(In)Stability of Test Scores

Stefan Merchant[1], Jessica Rich[1], & Don A. Klinger[2]
[1]Queen's University, [2]University of Waikato

Abstract
Both school and district administrators use the results of standardized, large-scale tests to inform decisions about the need for, or success of, educational programs and interventions. However, test results at the school level are subject to random fluctuations due to changes in cohort, test items, and other factors outside of the school's control. This study examined year to year changes in school level results on standardized tests delivered in Ontario, Canada. G-theory analyses found that test scores are not stable enough for meaningful conclusions to be made based on year to year changes in school level results. For small and medium sized schools, years of data need to be collected before defensible decisions can be made about trends in test scores. The authors introduce a 'bounce' statistic that provides a simple, easy to interpret measure of test score stability.

Keywords: large-scale testing, G-theory, educational policy, test reliability

Introduction
All provinces have testing programs to help them evaluate student achievement in relation to curriculum expectations (e.g., British Columbia, 2021; Prince Edward Island, 2019). These tests may be implemented in early years (e.g., Saskatchewan, 2021), in middle school (e.g., Manitoba, n.d.), or for graduation (e.g., Alberta, 2021). In Ontario, Grade 3 and 6 students are required to write province-wide literacy and numeracy tests. These tests originated from the Education Quality and Accountability Office (EQAO), which implement a variety of standardized tests in Ontario schools. While some of those tests are high stakes for students (e.g., Ontario Secondary School Literacy Test), the grade 3 and 6 tests are low stakes, as there are no sanctions, promotions, or report card grades associated with student performance. The primary purpose of the EQAO tests is to monitor overall student achievement against the provincial standards for reading, writing, and mathematics (Educational Quality and Accountability Office, 2020), but the intended and unintended purposes for the tests have expanded. As examples of intended uses, the Ontario Ministry of Education has directed funds and resources to schools with very low student performance on these tests. Test results have also been used to implement educational initiatives and policies within the province and to monitor the subsequent impact of these initiatives (Bolden et al., 2014; Canadian Language and Literacy Research Network, 2008; Klinger & Rogers, 2011; Klinger & Wade-Woolley, 2009). At the same time, unintended uses include examples of organizations that use the EQAO results to rank schools, to infer differences in educational quality between schools, or to infer which neighbourhoods have the best schools (e.g., Cowley & Emes, 2020; Scholarhood, 2017; Volante, 2004). One Ontario realty website (http://www.realtyforsale.ca/best_schools/) even goes so far as to list local elementary schools using a traffic light system of red (for low ranking schools), yellow, and green (for high ranking schools).

Ontario schools often use EQAO test results to establish, monitor, and evaluate school improvement plans. Hence it is common for schools to report increases in their EQAO test results as evidence of the impact of improvement plans and the lack of improvement as the need for renewed efforts (e.g., Calder,

2015; Renfrew County District School Board, 2016; Toronto District School Board, 2018). Such uses can result in high stakes for teachers and administrators. Elementary school principals have anecdotally told us they are evaluated based upon EQAO test results; consequently, they make leadership decisions designed to augment the scores. These decisions include which teachers get placed into Grades 3 and 6 (grades in which the EQAO tests are administered) and what types of resources are purchased for the school. Despite evidence that large-scale tests do not yield data suitable for school improvement, schools and districts continue to focus their school improvement efforts on improving their large-scale test scores (Klinger et al., 2006; Rogers, 2014; Ungerleider, 2006).

As a result, such testing has "dramatically reshaped the responsibilities of school leaders" (Leithwood, 2012, p. 17). Canadian educational experts have exhorted school leaders to become data-driven (e.g., Earl & Katz, 2006). As Earl (2008) noted, "real benefits can accrue from 'getting to know data'" (p. 43). While the potential benefits of such information may be real, this requires school leaders to have some baseline proficiency in data literacy (Goren, 2012). In the case of large-scale testing, the ability to properly interpret score results requires a foundational understanding of statistics and an appreciation for the limitations of what kinds of interpretations can be meaningfully derived from test results. As an important example, it is tempting for a school administrator to conclude from the improved test scores that their school improvement plans are having the desired effect. However, prior research on large-scale testing has identified that the primary source of score variance is the students themselves (Anderson, 2012; Hollingshead & Childs, 2011; Klinger et al., 2006; Ungerleider, 2006). Students' prior achievement, demographic characteristics, and personal qualities are all associated with achievement on standardized tests (Hattie, 2008). Since the cohort of students in any given class changes from year to year, this leads to instability of the results (Leithwood, 2012). This instability makes meaningful comparisons between years difficult, adding complexity to the analyses needed to determine whether or not school reforms have led to changes in test scores or if said changes are due to other factors. Leithwood (2012) categorized school scores on the EQAO tests as "stable", "increasing" or "decreasing" and found that few schools (7.4%) demonstrated continuous improvement. Instead, the most common pattern of performance was categorized as "no predictable direction" (p. 23). While Ontario educators can take heart in the fact that fewer than 1% of schools demonstrated a continuous decline in their EQAO test scores, the main finding is that at the school level, test scores fluctuate in ways that are not easily understood.

