Journal of Educational Change (2020) 21:393–420
https://doi.org/10.1007/s10833-020-09380-5

Large-scale assessments and their effects: The case of mid-stakes tests in Ontario

A. Hargreaves[1,2]

Published online: 11 May 2020

Abstract

This paper analyzes the nature and perceived effects of mid-stakes testing (known as the EQAO) in Ontario, Canada. Ontario's mid-stakes tests were meant to ensure accountability and transparency, and assure system-wide improvement, while avoiding the negative effects and perverse incentives of their high-stakes counterparts. The paper provides new evidence from two projects covering almost a 10-year timespan in 10 of Ontario's 72 school districts. It shows that even though mid-stakes testing is milder in its manifestations and effects than high-stakes testing, concerns remain about the need for and side effects of such testing. The findings concern two periods of Ontario educational reform. In the first period, with a specific focus on improving performance in literacy and mathematics, administrators and special education support staff felt that the assessments raised teachers' expectations and sense of urgency leading to steady improvements in measured achievement, but that there was also evidence of negative effects, especially on paying undue attention to "bubble" students just below the threshold for minimum proficiency. In the second reform period focused on broad excellence, well-being and equity as inclusion, mid-stakes tests were perceived as having more widespread negative effects. These included teaching to the test, cultural bias, avoidance of innovation, dilemmas of whether to include highly vulnerable students in the testing process or not, and emotional ill-being among students and teachers. The paper concludes that Ontario's twentieth century system of large scale, mid-stakes assessment has not kept pace with its twenty first century commitments to deeper learning and stronger well-being.

Keywords Testing · Educational reform · Ontario

✉ A. Hargreaves
hargrean@bc.edu

[1] Lynch School of Education, Boston College, Chestnut Hill, MA, USA

[2] University of Ottawa, Ottawa, ON, Canada

Introduction

Twenty years ago, the leading article in the very first issue of the *Journal of Educational Change,* of which I was founding Editor-in-Chief, was by Michael Fullan, who invented the entire field of educational change. Fullan presented an argument that both noted and nudged forward what he called "The Return of Large Scale Reform" (Fullan 2000). Commenting on and drawing together evaluations of large-scale reform efforts in large school districts, across huge networks or *collaboratives* of schools, and within the policies of whole nations, Fullan set out 8 key lessons that could be derived from this collection of reforms. One of these lessons was the importance of combining *pressure and support* in implementing change. Pointing to research by Hill and Crévola (1999) on literacy improvement in a Catholic system in Victoria, Australia, and also on the National Literacy and Numeracy Strategy in the UK (Barber 2009), Fullan (2000: 24) argued that accountability, targets and testing were a big part of the pressure-support mix. A standards-based approach to systemic change on a significant scale, Fullan claimed, held promise if schools, districts and nations set out to

- "Determine standards and set system-wide and school-specific, year-by-year targets";
- "Focus school support services and available funds on achieving the standards and targets";
- "Put in place accountability and incentive arrangements linked to performance against standards and targets"; and
- "Conduct periodic full-cohort testing to monitor performance against the standards and targets".

Somehow, Fullan went on, the task over time was then to develop capacity in schools to develop their own approaches to and ownership of change in ways that met and meshed with the systemic priorities linked to the targets and tests. However, from the late 1990s onwards, large-scale educational reform in many countries became synonymous with only one side of Fullan's pressure-support equation, and emphasized standards, targets and what came to be known as *high-stakes testing*, within systems driven by top-down accountability or pressure.

From the beginning, this movement attracted many critics, including myself (Hargreaves 2003). But especially in the past decade, the United States, England, Australia and parts of Asia have seen escalating criticisms about the negative effects of high-stakes testing on students' learning, teachers' wellbeing, and on attempts to introduce curriculum innovation and develop global competencies.

One widely cited alternative that is meant to avoid the effects of such high stakes testing involves *testing of student samples* rather than census or full-cohort testing of whole student cohorts—as in Finland (Sahlberg 2011). Another alternative has been to use test data to inform teachers' *professional judgment* rather than as raw measures of student or school performance—as in Scotland (Reedy 2018). A third alternative to high-stakes testing and its effects is *mid-stakes testing*.

Mid-stakes testing is not prevalent in the research literature although it is popular in systems such as Ontario in Canada, Australia and (in some of its reform periods) South Korea (Lee and Kang 2019). Like high stakes testing, mid-stakes testing uses large-scale assessments to report on the progress of schools and systems, and to guide districts' and schools' improvement efforts. It also reports these results online to inform parents' choice of school and, in various ways, to compare schools' performance. However, mid-stakes testing does not come with the punitive consequences of high-stakes testing, such as firing principals and teachers, or closing low-performing schools. Writing about mid-stakes testing in South Korea, Lee and Kang (2019: 4) describe the government's move from a high-stakes to mid-stakes strategy in 2013 as one involving reduction in the number of testing points, and abolition of uses of the tests for evaluating schools.

This paper analyzes the nature and perceived effects on students and educators of mid-stakes testing in Ontario, Canada—one of the world's highest performing and most equitable education systems according to the OECD's international PISA assessments (Campbell et al. 2017; O'Grady et al. 2016). It provides new evidence from two projects covering almost a 10-year time-span (Hargreaves et al. 2012, 2018) which shows that even though mid-stakes testing is milder in its manifestations and effects than high-stakes testing, legitimate concerns remain about the need for and side effects of such testing. These concerns have not yet been expressed in most of the international research and policy analysis concerning Ontario's educational system where Ontario has largely been positively presented as an exemplary case in which "empowered educators" demonstrate how excellence and equity can be combined to raise performance and narrow achievement gaps (Campbell et al. 2017; Tucker 2019: Organization for Economic Cooperation and Development 2011). Problematic aspects of Ontario's system—in particular, the nature and side-effects of its mid-stakes testing processes—have largely been omitted from these accounts, in ways that could prove misleading for other systems interested in adopting some or all of Ontario's reform model.

No high performing systems are perfect. Singapore, South Korea, and Hong Kong, for example, have globally renowned strong teaching professions and unified governmental and societal commitments to education for public good (Tucker 2019: Organization for Economic Cooperation and Development 2011), but their high performance is also accompanied by a vast shadow system of private after-school tutoring and cramming for traditional examination and test success (Ng 2017; Bray 2006). Without careful and critical appraisals of high performing systems based on all relevant scientific evidence, there is a risk that global policy advocacy regarding these systems and their characteristics will be based on political ideology and opinion rather than on the full span of empirical research. In this respect, it is time for a critical appraisal of mid-stakes testing and how it is viewed by Ontario's educators so that a more balanced representation of Ontario's strengths and weaknesses as a high performing system is attained.

