Pang, X., & Saunders, K. (2004, April). *Isolating item and rater variance in the grade 3 assessment of reading*. Paper presented at the annual meeting of the American Educational Research Association, San Diego.

Educational Quality and Accountability Office. (n.d.). *Scoring guides and anchors, 2002-03: Grade 3 and grade 6 assessment of reading, writing and mathematics*. Retrieved March 2, 2006, from: http://www.eqao.com/Educators/Elementary/036/036_7.aspx?Lang=E&gr=036&Aud=Educators&yr=03&gb=true

Educational Quality and Accountability Office. (2006a). *Grade 3 assessment of reading, writing, and mathematics: Framework*. Retrieved February 22, 2006, from: http://www.eqao.com/pdf_e/06/06P008e.pdf

Educational Quality and Accountability Office. (2006b). *Grade 6 assessment of reading, writing, and mathematics: Framework*. Retrieved February 22, 2006, from: http://www.eqao.com/pdf_e/06/06P008e.pdf

English, F.W., & Steffy, B.E. (2001). *Deep curriculum alignment. Creating a level playing field for all children on high-stakes tests of educational accountability*. Lanham, MD: Scarecrow Press.

Firestone, W., Winter, J., & Fitz, J. (2000). Different assessments, common practice? Mathematics testing and teaching in the USA and England and Wales. *Assessment in Education, 7*(1), 13-37.

Forster, M., & Masters, G. (2004). Bridging the conceptual gap between classroom assessment and system accountability. In M. Wilson (Ed.), *Towards coherence between classroom assessment and accountability. 103rd Yearbook of the National Society for the Study of Education* (Part II, pp. 51-73). Chicago, IL: University of Chicago Press.

Frederickson, J.R., & White, B.Y. (2004). Designing assessments for instruction and accountability: An application of validity theory to assessing scientific inquiry. In M. Wilson (Ed.), *Towards coherence between classroom assessment and accountability. 103rd Yearbook of the National Society for the Study of Education* (Part II, pp. 74-104). Chicago, IL: University of Chicago Press.

Frey, B.B., Schmitt, V.L., Petersen, S.E., & Peyton, V.D. (2004, April). *Validity of teacher-made tests*. Paper presented at the annual meeting of the American Educational Research Association, San Diego.

Hamilton, L. (2003). Assessment as a policy tool. In R. Floden (Ed.), *Review of research in education* (Vol. 27, pp. 25-68). Washington, DC: American Educational Research Association.

Johnson, D.R. (2005). *Signposts of success: Interpreting Ontario's elementary school test scores*. Toronto, ON: C.D. Howe Institute.

LeMahieu, P.G., & Reilly, E.C. (2004). Systems of coherence and resonance: Assessment for education and assessment of education. In M. Wilson (Ed.), *Towards coherence between classroom assessment and accountability. 103rd Yearbook of the National Society for the Study of Education* (Part II, pp. 189-202). Chicago, IL: University of Chicago Press.

Linn, R. (1994). Performance assessment: Policy, promises and technical measurement standards. *Educational Researcher, 23*(9), 4-14.

Lipsey, M.W., & Wilson, D.B. (1993). The efficacy of psychological, educational, and behavioral treatment: Confirmation from meta-analysis. *American Psychologist, 48*(12), 1181-1209.

Lloyd, J.E.V., Walsh, J., & Yailagh, M. S. (2005). Sex, differences in performance attributions, self-efficacy, an achievement in mathematics: If I'm so smart, why don't I know it? *Canadian Journal of Education, 28*, 384-408.

Marso, R., & Pigge, F. (1991). An analysis of teacher-made tests: Testing practices, cognitive demands and item construction errors. *Contemporary Educational Psychology, 16*, 279-286.

McDonnell, L.M. (1994). Assessment policy as persuasion and regulation. *American Journal of Education, 102*, 394-420.

Moss, P. (2004). The risks of coherence. In M. Wilson (Ed.), *Towards coherence between classroom assessment and accountability. 103rd Yearbook of the National Society for the Study of Education* (Part II, pp. 217-238). Chicago, IL: University of Chicago Press.

Ontario Ministry of Education and Training. (1999). *Guide to the provincial report card, grades 9-12*. Retrieved February 22, 2006, from: http://www.edu.gov.on.ca/eng/document/forms/report/sec/srepgde.pdf

Ontario Ministry of Education and Training. (2000a). *Guide to the provincial report card, grades 1-8*. Retrieved February 22, 2006, from: http://www.edu.gov.on.ca/eng/document/forms/report/1998/repgde.pdf

Ontario Ministry of Education and Training. (2000b). *The Ontario curriculum grades 9-12: Program planning and assessment*. Toronto, ON: Author.

Pajares, F. (1996). Self-efficacy beliefs in academic settings. *Review of Educational Research, 66*, 543-578.

Pellegrino, J.W., Baxter, G.P., & Glaser, R. (1999). Addressing the "two disciplines" problem: Linking theories of cognition and learning with assessment and instructional practice. In A. Iran-Nejad & P.D. Pearson (Eds.), *Review of research in education* (vol. 24, pp. 307-353). Washington, DC: American Educational Research Association.

Porter, A. (2002). Measuring the content of instruction: Uses in research and practice. *Educational Researcher, 31*(7), 3-14.

Resnick, L.B., Rothman, R., & Slattery, J.B. (2004). Benchmarking and alignment of standards and testing. *Educational Assessment, 9*(1/2), 1-28.

Roach, A.T., Elliott, S.N., & Webb, N.L. (2005). Alignment of an alternate assessment with state academic standards: Evidence for the content validity of the Wisconsin alternate assessment. *Journal of Special Education, 38*, 218-232.

Schafer, W.D., Swanson, G., Bené, N., & Newberry, G. (2001). Effects of teacher knowledge of rubrics on student achievement in four content areas. *Applied Measurement in Education, 14*, 151-170.

Shepard, L.A., Flexer, R.J., Hiebert, E.H., Marion, S.F., Mayfield, V., & Weston, T.J. (1996). Effects of introducing classroom performance assessments on student learning. *Educational Measurement: Issues and Practice, 15*(3), 7-18.

Smithson, J.L., & Porter, A.C. (2004). From policy to practice: The evolution of one approach to describing and using curriculum data. In M. Wilson (Ed.), *Towards coherence between classroom assessment and accountability. 103rd Yearbook of the National Society for the Study of Education* (Part II, pp. 105-131). Chicago, IL: University of Chicago Press.

Stecher, B. (1998). The local benefits and burdens of large-scale portfolio assessment. *Assessment in Education, 5*, 335-351.

Willingham, W.W., Pollack, J.M., & Lewis, C. (2002). Grades and test scores: Accounting for observed differences. *Journal of Educational Measurement, 39*, 1-37.

Wilson, M., & Draney, K. (2004). Some links between large-scale and classroom assessments: The case of the BEAR assessment system. In M. Wilson (Ed.), *Towards coherence between classroom assessment and accountability. 103rd Yearbook of the National Society for the Study of Education* (Part II, pp. 132-154). Chicago, IL: University of Chicago Press.

Wolf, S.A., Borko, H., Elliott, R.L., & McIver, M.C. (2000). "That dog won't hunt!": Exemplary school change efforts within the Kentucky reform. *American Educational Research Journal, 37*, 349-393.

Wolfe, R., Childs, R., & Elgie, S. (2004). *Ensuring quality assessments: A project to refine and affirm assessment processes. Final report of the external assessment of EQAO's assessment processes*. Toronto, ON: OISE/UT.