The variability in student cohorts may also further impact the stability of results in the presence of other factors. For example, schools with smaller class sizes will likely have greater instability than schools with large class sizes. Hollingshead and Childs (2011) noted that the majority of Ontario elementary schools have fewer than 40 students writing the Grade 6 literacy and numeracy tests each year. Given that score stability is lower when student populations are smaller, the large number of Ontario schools with small student populations poses a challenge in examining year-to-year changes in test results. Hollingshead and Childs (2011) further showed that how test results are reported can impact stability. For the grade 3 and 6 EQAO examinations, students are categorized into 4 different levels and achieving Level 3 or 4 is considered as "Meeting provincial standards". The percentage of students who meet provincial standards is the number that is reported at the school level by EQAO. However, as Hollingshead and Childs (2011) highlighted, reporting the percentage of students who achieve above a predetermined cut score leads to greater year-to-year variation in the results than reporting a mean score on the test. They found that, in a school of 60 students, there is a 1 in 3 chance the percentage of students who achieve a level 3 or 4 on the test will change by 10% from one year to the next due to random fluctuations in student achievement. Averaging data over more than one year increased the stability of scores, but their research did not uncover how much data are needed to make robust, defensible conclusions about the stability of school-level test results.

The purpose of our research was to look at sources of variation and stability of school achievement scores on the large-scale literacy and numeracy tests delivered to Grade 3 and 6 students across Ontario. In particular, our research aims to determine the sources of variance that affect the school results and the amount of data required to make defensible judgements about trends in school-level performance. Our findings will better inform debates surrounding the impacts of school improvement efforts and the ranking of schools, and will also be important to educators and policymakers seeking to connect changes in practice with large-scale test results.

Generalizability Theory

Generalizability theory (G-theory) is a statistical method with roots in classical test theory that allows researchers to identify the contributions of different sources of variance and the amount of data needed to achieve score stability (Brennan, 2010). One common application of G-theory is to identify sources of variance in test scores. Examples of such sources include; raters, test items, and testing occasions. For our study, the sources of variance investigated were the school, the year the test was written, and the subject of the test (mathematics, reading, and writing).

Briesch et al. (2016) conducted a review of studies in K-12 education using G-theory and found that of the 45 studies reviewed, 33 investigated raters and their impact on scores and all of them used the student as the object of measurement. Our study is unique in that we used schools as the object of measurement, meaning we used G-theory to identify the sources of variance in school performance (and not student performance) on large-scale tests.

A G-theory analysis is essentially a reliability analysis. Instead of calculating a reliability statistic such as Cronbach's alpha, a generalizability coefficient is calculated. As with Cronbach's alpha, the closer the generalizability coefficient is to 1.0, the more generalizable (reliable) the measure. While other statistical techniques are available to estimate reliability and identify sources of variance, G-theory has the advantage of using the identified sources of variance to predict the number of facets (e.g., items, raters, occurences) required to obtain consistent results. This is done when information from the G-theory analyses is used to inform a decision or "D-study". The purpose of a D-study is to determine how many of each facet are needed to obtain predetermined levels of generalizability. As an example, a D-study may determine how many test items are needed to obtain desired levels of generalizability for a test. In our case, we used the D-study to determine how many years of test data would be needed to obtain dependable (generalizable) EQAO results.

Study Design

In the language of G-theory, our research incorporates a fully-crossed three-facet design (school x year x test) with 6 sources of variability plus residual (unexplained) error (Table 1).

Table 1

Sources of Variability in our Three-Facet, Fully Crossed Study Design

Source of Variability (facet)	**Description**	**Variance Notation**
Schools (*s*)	Universe-score variance (object of measurement)	σ^2_s
Years (*y*)	Constant effect for all schools due to behavioural inconsistencies of students and teachers from one year to another (different cohorts)	σ^2_y
Tests (*t*)	Constant effect for all schools due to differences in test domains	σ^2_t
s x *y*	Interaction between schools and years representing inconsistencies in school achievement from one year to the next	σ^2_{sy}
s x *t*	Interaction between school and test representing inconsistencies in school achievement across test domains (reading, writing, mathematics)	σ^2_{st}
y x *t*	Interaction between year and test representing inconsistencies in test domains across years	σ^2_{yt}
s x *y* x *t*, *e*	Residual consisting of a three-way interaction between school, year, and test, plus remaining residual measurement error (due to unmeasured facets that affect the measurement and/or random events).	$\sigma^2_{syt,e}$

In G-theory, the sources of variance are called "facets" and may be "random" or "fixed". Strictly speaking, a random facet means the sample comes from an infinitely large universe of possible samples.

Practically speaking, as long as the universe of possible samples is potentially large, a facet can be considered random. For example, students are usually treated as a random facet because the population of students is very high and because researchers usually want their G-theory analyses to generalize to all students. In our study, the facets 'school' and 'year' are treated as random. A fixed facet is one in which contains a small and limited number of possible samples. In our study, 'Test' is treated as a fixed facet because there are only three domains measured by EQAO tests (reading, writing, and mathematics).