Our research in 10 of Ontario's 72 school districts has, in this respect, been both affirming of and critical about different aspects of Ontario's strategy. On the one hand, it has provided empirical support for many aspects of Ontario's approach to public education and its improvement—especially in relation to student well-being

(Hargreaves and Shirley 2018a), teacher well-being (Shirley et al. 2020), collaborative professionalism (Hargreaves 2019; Hargreaves and O'Connor 2018), and leadership from the middle (Hargreaves and Shirley 2020). On the other hand, our evidence-based appraisal of Ontario reforms and their impact has also pointed to imperfections such as potential sources of cultural bias in definitions of well-being (Hargreaves and Shirley 2018b), unsustainability of middle-level leadership strategies (Hargreaves and Shirley 2020), and educators' reluctance to acknowledge hierarchies of expertise in the development of collaborative professionalism (Hargreaves and O'Connor 2018). The problematic aspects of Ontario's reform practices have perhaps been most evident in its adoption of and continuing commitment to mid-stakes testing.

Literature review

There is nothing new about the frequently cited problems that arise from examinations and testing. In his 1922 classic text on *Economy and Society*, German sociologist, Max Weber (1968, p. 999), pointed out that examinations or tests of specialized expertise, such as those that were systematically used in China beginning more than a Millenium ago, were indispensible to modern bureaucracies. They ensured that appointments and promotions in state agencies would be based on objective merit rather than on claims to noble lineage and patronage.

However, while one-time selective examinations are gateways to success, they are also drawbridges of failure that can shut out the rejected from citadels of opportunity and recognition. For these reasons, educational assessment and policy experts have described how selective examinations and testing can have negative "*backwash*" effects on teaching and learning. These effects included failing to address or achieve deeper learning outcomes that were not easily examined (Biggs 2003), diverting teaching and learning time to rehearsing examination questions or upcoming continuous assessments (Newton 2008), and overemphasizing evanescent knowledge that is easily memorized and quickly forgotten (Board of Education 1911).

From the 1980s onwards, in a number of countries, beginning with the US and UK, one-time examinations and tests began to take on another function as well as selection and certification for employment and higher education. Assessments were created, supported, and expanded not just to pass judgments on individuals, but also to deliberate on the effectiveness of whole schools and entire educational systems (Goldstein 2001). This is what happened to the General Certificate of Secondary Education (GCSE) examination in England, for example, where the percentage of students gaining Grades C and above was used as one crucial metric of whether schools were deemed to be failing, required intervention, or even warranted closure altogether so they could be forcibly replaced by semi-private academies.

In other cases, tests were introduced—especially for children who had not yet reached adolescence—that had no selection or certification value for individuals at all. This occurred in elementary and primary schools in the US and England respectively, and also overseas in Singapore, Hong Kong, and Australia. The tests were designed to hold schools and systems responsible for public and political

accountability in meet learning standards (Smith and O'Day 1990; Darling Hammond 2004). If schools failed to meet mandated benchmarks or targets for proficiency, then, in the US and UK at any rate, severe consequences could follow—replacement of the principal, firing some or all of the staff, takeover by the state or an entity designated by the state such as a private company, or complete closure to enable replacement by a totally different school. For these reasons, large-scale assessments in England and many parts of the US came to be known as being *high-stakes* in character.

Nowhere was high-stakes testing more evident than in the public policy strategies of the Blair government in the UK. The introduction of performance targets into educational improvement and accountability processes underpinning the UK's National Literacy and Numeracy Strategy led to criticisms that not only did this divert the energy of teachers and schools away from their core practices of quality teaching and learning, but it also produced results by cheating and fabrication (Tymms 2004). According to the UK Royal Statistical Society, the resulting rates of alleged improvement in achievement scores were, in any authentic sense, statistically improbable (Bird et al. 2005). These metrics of failure and success were and still are also used as pretexts to identify and close down so-called failing schools so they could be reopened as semi-privately sponsored and controlled academies (Hargreaves and Harris 2011).

These patterns repeated themselves in the US where the legislation of *No Child Left Behind* (NCLB) and then *Race to the Top* placed public schools and their teachers on a schedule of Adequate Yearly Progress to reach eventual full proficiency for all students. As the targets became increasingly unrealistic, schools were closed down, and there was an epidemic of system-wide cheating (Severson 2011).

The use of student tests for accountability purposes to monitor and report on the performance of whole institutions and systems, as well as their teachers, has led to critiques of high stakes testing in books (Koretz 2017; Ravitch 2011; Sahlberg 2011; Kohn 2015; Zhao 2018; Hargreaves 2003; Hargreaves and Shirley 2012), peer reviewed journal articles (Booher-Jennings 2005; Braun 2015; Goldstein, 2001; Baker and Foote 2006; Baker et al. 2013; Falk and Drayton 2004; Daly 2009; Darling Hammond 2004; Fuhrman and Elmore 2004), professional magazines and blogs (Rothstein 2014; Fullan 2011), the national statistical societies or associations of both the UK (Bird et al. 2005) and US (American Statistical Association 2014), as well as the American Educational Research Association (2000).

These critiques refer to what Donald Campbell (1976) famously called the *perverse incentives* of high stakes testing. These include narrowing the curriculum; teaching to the test; lowering of teacher morale and retention rates; allocation of disproportionate time and effort to students just below the required levels of proficiency at the expense of those with more severe needs; expulsion and exclusion of second language students or students with special needs judged unlikely to perform well on the tests; diversion of professional learning communities and improvement efforts onto analyses and actions that might secure quick gains in measured performance; bullying of staff by administrators in cultures with high degrees of threat; explosions of after-school tutoring in entire shadow systems of education; and escalating levels of anxiety and even suicide among adolescents taking the tests.

In the face of these well-documented problems, some systems, like Wales and Alberta, pulled away from high-stakes tests. Others, like Singapore (Ng 2017) and Hong Kong (Chiu 2018), are reducing the prevalence of these tests by delaying the grades in which they are administered, by making participation voluntary among younger children, or by testing samples rather than entire student cohorts, for example. The 2015 US *Every Student Succeeds Act* withdrew Federal pressure on states to implement high stakes testing (US Department of Education 2015), and states like California have been developing broad data dashboards instead (Torlakson 2015; California Department of Education 2016). A number of other systems, like Finland, avoided being drawn into high stakes testing in the first place (Sahlberg 2011). Meanwhile, systems like Ontario adopted another solution—*mid-stakes tests*.

In mid-stakes tests, system-wide tests are still administered to check how schools and systems are performing and to determine whether intervention is required, but the consequences are less severe and life changing than in a high stakes system (Lee and Kang 2019). In the case of Ontario, the presence of mid stakes is evident in the fact that test results have been made known to the public as a basis for parental choice of school, and also to district directors and superintendents who can use them to inform and justify prompt interventions. In particular, data-driven improvements through 6-week teaching–learning cycles of performance review involving interim assessments have identified students who were yellow (at risk of not achieving proficiency targets) or red (performing far below proficiency) rather than green (at or above proficiency) so that just-in-time strategies could be used to raise performance.

At the same time, Ontario's mid-stakes tests do not possess the high-stakes of punitive systems. Principals and teachers are not hastily replaced, schools are not closed down, and there are no semi-private systems of academies or charter schools waiting in the wings to take over from traditional public schools when results fall short.