G-theory analyses are based on analyses of variance (ANOVAs) and as such, they examine not only main effects but interactions. Thus, we estimated variance components for six sources of variability (facets): school, year, test, school*year interaction, school*test interaction, and year*test interaction (Table 1). The percentage of variance accounted for by each facet was calculated so that relative comparisons could be made amongst facets based on size. The variance estimates were then used to calculate generalizability coefficients using the following formula:

$$\text{G-coefficient} = \frac{\sigma^2_s}{\sigma^2_s + \sigma^2_{sy} + \sigma^2_{st} + \sigma^2_{syt,e}}$$

Looking at the formula, one can see that, in this case, the G-coefficient indicates the proportion of the total variance that is due to the school variance.

Bounce – A New 'Pragmatic' Statistic

The abstract language and conceptual demands of G-theory (e.g., universe of admissible observations, fixed and random facets, etc.) make this statistical approach daunting and inaccessible for teachers and administrators. To better mobilize our research and make the findings more accessible to our intended users, we created an alternative means of calculating variability in school achievement over time. We have called this statistic "bounce" (i.e., the 'bounce' in school-level achievement from one year to the next).

For each school in a sample, we calculated the absolute value of the change in reading, writing, and mathematics test results from one year to the next, and used those values to calculate a mean value for the year-to-year change in the percentage of students who meet provincial standards on the EQAO tests. This represents the average change in a school's results on a particular test from one year to the next. Readers familiar with summation notation will recognize the bounce formula given below as the formula to calculate a mean value.

$$Simple\ Bounce = \frac{1}{k}\sum_{n=1}^{n=k} |Score_n - Score_{n+1}|$$

While this statistic is simple and easy to understand, it is important to recognize that the average change in schools' test results includes both signal and noise. Signal represents systematic changes in school achievement that may be due to provincial-, district- or school-based initiatives. Noise constitutes the random variability in school achievement scores that can be attributed to other factors including, but not limited to, different cohorts of students, changes in teaching staff, number of inclement-weather days, etc. To separate signal from noise, we calculated the slope of the line of best fit for each school's achievement over time. This slope was then subtracted from the simple bounce calculation. In doing so, we obtain a better estimate of random fluctuations in achievement by correcting for the sustained improvements or declines schools may demonstrate. Mathematically, this is expressed aa

$$Corrected\ Bounce = Simple\ Bounce - |Slope\ of\ line\ of\ best\ fit\ for\ achievement\ over\ time|$$

As will be shown in the results section, the sustained change that schools demonstrated over time was usually very small compared to the random fluctuations. This means educators could use the simple version of the bounce calculation and obtain meaningful results. We believe this is important if the statistic is to be used by professionals and not just researchers.

Data Source

We applied to the EQAO Data Portal for Researchers and gained access to school-level achievement data

for eight consecutive reporting periods between the years of 2005 and 2012. These years were chosen as this was a period of time in which EQAO provincial assessments were central to school improvement efforts, with little change to curriculum. More recent data have been impacted by major curriculum reforms beginning in 2013, and ongoing changes to the provincial assessment programme. EQAO aggregates student-level results into school-level results and reports the percentage of students achieving or exceeding the provincial standard in each separate test domain (reading, writing and mathematics). In Ontario, the provincial standard is Level 3; equating to letter grade B, or a percentage mark within the range of 70 to 79% (Ontario Ministry of Education, 2010). From the school-level (Grades 3 and 6) data, we extracted the following variables: school names; school ID numbers; the percentage of students achieving or exceeding the provincial standard in reading, writing, and mathematics; and the number of students in Grades 3 and 6. The total number of schools for which data were reported exceeded 4000 schools, but we excluded schools for which achievement data were missing between 2005 and 2012 leaving us with 2947 schools to sample from. Schools that were excluded from the analyses tended to be either small and/or special schools with inconsistent participation in EQAO testing and/or suppressed data. It should be noted that any schools where the testing population was below N = 15 did not have data reported and so our sample does not include any of these schools.

From our final population of 2947 schools, we needed to create samples suitable for our G-theory analyses. Given we were interested in determining how school size (student population) affected the stability of EQAO results, we needed to sample schools based on their population of test-takers. To do this, we sorted schools in descending order (high population of test-takers to low) and sampled groups of schools from the top (largest population), middle (median sized-population), and bottom (smallest population) of the list. Sample sizes of 100 were chosen because this number struck the appropriate balance between having a large sample and minimizing facet variability within that sample.

Analyses

We used the Statistical Software for the Social Sciences (SPSS, Version 24) to estimate variance components and then Excel to calculate generalizability coefficients and bounce. Before estimating variance components in SPSS, data for each sample of schools had to be reformatted from multivariate to univariate format. In univariate format, the object of measurement (school) has multiple rows of data for each case. Within SPSS, the VARTOCASES command was used to make the multivariate-to-univariate conversion. Some sample calculations were run independently using a G-theory specific software called GENOVA to check for consistency with the Excel results. GENOVA and SPSS yielded either identical or very slightly different results for the variance components. Differences in the variance components were likely due to the different algorithms each software package used to calculate the variance components. Any numerical differences found in the variance components made a minimal difference in the G-coefficient and no differences in our overall findings or conclusions.