These mid-stakes tests in Ontario were meant to ensure accountability and transparency, and assure system-wide improvement, while avoiding the negative effects and perverse incentives of their high-stakes counterparts. Instead of negative pressure, there has been positive pressure and support. Instead of being punished, struggling schools or districts have received assistance in an environment of "non-punitive accountability" (Fullan 2012). Has this process of *mid-stakes* testing avoided the perverse consequences that have afflicted schools in high-stakes environments elsewhere?

The Ontario policy context of mid-stakes testing: 2004–2011

Building upon findings presented in two detailed reports (Hargreaves et al. 2012, 2018) of collaborative research with 10 out of 72 Ontario school districts, this article examines the perceived impact and consequences of a mid-stakes test known as the EQAO (*Education Quality and Accountability Office*) on students, teachers, and schools in the province. Its original contribution is to advance a new line of interpretation that synthesizes the evidence from Ontario for a reconsideration of mid-stakes testing and its effects as a policy strategy. Especially given the prominence

and popularity of Ontario as a system that is frequently cited by global policy organizations such as the OECD (2011) and the US National Center for Education and the Economy (Tucker 2011) as one of a very small number of educational systems that are both high performing and quite equitable, and that hosts numerous teams of educators and policy makers from overseas, a critical, balanced and evidence-based appraisal is long overdue.

The EQAO is an arms-length, quasi-autonomous agency established by the Ontario Government. Following the recommendations of a Royal Commission on Learning in 1995, it was set up under a new Progressive Conservative government in 1996 (Royal Commission on Learning 1995). It developed and implemented large-scale assessments in reading, writing and mathematics in Grade 3 from 1996 to 1997, in Grade 6 from 1998 to 1999, in Grade 9 mathematics from 2001, and in the form of a secondary school certificate from 2002. The original mandate and purpose of EQAO in the Educational Quality and Accountability Office Act of 1996 was unambiguously one of public reporting of assessment results to ensure accountability, "to evaluate the quality and effectiveness of elementary and secondary school education", "to evaluate the public accountability of boards", and to report the results and implications of the tests "to the public and to the Minister" (Campbell et al. 2018: 43).

In 2003, a Liberal Government was elected in Ontario. Following the example of Prime Minister Tony Blair in the UK, a central plank of its platform was reforming and raising standards in education. This policy direction followed a period of Progressive Conservative Government that had been characterized by top-down educational change, hastily imposed curriculum reforms, and a climate of blaming and shaming the teaching profession for alleged educational failures (Hargreaves 2003).

Michael Fullan was soon approached by the new Government to become the Special Adviser on Education to the Premier. Fullan had led an evaluation of the Literacy and Numeracy Strategy in the UK. An article on Fullan's appraisal of it and its implications for Canadians was published in the nation's leading current affairs magazine, *MacLeans* (Schofield 2001).

With Fullan's guidance, the province developed a strategy of tri-level reform that Fullan had been advancing from the late 1990s (Fullan 2010) and that was consistent with his vision of large-scale reform in the inaugural issue of the *Journal of Educational Change*. At the top, following a strategic concept developed by Harvard Business professor John Kotter (1996), the overall vision and direction of policy was built by a *guiding coalition* of government leaders in partnership with key stakeholders such as unions, school boards and business interests. Closely following the UK's Literacy and Numeracy strategy, two priorities were those of literacy and numeracy achievement where the key goals were to "raise the bar" and "narrow the gap" in measured achievement. In addition, there was a determination to improve high school graduation rates and restore public confidence in education—all within one election term.

Again, following the UK's example, a system-wide target was set in which 75% of students would meet or exceed a Level 3 threshold cut-score for proficiency. Among educators, this focus became known, colloquially, as the "Drive to 75". In explicit contrast with the UK model, though, the province pledged and provided significant

support to the middle and lower levels in the tri-level model. This took the shape of training, coaching and resource materials for literacy improvement for teachers at the bottom level, and significant investments in the middle levels of $25 million dollars to each of three education groups: the teacher unions to provide professional development, principals' organizations to further initiatives for student success, and school district directors to develop and implement a strategy for inclusive special education reform.

The implication for EQAO testing was that its large-scale assessments that were initially designed for accountability purposes assumed improvement and intervention functions (Campbell et al. 2018). Depending on their results, schools and school districts could be given notice to improve and to receive various forms of intervention and assistance. Although these interventions were not as high stakes as in the US or England, they could and did lead to considerable effort and focussing of attention on raising achievement scores. This was especially the case for students and schools that were falling short of measured proficiency or that were failing to improve high school graduation rates.

In particular, in schools and districts, EQAO test scores were combined with other indicators of student and school progress to guide data-driven improvements in 6-week teaching–learning cycles. Adopting this UK strategy that first originated in Australian research and development work by Peter Hill and Fullan's future co-author, Carmel Crevola, (Hill and Crévola 1999), teachers would track student progress during each cycle, often on a weekly basis, identifying students making satisfactory progress as green, and those requiring inquiry and possible intervention as red or yellow. In the latter cases, rectifying action would be initiated to raise achievement in real-time. EQAO and other data were also used to compare schools in similar circumstances but with different performance levels over 1-year and 3-year periods to enable struggling schools to locate and leverage peer support as needed.

In the coming years, outside EQAO or the Government's own educational research department, little evidence surfaced about the side-effects of EQAO testing on Ontario's educators and their students. The collection and presentation of our data, embedded in a wider study of the impact and implications of special education reform, was thus of particular interest (Hargreaves et al. 2012).

My own positionality on educational change and assessment circa 2000

At this point, it is probably helpful to provide some detail regarding my own positionality on educational change and assessment around the time I founded the *Journal of Educational Change* and in the lead up to Ontario's educational reform strategy in the early 2000s, that was consistent with Fullan's journal paper in 2000.

During this period, I was deliberately constructing a platform to support and advance educational change as an open and independent field of study and action. In 1995, with my colleague Lorna Earl, I co-founded the International Centre for Educational Change at the Ontario Institute for Studies in Education, as a group of researchers and developers, (including Michael Fullan) with a mission "to investigate, initiate, support and speak out with integrity and authority on changes and

reforms in education…across the world". In 1997, I edited the annual Yearbook for the Association of Supervision and Curriculum Development (ASCD) on *Rethinking Educational Change with Heart and* Mind (Hargreaves 1997), and a year later, as lead editor of a team of four, also including Michael Fullan, published the double-volume *International Handbook of Educational Change* (Hargreaves et al. 1998).

This strategic work arose out of and in parallel to a programme of educational research on educational change and its effects. In the mid 1990s, a series of papers, culminating in an eventual book, *Learning to Change*, documented the pilot project efforts of Ontario's only socialist government in history to implement reforms for early adolescents in the early 1990s that were directed at broad learning outcomes rather than narrow and highly prescriptive learning standards, at organizing the curriculum around interdisciplinary projects rather than traditional school subjects, and at developing alternative forms of assessment in the form of portfolio and performance-based assessments. Two outcomes of our research were significant at this time.