Decision Study (D-Study)

We used the variance estimates from the G-Study to make decisions about the number of tests and years of data we would need to make reliable estimates of schools' true achievement scores. To do this, we computed G-coefficients (indicators of reliability) for different scenarios involving different numbers of tests and years of data (facets). Scenarios producing G-coefficients greater than the criterion level G (e.g., 0.80) were compared based on relative feasibility and associated costs.

Results

The results presented here (Table 2) were calculated based upon the three sub-samples of 100 schools (large, medium, and small test-taking populations).

Table 2

Descriptive Statistics of Each Sub-sample

School size	Sample size	Mean # of students	SD
Grade 3			
Small	100	21.71	1.12
Medium	100	39.67	0.36
Large	100	101.58	19.91
Grade 6			
Small	100	21.66	1.27
Medium	100	42.96	0.50
Large	100	161.18	63.51

Variance Components

Generalizability theory was used to estimate variance components for the facets 'school,' 'year,' 'test,' and their interactions. Variance components were calculated using the restricted maximum likelihood method (McNeish, 2017). Summaries of the variance estimates for each of the sub-samples can be found in Tables 3 and 4. The results across all sub-samples showed that the variance across tests and years did not contribute significantly to the total variance. Consistently across sub-samples and grades, the facets 'school,' 'school*year' interaction, and residual 'error' account for roughly 80–95% of the total variance.

Table 3

Percentage Variance Estimates for School Size Analyses

	Grade 3		
Component	**Small**	**Medium**	**Large**
School	27.13	37.24	39.78
Year	1.9	1.82	8.18
school*year	42.25	35.59	29.97
school*test	1.62	3.6	5.7
year*test	1.91	1.33	2.56
error	25.2	20.43	13.81
	Grade 6		
Component	**Small**	**Medium**	**Large**
school	30.3	50.3	63
year	0.91	0.43	2.45
school*year	36.66	22.66	16.19
school*test	3.97	3.88	4.48
year*test	4.9	4.33	4.6
error	23.26	18.41	9.28

Generalizability Coefficients

The mean G-coefficient for one test and one year is reported along with the standard deviation obtained from bootstrapping. Because bootstrapping had to be done manually, we limited ourselves to 20 different "runs" using 50 schools selected at random from each sub-sample of 100 schools. The standard deviations of the runs indicate that further bootstrapping would not yield substantially different results. We reported the G-coefficient for one year and one test as it is common for educators to make year-to-year comparisons based on a single subject area (e.g., "Our mathematics scores were 7% higher this year compared to last year"). As expected, the school size analyses demonstrated that smaller schools have less stability in the percentage of students who achieved Level 3 or 4, compared to larger schools. This is consistent with the findings of Hollingshead and Childs (2011).

Table 4
Mean G-coefficients

School size	Grade 3 G-coefficient		Grade 6 G-coefficient	
	Mean	SD	Mean	SD
Small	0.28	0.06	0.32	0.05
Median	0.38	0.05	0.53	0.06
Large	0.45	0.06	0.68	0.05

Bounce Calculations

While G-theory is well known among psychometricians, it is a difficult statistical procedure for school-based administrators. Thus we also calculated the bounce statistic described earlier. Our expectation was that the bounce statistic would capture some of the same information as the G-coefficient in a more practical manner. As shown in Table 5, the bounce results follow the same pattern as the G-coefficients—larger schools have less bounce, and therefore more stable scores. There is less bounce in Grade 6 results than in Grade 3.

To investigate to what extent G-coefficients and bounce calculations yield the same information, we graphed the G-coefficient and bounce statistic for 30 different samples of Grade 6 schools. As can be seen from Figure 1, G-coefficients and bounce values correlate strongly (r = -0.95), but not perfectly. The bounce values are clustered together because they are the mean bounce of all schools in each sample and averaging such a large number of values reduces variation.

One advantage of the bounce calculation is that the units are the same as the units in which the test scores are reported. This makes for easier interpretation by school administrators. In this example, the units for the bounce calculation are the percentage of students who achieved Level 3 or above on the EQAO tests. Looking at the Math test results for our Grade 6 medium schools we see a bounce value of 12.28 (SD = 3.93). This means that between one year and the next, the average difference in the percentage of students who achieved Level 3 or above on the math test was 12.28%. In other words, a change from 62% achieving Level 3 or above to 72% in the next year would be considered normal. If we use within two standard deviations as our definition of "normal", we can see that any deviation of less than 24.56% could be considered normal. These figures come from our uncorrected bounce calculation, but even when the corrected bounce calculation is used, we still have a bounce value of 11.61 percent meaning that while a change from 62% to 45% appears to be a drastic decline in achievement, it could be a normal fluctuation in scores that occurs over time, and not due to school factors such as changes in instructional quality.