First, inconsistency in implementation resulted from lack of clarity about how outcomes and projects should be interpreted. This was a result of failure to invest in leadership development and collaborative teacher cultures that would have been able to make collective and local sense of the outcomes and frameworks, school by school. Second, given the prospect that these and similar efforts at implementing broad outcomes might fail, hovering in the background in places like the US and UK, was a wider educational reform movement which I called the *New Orthodoxy of Educational Change* (Hargreaves et al. 2001). Later renamed by Pasi Sahlberg (2011) as the *Global Educational Reform Movement* or GERM, this movement or orthodoxy, I argued, was becoming characterized by specific standards rather than broad outcomes, standardization and prescription of content, a narrow focus on literacy and mathematics achievement, and increased testing and accountability.

A second project funded by *The Spencer Foundation* from 1998 and co-directed with Ivor Goodson who was then at the University of Rochester in the US, examined not what happened to the implementation of particular reforms (where most educational change research was concentrating its energies, especially as it followed Federal dollars in the US), but on how over 300 teachers in 8 US and Canadian secondary schools had experienced multiple educational changes over 30 years (Hargreaves and Goodson 2006). These three decades included the most current one within the study which addressed and analysed the impact on teachers of the onset of the Regents Examinations in New York State, and the reform agenda of a Progressive Conservative Government in Ontario in the second half of the 1990s through to the early 2000s.

The results were recorded in two chapters of my 2003 book *Teaching In The Knowledge Society* (Hargreaves 2003). In New York State, the creation of a new magnet school, led to rapidly declining performance in its non-magnet neighbour. The introduction of increased examination credit requirements led to losses of curriculum choice, standardization of content, and diminished capability among teachers to respond to student diversity. Students became too test conscious and teachers found it harder and harder to connect with students' interests. Students with special needs and language learning difficulties were especially disadvantaged when they

were required to take the tests. The resulting effect on teachers was demoralization and burnout.

In Ontario, the five schools in the Spencer study combined with four more in an on-going improvement project we were conducting, revealed the impact of the new reform and testing policies on teachers. Survey responses from 480 teachers in the seven schools produced results including the following. Only 20% of teachers felt that a new Grade 10 literacy test promoted student improvement. Nine out of ten teachers believed the test did not motivate the students; nor did it increase their own confidence as teachers. Fewer than a quarter of teachers supported the tests or felt they made them more accountable as teachers. I concluded: "On this evidence, systemwide testing conducted on a census rather than a sample basis does not help, and in some ways actively hinders, teachers in supporting their students to learn in a knowledge society" (Hargreaves et al. 2001, pp. 100–101).

Given the weight of this prior evidence, and the 2003 Liberal Government's decision to persist with and indeed expand EQAO testing as part of its own reform program, along with Fullan's own declared commitment to census or full-cohort testing as reported in his 2000 paper and elsewhere, my own early response, with my colleague and co-author Dean Fink, was to predict that a target-driven culture would lead to the perverse incentives and strategies of gaming the system that had, by then, been exposed in the UK (Hargreaves and Fink 2005). In response, Fullan and other architects of the reform insisted theirs was a made-in-Ontario solution that provided far more support to professionals and avoided the punitive sanctions for failing to meet targets that had characterized the UK. We debated these differences of opinion vociferously, including in public, over the course of the next few years.

Meanwhile, with specific reference to EQAO, assessment specialists issued early warnings that using large scale assessments for both improvement and accountability purposes could prove problematic (Earl and Torrance 2000) and that "the proliferation of purposes" could "compromise how well EQAO's assessments can accomplish any one of them" (Wolfe et al. 2004, p. 5). The long-standing concern that data collected for one purpose should not be used for another, was casting a shadow over the future use and effects of EQAO data.

The perceived effects of mid-stakes testing: 2008–2011

In 2008, my colleague, Henry Braun, and I, were approached by the Ontario Council of Directors in Education (CODE) to study the special education strategy they had been charged with implementing by the Ontario Government. The Directors called their strategy *Essential for Some, Good for All (ESGA)*, based on their guiding belief that the practices that were essential for some students who had disabilities or other special needs, were also good for all students.

The province's leaders in central government realized that a special education strategy would be complicated and not amenable to the reform model that it had adopted to improve literacy and mathematics achievement in the mainstream. It allocated $25 million to CODE to take responsibility for this reform, conditional on the resources being spent to improve achievement outcomes and narrow achievement

gaps in ways that were consistent with the province's broad special education philosophy that was built around the principles of *Universal Design for Learning* (Meyer et al. 2013). The context and methodology are described fully in the project report (Hargreaves et al. 2012). The research team worked with a self-selected but also representative sample of 10 Ontario school districts to interpret and explain the core principles and practices that had been adopted in ESGA. The research design employed a mixed-methods approach with both quantitative and qualitative components.

A survey was conducted of self-reported perceptions and practices related to ESGA among a sample of school principals, teachers and special education support staff in nine of the 10 districts. This survey included items about educators' uses and perceptions of EQAO and other assessments. The web survey was conducted in order to collect evidence from a larger sample of teachers and other school professionals beyond the schools from which data were collected as part of each school district site visit. Schools asked to participate in the survey were ones that had been involved to some extent in ESGA.

There were two main response formats. One was a standard Likert scale, usually having 5 choices: 1 = Strongly Disagree; 2 = Disagree; 3 = Neutral; 4 == Agree; 5 = Strongly Agree. The survey instrument also had open-ended questions covering a broad range of issues, including uses of assessment. These open-ended questions gave respondents opportunities to make more extended statements about their perceptions of the positive and negative effects of ESGA, including the impact of EQAO.

Overall, survey respondents registered disagreement with the statement that EQAO results provided an accurate measure of academic competencies (mean = 2.85) and strongest agreement with the statement that EQAO results were "*not* an appropriate measure of what students with special needs know and can do" (mean = 4.04). There was modest agreement that district concerns with EQAO results were driving too much practice and yet disagreement that those concerns distracted teachers from helping the students who needed them the most (mean = 2.77). Clearly, many respondents had concerns regarding EQAO and its impact, yet most still asserted that it did not influence their own allocation of effort.

Open-ended questions addressing EQAO elicited considerable comments that complemented the closed survey responses. A number of respondents indicated that they found EQAO results useful, including for students with special needs. As one teacher put it, "it gives me somewhat of a focus on areas that are lacking, such as problem solving in math, inferring in language etc." Such teachers saw EQAO results as just one source of evidence among many that could help them in crafting their strategies to help all students:

> The Board (District)-level focus on EQAO results actually helps me compare my identified students to all students that have written the EQAO. We try to move our level 2 students to level 3, 3 to 4, etc. Teachers are trying to identify the gaps and try to close it. It is no different for our identified students. The individual EQAO results are considered one 'piece of the puzzle' because we also observe other informal testing.

When asked to comment on negative effects of EQAO, teachers expressed a number of concerns. "What is frustrating is special needs students are given many accommodations to succeed in the classroom, but when tested on the EQAO, nothing is put in place to support them", one teacher remarked. Another commented:

> I find that because EQAO is a paper and pencil test, it does not truly assess and evaluate our special needs students. These students are not able to show what they really do know. Throughout the year, these students have various choices for their assessments and evaluation and usually it is not a paper and pencil task.

Mirroring the findings of the closed ended questions concerning EQAO, some respondents remarked on the undue importance that was placed on it.

> I personally think the District and Administrators put too much emphasis on EQAO results. I teach what my students need to know (whether mainstream, accommodated, or modified) based on curriculum and in a way that meets their needs. If that helps them in EQAO, great, if not, there are more important lessons to learn than getting a good mark on EQAO.

However, while some felt that the Drive to 75 had led to distorted priorities, in closed-ended responses, educators indicated that this had not ultimately affected their own practices.

In summary, survey respondents were in general agreement that there was too much attention paid to EQAO results and, that for many students with special needs it was not an appropriate instrument for determining what they could accomplish. At the same time, a number of teachers indicated that the EQAO did not impact their own day-to-day teaching. Moreover, special education resource teachers, and administrators at the school and system levels, were likely to state that they found EQAO results helpful in pointing to areas and students needing more attention.

Qualitative interview data from districts where ESGA projects and initiatives involved or backed on to the grades (3 and 6) where the tests took place yielded further insights into educators' perceptions of assessment in general and of EQAO assessments in particular, and their effects. Full details of the data are provided in the project report (Hargreaves et al. 2012). A brief summary is presented here.

As in the survey data, there was evidence to support the role that EQAO played in promoting improvement. System administrators valued EQAO because it enabled them to monitor progress in their schools and to have data that could justify action and intervention. EQAO was also valued by many principals and by special education resource teachers who felt the data pushed classroom teachers to take more responsibility and set specific goals for all their students, including those with learning disabilities. This occurred when teachers had to review results with their colleagues and to incorporate what they learned into their teaching–learning cycles. The focus on data, including EQAO data, had "increased the capacity of all teachers working with special needs students", one respondent reflected, by encouraging teachers to set "specific, measurable and attainable goals" for these students.

Classroom teachers were more critical of EQAO because they felt their efforts were being diverted by the Ministry and by school and system administrators towards students just below the Level 3 borderline (between 2.7 and 2.9) whose scores were treated as more important and having greater priority than those of other students down at Level 1 or 2—too far from the point of proficiency to benefit the overall scores from a quick lift or intervention. One principal's office had a chart calculating the percentage of students at Level 3 and those at Levels 2.7–2.9 who might be quickly moved up to 3 with concentrated and intense effort. As one Ministry official discreetly confided, getting students up to functional literacy at Level 2 would do nothing to help the government meet its stated targets for Level 3 proficiency. Even if the distortions were not so egregious as this, schools still had to "interrupt the flow of learning" to complete the tests which, by the time results were returned, months later, were no longer of value for helping the students being tested who had normally moved on to other classes, teachers and schools.

The results and their implications concerning EQAO assessments derived from this first phase of our research are therefore mixed. Advocates of mid-stakes testing who were eager to find an assessment system that would achieve systemic goals of improvement and accountability without the negative side effects of high-stakes systems like England's could point to evidence supporting their use. The tests received support from administrators and from special education resource teachers. They developed a sense of urgency about expectations and equity by all teachers for all students. They stimulated data-driven collaborative inquiry devoted to diagnosing student difficulties and making interventions in real time. They also enabled system leaders to discharge their responsibilities properly as a result of knowing how the students they served were or were not performing. Teachers and some administrators complained about various aspects of the tests, but none admitted that this adversely affected their own instructional practice. And at the end of the day, the Drive to 75 increased proficiency levels to around 70% by 2011. These new levels were up from 54% at the beginning of the drive in 2004 and were accomplished at gradual rates that were statistically defensible and educationally sustainable, compared to questionably rapid gains in schools and systems in England and the US.

At the same time, the project data lent support to critics of EQAO. Evidence pointed to perceived distortions of and interruptions to the flow of learning. There were diversions of disproportionate efforts on to what US researchers have called "bubble" students whose scores were just below the targeted levels of proficiency at the expense of students with more serious learning needs (Booher-Jennings 2005). Mid-stakes tests, according to these data, therefore still seemed vulnerable to a number of the same negative and distorting side effects of their high-stakes counterparts. This was despite the fact that they also attained greater success in achieving authentic and sustainable gains, and in securing the support of administrators and special educators as a way to stimulate and focus all teachers efforts on all their students' achievement.

Given credible cases both for and against Ontario's mid-stakes testing in this first period focused on raising attainment in literacy and numeracy (mathematics), the judgment about whether the gains justified the collateral damage of negative side effects is ultimately an important ethical, professional and political one. It cannot

be decided by the evidence alone, but only by weighing off the clear gains against unintended negative consequences that nonetheless must be taken seriously in their consequences for students and the cultures of schools. However, when the goals for Ontario's schools as established by the Ministry of Education became more complex after 2013, and the administration of EQAO became increasingly embedded into the system, this balance of evidence regarding the large-scale assessment of EQAO and its combined use for accountability and targeted improvement purposes started to shift. This called for even greater attention to EQAO's problematic aspects.

The Ontario policy context of mid-stakes testing: 2014–2018

In 2013, although the Government of Ontario remained Liberal, the Premier stepped down and was succeeded by his previous Minister of Education, Kathleen Wynne. The succession signaled a shift in policy and strategy. With a background of graduate study in Canadian First Nations languages, and given her concerns about bullying and marginalization of LGBTQ and other vulnerable groups, Wynne's education policy took the province in new directions.

Achieving Excellence, released in 2014, set out four priorities (Ontario Ministry of Education 2014). One was to maintain public confidence in a system that had raised student performance on EQAO results by 17 percentage points over the previous decade. As the report's title made clear, excellence was also still a clear priority. Partly this would take the form of continuing the movement towards 75% proficiency levels on EQAO assessments, with a particular focus on mathematics, where results had flat-lined and were starting to fall. Other areas of excellence were now added to literacy and mathematics as priorities, like the Arts and STEM. Equity was a third priority alongside excellence and public confidence, but it was no longer interpreted just as narrowing achievement gaps. Equity now encompassed inclusion of diverse and vulnerable groups and their identities, such as indigenous, refugee and LGBTQ students, so they could see themselves, their communities, and their needs reflected in the life and learning of their schools. The fourth pillar of Ontario education reform was well-being. The province's mission statement proclaimed that "Ontario is committed to the success and *well-being* of every student and child".

No changes were proposed to EQAO or large-scale assessments, generally, though. Elsewhere, outside Ontario, however, reservations about the limits and risks of top-down accountability had been deepening. Harvard Professor Richard Elmore (2004) and Stanford Professor Linda Darling Hammond (2004) advocated lateral, internal, professional accountability over vertical, external, bureaucratic accountability. Finnish expert Pasi Sahlberg (2011) identified accountability and testing as one of the dangerous elements of the Global Education Reform Movement (GERM) that was sweeping the world. In *The Fourth Way*, Dennis Shirley and I argued that accountability should be the small remainder that would be left after responsibility had been subtracted (Hargreaves and Shirley 2009).