Table 5

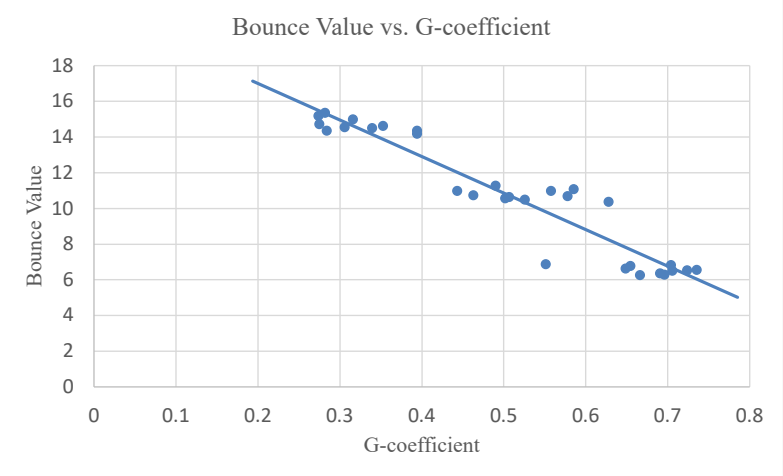
Bounce Calculations for Different School Sizes

	Math				Reading				Writing			
Grade 3	Mean	SD	Slope	Corrected	Mean	SD	Slope	Corrected	Mean	SD	Slope	Corrected
Small Schools	16.25	6.82	0.38	15.87	15.40	6.21	0.79	14.61	14.92	4.96	2.16	12.76
Medium Schools	12.52	4.88	0.39	12.13	11.28	4.15	0.79	11.49	11.28	4.62	1.65	9.63
Large Schools	7.59	2.80	0.53	7.06	8.34	3.27	1.18	7.16	7.42	3.03	1.92	5.50
Grade 6												
Small Schools	15.09	5.42	-1.14	13.95	14.18	5.36	1.97	12.21	15.16	5.29	1.97	13.19
Medium Schools	12.28	3.93	-0.67	11.61	10.19	3.33	1.60	8.59	10.37	4.03	1.72	8.65
Large Schools	7.23	3.14	-0.04	7.19	6.21	2.42	1.72	4.49	5.96	2.25	2.24	3.62

Bounce calculations look at all changes in achievement, including sustained improvements or declines. Because we were interested only in the random variation in the school scores, we calculated the regression line for school achievement over time for all our subsamples. The slope of this regression line is reported and can be subtracted from the bounce calculation to correct for systematic changes in achievement. This is represented in Table 5 as corrected bounce. While this is an improvement in the calculation, we note that slopes are usually small compared to the bounce value, meaning that systematic changes in achievement are generally much smaller than random fluctuations. The most obvious exception to this finding is the Grade 6 writing results for large schools in which the slope value was 37% of the bounce value. However, even in this case, we would argue that reporting the simpler, uncorrected bounce value will yield valuable information that is precise enough to inform administrative decision-making in Ontario schools. We recognize this may not be true for all samples and so there is value in calculating the corrected value for the bounce statistic.

Figure 1

Bounce Statistic vs. G-Coefficient for Grade 6 Schools



D-studies

A significant advantage of G-theory is its ability to make predictions about dependability under different conditions. These predictions are known as D-studies and, in our case, give us information about how many years, or how many tests, are required to achieve a desired level of dependability. The results of the D-studies for our sub-samples of interest are reported in Table 6.

Generally, a D-coefficient value of 0.80 or above is considered good (e.g., Broglio et al., 2009; Gagnon et al., 2009). As can be seen from Table 6, three tests (reading, writing, mathematics) yields more dependable school achievement results than one test alone. To reach a 0.80 threshold in Grade 3, small- and medium-sized schools need more than five years of data and three tests per year, whereas large schools could make dependable conclusions based on five years of data and one test per year. In Grade 6, small schools still need more than five years of data and three tests per year, whereas medium and large schools could make generalizable conclusions based on 3 years and 3 tests. Large schools crossed the 0.80 threshold after three years with only one test.

Table 6

D-coefficients for Different School Sizes

School Size	**Grade 3**					
	1 year, 1 test	3 years, 1 test	5 years, 1 test	1 year, 3 tests	3 years, 3 tests	5 years. 3 tests
Small	0.28	0.53	0.64	0.35	0.60	0.71
Medium	0.38	0.62	0.71	0.46	0.71	0.79
Large	0.45	0.66	0.73	0.52	0.75	0.82
	Grade 6					
	1 year, 1 test	3 years, 1 test	5 years, 1 test	1 year, 3 tests	3 years, 3 tests	5 years. 3 tests
Small	0.32	0.56	0.65	0.40	0.65	0.74
Medium	0.53	0.74	0.80	0.62	0.82	0.88
Large	0.68	0.83	0.87	0.75	0.89	0.92

Discussion

The purpose of our research was to determine the generalizability (dependability) of school-level results on Grades 3 and 6 EQAO tests (reading, writing and mathematics) for different sized schools, allowing us to estimate the number of occurrences of assessment data (years and test domains) required to make defensible judgements about school-level provincial test performance. The results presented here indicate that variance in school achievement is mostly due to the size of the school (test taker population), interactions between school size and the cohort of test-takers (i.e., inconsistencies in school cohorts from one year to the next), and residual error (i.e., the three-way interaction between school size, year, and test, and unexplained measurement error).

Our analyses demonstrated that Grade 6 scores are more stable than Grade 3. We are unsure as to why this is, but hypothesize that several factors could be at play. The first is that class sizes tended to be larger in Grade 6 than in Grade 3, and, as we have shown, greater numbers of students lead to greater score stability. Secondly, it is possible, that by the time students reach Grade 6, school effects have become stronger and so the impact of individual differences of students within cohorts is diminished. This hypothesis is supported by the fact that the unexplained variance in the Grade 3 schools is slightly higher than in Grade 6 schools. These findings provide an interesting focus for future work.

Based on our findings, meaningful judgments about the impact of school improvement efforts on EQAO scores at the Grade 3 level cannot be done in a single year, nor even after five years. Thus, for small- and medium-sized schools (which are most schools), EQAO results are not good measures of the

impact of school improvement efforts in a single subject, or over short periods of time (i.e., five years or less). Some schools and school districts (e.g., Rainbow Districts School Board, 2016; Upper Canada District School Board, 2018) recognized that making year-to-year comparisons in test results is problematic and so use rolling averages. Our D-study results provide an estimate with respect to the time period these rolling averages should cover. For instance, medium-sized schools could use a five-year rolling average for their Grade 6 EQAO results. The D-study results show that for Grade 3 schools (of any size) rolling averages would need to be longer than five years. Given that many neighbourhood schools will see demographic shifts during periods greater than five years, even the use of 5-year rolling averages may not provide a dependable measure of school improvement efforts. Short time spans do not provide reliable data and longer time spans are affected by demographic shifts and other external factors that will affect a school's average achievement on these external tests.

Of course, Ontario schools can use results from all three tests (mathematics, reading, and writing) to shorten the time frame needed to achieve dependable rolling averages, but there are two problems with this approach. The first is that most schools will need still need more than five years for their Grade 3 results to be reported with acceptable levels of dependability. The second, and more important, problem is that reporting the results of three separate tests in aggregate format combines constructs in a way that is not useful or helpful. If a school implements a new reading program, it is only the scores on the reading test that should be considered when evaluating the impact of the program and so aggregate scores across all three tests are not useful for gathering information about school-level initiatives targeted at specific subjects like mathematics, reading or writing. It is possible that results from all three tests may provide information or insight about the effectiveness of school-based initiatives targeted at *general* learning (e.g., a program designed to improve students' self-regulated learning), but they should not be used to make decisions about subject specific initiatives.

Ultimately, our results imply that if a school administrator wishes to evaluate the impact of an initiative (e.g., a strategy to improve mathematics learning), test scores need to be supplemented with other data collection tools such as classroom assessment data, interviews with students, discussions with teachers, and parental feedback. Test scores by themselves are not dependable enough to evaluate the effectiveness of school initiatives and programs over the short term (i.e., one to three years) and, for smaller schools (e.g., 20 to 25 test-takers), even five years of data would not yield sufficient generalizability to make defensible decisions. It should also be noted that our analyses ignore the smallest schools since schools with less than 15 students writing the EQAO tests in any year were not included. As a result, many schools in Ontario are smaller than the "small" schools included in our sample. These schools will have score stability even lower than what we report here.

While analyses of variance, and G-theory in particular, can yield rich information about the reliability of test scores, practicing educators need a simpler statistic. The bounce statistic we introduced provides a simple and easily understood measure of score stability. The advantages of this statistic are its ease of calculation, along with the fact that it is expressed in the same units as the test scores, making interpretation and comparison easier. As expected, large schools had the lowest bounce, meaning they had the most stable scores. The lowest (corrected) bounce statistic we found was for Grade 6 writing in large schools with a value of 3.62. This means that from one year to the next, a typical change in the percentage of students who achieved Level 3 or 4 in writing was under 4%. While a corrected bounce value of 3.62 indicates good stability, it must be remembered this result comes from the 100 largest schools in our sample and is not representative of most schools in the province. Our highest (corrected) bounce statistic had a value of 15.87 for Grade 3 mathematics in small schools meaning that for these schools, the percentage of students achieving level 3 or 4 typically fluctuates by almost 16% from year to year.

Despite the inherent instability in EQAO test results, some Ontario schools report declines in EQAO achievement of even two or three percent as being worthy of concern and needing to be addressed through school improvement initiatives (e.g., Waterloo Region District School Board, 2016). Similarly, districts report gains of as little as one percent as being noteworthy and evidence of improvement (e.g., Hamilton Wentworth District School Board, 2019; Hastings Prince Edward District School Board, 2012; Limestone District School Board, 2017). As our bounce calculations show, at a school level, changes of a few percent are well within what would be expected from random fluctuations in test results. As former educators and administrators, we understand the political utility of using changes in test results to motivate teachers and students, instill confidence in parents, and give a rationale for new initiatives.