Previous advocates of testing and accountability as a way to maintain public confidence were also beginning to raise questions about top-down accountability, including Michael Fullan himself. In his highly cited and widely

used paper "Choosing the Wrong Drivers for Whole System Reform, Fullan (2011) had described a wrong driver as "a deliberate policy force that has little chance of achieving the desired result" (p. 3). One of his four wrong drivers was accountability with its use of test results to reward and punish schools and teachers. Accountability, he said, "assumes that educators will respond to these prods by putting in the effort to make the necessary changes" (p. 8). But, he continued, "leading with accountability is not the best way to get accountability" because it "does not highlight the instructional improvement to bring about needed changes" (p. 8). Fullan retained the view that systems could "do testing, but less of it" (p. 9), but should mainly use assessment to drive improvement, not to enforce accountability.

Ontario received barely a mention in Fullan's paper except in support of his argument promoting capacity building. It was therefore not clear whether Ontario's EQAO was regarded as part of the right drivers of capacity building because it had "less" testing compared to England and the US, or as a wrong-driver of top-down accountability. The evidence of our first phase of research on Ontario's mid stakes testing system (Hargreaves et al. 2012) had been mixed, and in the absence of any new research, the balance of benefits and negative effects, of right and wrong drivers in Ontario and embedded within EQAO and its uses, had yet to be determined.

The EQAO was now also operating in a provincial and global context of policy goals that were more complex than simply raising results in literacy and mathematics or increasing graduation rates. Like a number of other systems, Ontario was moving towards a twenty first century curriculum in inquiry-based learning, new pedagogies for deep learning, technology assisted instruction, and attention to children's overall well-being—all within a province of manifest diversity. How did its large-scale assessment system of EQAO fit with these developments? Moreover, with evidence mounting about a global crisis in the teaching profession, the intellectual climate surrounding the value of top-down accountability and testing was also changing rapidly. What were EQAO's uses and consequences within this rapidly changing context?

The perceived effects of mid-stakes testing: 2014–2018

In 2014, our work with the Consortium of 10 Ontario school districts (9 of which were the same as before) entered a second phase. The project's earlier phase had identified a strategy of *Leading from the Middle* (LfM) where school districts worked together in conditions of transparency and collective responsibility to support all students' success. In this second phase, the Consortium wanted the research team to undertake a second period of research, to explore with them how they were using LfM to pursue projects in their districts such as developing number sense in young children, or implementing programs of emotional self-regulation, that they had initiated in their districts and that were connected to the Ministry's four pillars. The study's goals were developed with due consideration for the Ministry of Education's priorities in *Achieving Excellence*. The agreed purpose was to "gather perceptions of the projects' strengths and weaknesses."

In May 2016, our research team conducted site visits to all 10 of the Consortium's districts across the province. At least two team members visited each district. Team membership was mixed and rotated among the 5 members in order to enhance cross-validation of interpretation. The research team conducted interview-based mini-case studies over 1–2 days with each school district. We interviewed 222 educators, selected project leaders, and project coordinators at the district and Ministry level. Interviews lasted approximately 1 h each and were conducted in private locations in each district office or school building.

As in the earlier study, our almost identical sample was representative of districts across the province in terms of geographical spread, urban and rural distribution, religious and non-religious schools, and standardized test scores. The districts volunteered to participate in the research and to provide funding through CODE. After initial coding, the team wrote individual case studies of 5000–10,000 words each for internal use only, describing findings from each individual district, based on the themes and also on the emerging narratives inherent to each district.

Our research includes evidence about the impact of EQAO assessments from half of the 10 boards. Where district projects did not include grades in which EQAO assessments were administered, educators were less likely to mention the assessment as a factor that impacted learning, achievement and wellbeing. We had less access to teachers compared to educators in administrative roles than we did in 2011. Given that in Phase 1, teachers were more critical of EQAO than other educators, this suggests that our findings in Phase 2 probably underestimate educators' perceptions of the negative side effects of EQAO. There was also no survey of educators' perceptions of factors affecting their work, including large-scale assessment factors, as there was in 2011. For these reasons, our findings on EQAO, like our other findings, cannot be generalized to the whole system, or compared exactly with the previous phase of our research with the Consortium.

As in 2011, senior administrators felt that EQAO "helped with accountability" and "helped drive standards," in one director's words. A Superintendent of Special Education concurred, asserting that it had a "place in terms of accountability." Improvements in test scores gave senior staff confidence they were moving in a positive direction. A member of the professional services staff in one board proudly referred to the success of an early literacy project in terms of how "the EQAO scores have steadily gone up and maintained."

Again, as in 2011, some principals were also supportive of the value of EQAO. They were proud when they showed gains on EQAO literacy scores for students with learning disabilities. EQAO also provided them with a way to know their students. One said,

> When I was teaching, it was certain I would have been saying "No" but, as an administrator, I see the need for it because it's that piece of the puzzle that you wouldn't have had if you didn't have that data. You can certainly see those students and where they're struggling. Sometimes you didn't even know they were struggling until you have that data.

A principal in another board felt that EQAO "drives the conversations." Another principal commented that her team was "feeling pumped and excited. They think that they're going to see some improvements."

Facing a new context in which more complex skills were being sought among students, a system administrator wrestled with the pros and cons of maintaining the assessment.

> Is it the perfect way to measure that? No, but can a standard like that drive the way that I might teach better and help kids be clear around expressing their thoughts? I don't know. I don't think it's a bad thing. The fact that it causes anxiety for kids or that our kids with language, ESL students - that our kids with learning disabilities - are challenged by that; does that bother me? Yes, it does. It bothers me immensely, but I don't know a better way.

Alongside the persistence of arguments in favor of EQAO's role in monitoring system-wide performance and improvement as a basis for intervention, there were also signs that the link of EQAO to proficiency targets was no longer leading districts to focus unduly on "bubble" students. Data that are missing are as important as data that confirm or disconfirm previous findings, and in this respect, this second phase of research yielded no evidence of schools concentrating on children just below proficiency to lift their scores quickly. Indeed, there were indications of a positive shift in this strategy. For instance, the two districts in our study that shared evidence of the work they were doing as part of the province-wide push to improve mathematics achievement showed no evidence of trying to make quick gains by focusing on students coming up to Grade 3 and Grade 6 tests. Rather, they concentrated on building confidence and competence in mathematics and mathematical learning among their elementary teachers by providing additional professional development in line with the province's model of "collaborative professionalism." (Hargreaves and O'Connor 2018).

Back in 2011, our report on one large district observed a pervasive "case management system" of individual students that made "more systematic use of diagnostic assessments and a strategy of tiered interventions." One early years/special student services consultant recalled that in this system, problematic students were described as *marker students*—students who were just below the point of proficiency—a designation that also appeared in Ministry documents at the time.

By contrast, in this second period of reform, the students who were chosen as a focus for collaborative inquiry were identified as "students of mystery" in Ontario Ministry of Education language, or "students of wonder", as at least one district expressed it. In the words of an early childhood/student services consultant, "Last year we called it a *marker student*. We changed it very consciously to a *student of wonder* this year." This is a student that "doesn't likely have a diagnosis of anything but that they [the teachers] have questions and wonderings about." "We went through a process of looking at strengths" in addition to "areas of need.'"