However, it is imperative that school (and district) administrators have a realistic sense of what changes are likely to be meaningful and what changes are likely due to random fluctuation. It would be a shame if promising new initiatives were abandoned because of declines in test results due to random fluctuation. The bounce statistic can help administrators put into perspective the magnitude of their year-to-year changes in test results.

Conclusion

Our results show that at a school level, results for the Grade 3 and 6 EQAO literacy and numeracy tests have not been stable enough to make year-to-year comparisons. The results of our D-study indicate that if schools are to use rolling averages to report changes in their EQAO test results, those averages would need to include more than five years of data for all but the largest schools to give results that are dependable enough to inform decision-making. While it is possible that ongoing changes to the examination program may lead to more stable estimates, it appears unlikely. Schools continue to report results in terms of achievement of levels and the technical reports for the tests have not indicated notable changes in the reliability of the tests themselves (e.g., EQAO, 2017). As a result, we are confident our results from a slightly earlier time period continue today, and that EQAO test scores are not good short-term indicators of the success of improvement efforts in Ontario elementary schools. The simple bounce statistic we present here provides administrators with a pragmatic approach to estimate the (in)stability of test scores from one year to the next. Hopefully, this will spur them to seek more diverse and fulsome data to collect and analyze to evaluate school improvement initiatives.

References

Alberta Ministry of Education. (2021). *Student learning assessments.* https://www.alberta.ca/student-learning-assessments.aspx

Anderson, J. O., Lin, H. S., Treagust, D. F., Ross, S. P., & Yore, L. D. (2007). Using large-scale assessment datasets for research in science and mathematics education: Programme for International Student Assessment (PISA). *International Journal of Science and Mathematics Education*, *5*(4), 591-614. https://doi.org/10.1007/s10763-007-9090-y

Artuso, A. (2016, February, 28). School rankings raise many questions. *The Toronto Sun*. http://www.torontosun.com/2016/02/27/school-rankings-raise-many-questions.

Bolden, B., Christou, T., DeLuca, C., Klinger, D. A., Kutsyuruba, B., Pyper, J., Shulha, L. M., & Wade-Woolley, L. (2014). *Collaborative inquiry in Ontario schools. An evaluation report for the Ontario Ministry of Education*. Literacy and Numeracy Secretariat.

Brennan, R. L. (2010). Generalizability theory and classical test theory. *Applied Measurement in Education*, *24*(1), 1-21. https://doi.org/10.1080/08957347.2011.532417

Briesch, A. M., Chafouleas, S. M., & Johnson, A. (2016). Use of generalizability theory within k–12 school-based assessment: A critical review and analysis of the empirical literature. *Applied Measurement in Education*, *29*(2), 83-107. https://doi.org/10.1080/08957347.2016.1138955

British Columbia Minstry of Education. (2021). *Foundation skills assessment.* https://www2.gov.bc.ca/gov/content/education-training/k-12/administration/program-management/assessment/foundation-skills-assessment.

Broglio, S. P., Zhu, W., Sopiarz, K., & Park, Y. (2009). Generalizability theory analysis of balance error scoring system reliability in healthy young adults. *Journal of Athletic Training*, *44*(5), 497-502. https://doi.org/10.4085/1062-6050-44.5.497

Calder, M. (2015). *Board working to improve grade 9 EQAO math scores.* http://www.ucdsb.on.ca/ucdsbnews/2015-2016SchoolYear/October/Pages/UCDSBGrade9MathEQAOScores.aspx

Canadian Language and Literacy Research Network. (2008). *The impact of the literacy and numeracy secretariat: Phase 2 program evaluation*. University of Western Ontario.

Cowley, P., & Emes, J. (2020). *Report card in Ontario's elementary schools 2020.* Fraser Institute. https://www.fraserinstitute.org/sites/default/files/ontario-elementary-school-rankings-2020-13385.pdf

Earl, L. (2008). Leadership for evidence-informed conversations. In L. M. Earl & H. Timperley (Eds.), *Professional learning conversations: Challenges in using evidence for improvement* (Vol. 1, pp. 43-52). Springer Science & Business Media.

Earl, L., & Katz, S. (2006). *Leading in a data rich world: Harnessing data for school improvement*. Corwin.

Educational Quality and Accountability Office. (2017). *Ontario student achievement: EQAO's provincial elementary school report: Results of the assessments of reading, writing and mathematics, primary division (grades 1–3) and junior division (grades 4–6), 2016–2017.* https://www.eqao.com/provincial-report-elementary-2017/

Educational Quality and Accountability Office. (2020). *About EQAO*. https://www.eqao.com/about-eqao/

Gagnon, R., Charlin, B., Lambert, C., Carriere, B., & Van der Vleuten, C. (2009). Script concordance testing: more cases or more questions? *Advances in Health Sciences Education, 14*(3), 367-375.

Goren, P. (2012). Data, data, and more data—What's an educator to do? *American Journal of Education, 118*(2), 233-237.

Hamilton Wentworth District School Board. (2019). *HWDSB EQAO results leads to investment in people, practice and progress.* https://www.hwdsb.on.ca/wp-content/uploads/2019/09/EQAO-Infographic-2019.pdf

Hastings Prince Edward District School Board. (2012). *EQAO results for grades 3, 6 and 9 continue to improve*. http://www.hpedsb.on.ca/archives/eqao-results-for-grade-3-6-and-9-continued-to-improve/

Hattie, J. (2008). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement.* Routledge.

Hollingshead, L., & Childs, R. A. (2011). Reporting the percentage of students above a cut score: The effect of group size. *Educational Measurement: Issues and Practice, 30*(1), 36-43. https://doi.org/10.1111/j.1745-3992.2010.00198.x

Klinger, D. A., DeLuca, C., & Miller, T. (2008). The evolving culture of large-scale assessments in Canadian education. *Canadian Journal of Educational Administration and Policy, 76*(3), 1–34.

Klinger, D. A., & Rogers, W. T. (2011). Teachers' perceptions of large-scale assessment programs within low-stakes accountability frameworks. *International Journal of Testing, 11*(2), 122–143. https://doi.org/10.1080/15305058.2011.552748

Klinger, D. A., Rogers, W. T., Anderson, J. O., Poth, C., & Calman, R. (2006). Contextual and school factors associated with achievement on a high-stakes examination. *Canadian Journal of Education, 29*(3), 771–797. https://doi.org/10.2307/20054195

Klinger, D. A., & Wade-Woolley, L. (2009). *Supporting low performing schools in Ontario. Technical report prepared for the U. S. department of education.* WestEd Organization.

Leithwood, K. (2011). School leadership, evidence-based decision making, and large-scale student assessment. In C. Webber & J. Lupart (Eds.), *Leading student assessment* (pp. 17-39). Springer.

Limestone District School Board. (2017). *EQAO results show achievement in some levels continuing to improve*. https://www.limestone.on.ca/news/news_releases_2017-2018/e_q_a_o_results_show_achievement_in_some_levels_co

Manitoba Ministry of Education. (n.d.). *Assessment and evaluation.* https://www.edu.gov.mb.ca/k12/assess/assess_program.html

McDonnell, L. M. (2005). Assessment and accountability from the policy maker's perspective. In J. Herman & E. Haertel (Eds.), *Uses and misuses of data for educational accountability and improvement (104th Yearbook of the National Society for the Study of Education)* (pp. 35–54). Blackwell.

McNeish, D. (2017). Small sample methods for multilevel modeling: A colloquial elucidation of REML and the Kenward-Roger correction. *Multivariate Behavioral Research, 52*(5), 661-670. https://doi.org/10.1080/00273171.2017.1344538

Ontario Ministry of Education. (2010). *Growing success: Assessment, evaluation and reporting in Ontario schools.* Author. http://www.edu.gov.on.ca/eng/policyfunding/growSuccess.pdf

Prince Edward Island Ministry of Education. (2019). *Provincial assessments.* https://www.princeedwardisland.ca/en/information/education-and-lifelong-learning/provincial-assessments

Rainbow District School Board. (2016). *School valuation framework*. https://www.rainbowschools.ca/wp-content/uploads/2016/04/School_Information_Profile.pdf

Renfrew County District School Board. (2016). *Board improvement plan for student achievement and well-being kindergarten to grade 12: 2016-2017.* https://www.rcdsb.on.ca/en/resources-General/RCDSBBIPSA2016-2017-1.pdf

Rogers, W. T. (2014). Improving the utility of large-scale assessments in Canada. *Canadian Journal of Education/Revue canadienne de l'éducation, 37*(3), 1-22.

Scholarhood. (2017). *Compare schools & neighbourhoods. We help families find homes in the boundaries of the best schools.* www.scholarhood.ca

Toronto District School Board. (2018). *Multi-year strategic plan*. https://www.tdsb.on.ca/Portals/0/leadership/board_room/Multi-Year_Strategic_Plan.pdf

Ungerleider, C. (2006). Reflections on the use of large-scale student assessment for improving student success. *Canadian Journal of Education, 29*(3), 873–873. https://doi.org/10.2307/20054200

Upper Canada District School Board. (2018). *Board improvement plan for student achievement and wellness 2018-2019.* https://p16cdn4static.sharpschool.com/UserFiles/Servers/Server_148343/File/Our_Board/District%20Plans/BIPSAW/BIPSAW%20UCDSB%202018-2019%20Full%20Version.pdf

Volante, L. (2004). Teaching to the test: What every educator and policy-maker should know. *Canadian Journal of Educational Administration and Policy, 35*, 1-9.

Waterloo Region District School Board. (2016). *Standardized test results show room to improve*. https://cle.wrdsb.ca/2016/09/22/eqao-message-from-our-director/