Gone was collaboration that focused mainly on traditional achievement and on using quantitative data to identity *marker students* with weaknesses that could be rectified. Educators now used a range of data and perspectives to illuminate the strengths as well as weaknesses of students who struggled in some way. Educators

inquired together, drawing on all of the evidence at their disposal, including what they had gathered with their colleagues, rather than simply using provincial testing data to drive everything forward.

While some administrators were continuing to support EQAO, and educators were able to rise above the previous tendencies to adopt short-term fixes that would raise the scores, by 2016, there was less support elsewhere for EQAO other than one teacher who phlegmatically conceded that "it's just part of my job; we just do it." When asked whether the test should be continued, teachers indicated that the cons outweighed the pros. In one teacher's words, "I don't think I could think of a teacher that would say, 'Well, no. We need to keep it. It's so useful and great.'"

Several problems with the test were mentioned in terms of its negative side effects for student learning and well-being.

First, was *cultural bias*. Test questions that had culturally specific content, such as those about Canada's internationally famous ice hockey star, Wayne Gretzky, might have been unknown to newcomers from the Global South. Items that referred to winter vacations in warm places might have made little sense to children from low-income families. Even efforts to represent greater cultural diversity like including an item we saw students responding to on Tai Kwon Do, could have had little meaning for some Indigenous or refugee students.

One teacher added that testing items did not account for the differences in verb usage or other grammatical constructions used in Indigenous languages. Educators in the same district also noted that some students were more expressive when allowed to type responses, but were not able to do this on the test when they had to write their answers longhand.

Second, educators were concerned about students, such as recently arrived immigrants with language or trauma issues, or ones with autism spectrum disorders, who had no chance of succeeding on the test, yet whose scores would be counted in the school's final profile. A coordinator explained: "They don't report on the participating students. They report on *all* students. The kids with developmental disabilities who do not write are still in the denominator. Students who don't write the test and who are exempt are then given a zero". One teacher was worried about fairness and equity.

> I have Grade 3 and Grade 6 students that are non-verbal and autistic, that there's no way, shape, or form, can write that test. It's ridiculous that they would even get on the list. It doesn't take into consideration the poverty in my school. It doesn't take into consideration the (child services) involvement, the families that are living in motels. All those things that set my families back are not even considered. It's hugely detrimental to my kids when we get into those scenarios. It's very stressful for them. It's very stressful for the teachers. And, quite frankly, it seems to be unfair.

Third, the standardized tests incurred excess instructional time devoted to *test preparation* rather than new learning. Even though one director suggested that neither he nor his teachers should perseverate on the EQAO, because the most important thing was the focused time on learning, every year they still moved the desks into rows. In a class where EQAO practice had been part of the school's

weekly routine from the start of the school year, students were reluctantly redoing a practice test on reading comprehension from the previous day because many of them had performed poorly. The principal stated that the purpose of this school-wide practice was to get students accustomed to testing.

However, some educators were quite content to focus on test preparation. One district leader said that

> EQAO really sets the bar. I find when you put grade 3 questions on the table in front of a group of primary teachers, K to 3… "Let's talk about, as a community here, how can we support the grade 3 teachers in the building. This isn't about one year captured on a test. This is an accumulation of the years". We started to talk about what are the things that you can do in Grade 4 to support your Grade 6 teacher? We talked about doing daily, if not weekly, multiple-choice experiences in your room so that the children learn the strategies to conquer those types of questions with ease. I think EQAO has been a driving force.

A teacher of Grade 2 in the same district was also aware of what she needed to do to prepare herself and her students for the test in Grade 3. For her, success on the EQAO was equivalent to success in general:

> I did give them a question from the EQAO because I have a couple of the Grade 2s, just to see how they did. Then I sat with them and looked at what were the barriers. Was it the language? Was it the vocab? Was it that it was written? Was it that they had to communicate it? That they had to write it in that box? I made notes as they did it. To me, that helped me understand maybe what I need to do next year to be able to have them be successful.

Fourth, constraints of large-scale assessment sometimes adversely affected *efforts to innovate*. Two districts reported undertaking significant innovations. One participated in Michael Fullan's *New Pedagogies for Deep Learning* (NPDL) network (Fullan and Quinn 2015). But its promotion of innovation was regarded as being at odds with the demands and constraints of the EQAO. "I feel like EQAO is preparing students for a very antiquated version of education," one grade 3 teacher said. A grade 5 teacher agreed. "The standardized testing is so far removed from what we're doing. There's nothing standard about what we're doing. We're taking each child where they come from," they said. "All the things that we establish in our classrooms, the accommodations, all the tools that we give our students cannot be used on EQAO," another grade 5 teacher observed. "I do have EQAO pending as a grade third teacher," another teacher in the same focus group added. "I do have content that I'm expected to teach and assess. Hopefully some of the critical thinking skills would come through when the students are presented with a pencil-paper test for three days in a row. There's a complete disconnect."

Teachers outside the EQAO tested years did not experience the pressures and constraints of EQAO to the same degree. In the districts that showed us projects in the early years, for example, EQAO was never raised as an issue. When

teachers moved out of Grades 3 or 6, they could suddenly feel liberated from the strictures of EQAO. "Last year I was in grade 6 when I did my *New Pedagogy* project and I was like, 'Come on, I've got to get it done. EQAO is coming," one teacher remarked. But "this year," in a different grade, "it was like, 'Let's fly with this!' It's a big difference. If we didn't do math today, it doesn't matter. We'll catch up with it. The kids are engaged."

Another district resolved the tension between innovation and traditional large-scale assessments systemically. It seemed to avoid Grades 3 and 6, and even Grades 2 and 5, as places to introduce major innovations such as inquiry-based learning, mathematics reform, and NPDL (see also Owston et al. 2016). It is tempting to put innovation aside when EQAO comes closer, and this can hinder the pursuit of broader and deeper learning.

During a year of teacher action when the EQAO tests were suspended, teachers in one district expressed how much they benefitted from looking at other kinds of data instead. "Wow, I think I really understand my school community now," said one. Another put it this way: "Not having that [EQAO] data this year made our school improvement planning much richer, because we were looking at different data, which we should have been doing all along."

Fifth, the movement towards inquiring into the wider aspects of students' learning and development placed strain on the 6-week cycles of data-driven intervention that had been carried over from the period when EQAO and other data were used to raise performance and drive up achievement in literacy and mathematics. In one district, for example, a teacher explained that there were concerns about how to handle the sheer volume of data:

> Everybody's gathering data, which is good, but what do we do with it, and what is the best data to gather? Now you see the teachers are taking pictures, they're observing, they have checklists, they're gathering it too, but what to do with it?"

Teachers in this district felt that 6 weeks was not a sufficient window of time to collect greater volumes of data, identify issues, set out goals and objectives, and meet periodically to assess progress - while still attempting to manage an elementary school classroom full of diverse learners and needs. One teacher said that

> Within the space of three weeks, we need to determine what our goal is, have built that up with our students and clearly established what the criteria are with our students, collect the data, and be working on it before that mid-point meeting. Then, we have another three weeks to keep going with that, to continue collecting data, to hopefully bring them to a successful conclusion of that project, to then have our data for the final meeting. I think my colleagues and I are all in agreement that 6 weeks is too short of a time. We'd like to see it doubled.

Another teacher struggled with finding the time "to just actually implement data collection and data analysis. It's difficult. The classroom's a busy place. There are always problems that need to be addressed immediately, so it's just a question of time." A principal added that

> the time between the meetings is sometimes too short because we establish learning outcomes, let's say at the first PLC, then three weeks later we have the mid PLC, and sometimes, with all of the school activities and other workshops, teachers find it very hard to establish the strategies to reach the learning outcomes that we've set. Sometimes maybe 6 weeks is a little short. For next year I might be discussing if we can add a week between meetings.

The 6-week teaching–learning cycle was instituted along with EQAO and related-interim assessments as part of an improvement design that originally was intended to raise achievement results and narrow achievement gaps quickly in literacy and mathematics. But broader goals that incorporate critical issues such as attention to students' well-being, deeper explorations into the nature and source of students' learning problems and strengths, and greater use of wider spans of data to inform professional judgment on these matters, have meant that the original large -scale assessment design no longer fits the agenda of *Achieving Excellence*. It has become an anachronistic remnant left over from a previous change strategy that has hindered the province in its recent pursuit of more ambitious goals.

Last, there was a perception from the educators that the EQAO is not a neutral assessment, but rather an intervention that can actively harm their students' well-being. "I have kids that suffer from anxiety, so putting them into a testing situation like this seems totally wrong," one teacher said. Another recalled, "I spent so much time all year long trying to build the confidence of these children, that they were learners, that they were good at what they were able to do, and then this test would roll around and I would have to then give these kids things that they weren't able to do. I couldn't support them." A principal concurred: "Kids feel a lot of stress about it. Even though they're not going to be punished for it, they feel a lot of stress and anxiety about writing it."

One educator had experienced test anxiety even in her own family:

> My son is in Grade 3 this year. Two nights ago (when he went to bed) it was, "What if I put a comma in the wrong place?" I was like, "It doesn't matter." I've never said anything one way or the other, or anti-whatever. I'm like, "So you put a comma in the wrong place." He's like, "But the teacher is saying…" And I get it, because the teachers feel badly when it's ranked in the paper and it's in Maclean's magazine and the school is going to be reflected poorly.

Experiencing reactions such as these from their students has led Ontario's educators to question the necessity of the EQAO altogether. "There's a lot of pressure," one principal remarked. "I can picture one of my Grade 3 teachers. She's carrying the weight of things she can't control." This teacher had a student with ADHD who spent hours each day "spinning in his chair." She needed the time and space to support the student, but she also "knows this [the EQAO] is coming." She found herself focusing on the test, rather than on creating a positive learning environment that would promote learning and well-being at the same time.

Discussion

The original purpose of EQAO was to be an instrument of educational accountability. With the *Drive to 75*, it also became a tool for tracking, monitoring and intervention as well as mid-stakes accountability. Most system administrators, many school administrators, and special education support staff saw value in this large-scale assessment and its associated collaborative processes of data-driven intervention for raising expectations for all students, increasing awareness about vulnerable groups of students, and stimulating just-in-time interventions that raised achievement and narrowed achievement gaps. Combined with a relentless focus on improving literacy instruction along with extensive supports in the form of materials, training and coaching, the results were evident in a 17-point improvement in literacy scores at a sustainable rather than statistically improbable rate. Achievement gaps were also narrowed for second language learners and students with learning disabilities.

At the same time, many teachers were critical of EQAO, and there was considerable evidence that the assessments and associated top-down pressures for improvement and achievement gains led to practices of placing excessive focus on students just below the provincial target of Level 3 proficiency. This indicates that the perverse incentives that Campbell warned against can still occur in mid-stakes as well as high-stakes testing environments.

By 2016, learning goals in Ontario had developed greater depth and complexity than concentrating on literacy and mathematics achievement alone, and, in line with global trends (OECD 2017), there were also growing concerns about children's mental health and emotional well-being. Our interviews revealed that the closer to the classroom that the roles of educators got, the more that the holders of those roles saw that large-scale testing was having detrimental effects, not just on well-being, but on learning and innovation too.

A new context is characterizing more and more educational systems around the world. It is prominent in the global directions expressed in UNESCO's sustainable development goals and OECD's global competencies. In light of these developments, the benefits of EQAO for monitoring progress and stimulating higher expectations for achievement now appear to have been outweighed by the harmful consequences for broad excellence, equity and well-being. The strengthening of Ontario's culture of collaborative professionalism has eased or even eliminated tendencies to concentrate undue attention on marker students just below the point of proficiency, in favour of more authentic inquiry into the genuine strengths of and struggles among students of mystery and wonder. But, aside from this, the twentieth century large-scale assessment system has not kept up with twenty first century goals for deeper learning and young people's development. Instead, there is teaching to the test, cultural bias, avoidance of innovation in the years during and prior to the ones where tests are administered, subjection of teachers to heartbreaking dilemmas of whether to include highly vulnerable students in the testing process or not, and active creation of emotional ill-being among students and teachers alike. A twenty first century movement in deeper learning and stronger

well-being that is embracing a range of innovative practices has been rapidly outpacing Ontario's twentieth century system of large-scale, mid-stakes assessment.

Conclusion, implications and epilogue

The value of mid-stakes tests for raising achievement and improving equity in Ontario has always been genuinely debatable. In the context of bolder goals for deeper learning and improved well-being, the perverse consequences of mid-stakes assessments now seem to be as serious and pervasive as they have been in the high stakes assessments of England, the US, and elsewhere. Promoting well-being while continuing to perpetuate ill-being through large-scale, standardized assessment practices, runs contrary to the goals of achieving systemic coherence, and corrodes the sense of moral purpose that high performing systems proclaim and promote. At the same time, attempts at curriculum innovation and personalization are being limited or undermined by a testing system that promotes standardization.

In 2017, some of the emerging findings of this study were reported to the Premier, the Education Minister, Mitzie Hunter, and their senior staffs, as part of a feedback meeting from the Premier's and Minister's six advisors, of whom I was one. Three pieces of feedback were presented from our research. Positive attention was drawn to how the system was handling its mathematics improvement strategy and to the widespread use of well-being initiative across all 10 districts being studied, without the heavy hand of top-down implementation. At the same time, it was pointed out that while the province and its Premier had an admirable and authentic commitment to improving children's well-being, its continued use of EQAO assessments was creating ill-being. Without proposing any particular solution, the Government was asked to take collective moral responsibility for the fact that it was currently producing ill-being among young children in an environment where one of its core priorities was promoting well-being.

Shortly afterwards, on September 6, 2017, the Premier and the Minister of Education announced an Independent Review of Assessment and Reporting, including the work of EQAO, by the six advisors, including Michael Fullan and myself. The review was chaired by Carol Campbell of the Ontario Institute for Studies in Education, who was formerly Director of the Ontario Ministry of Education Research Strategy and Evaluation Branch. The review was based on analysis of relevant provincial, national and international evidence; meetings with stakeholder representatives and assessment experts; a web-based survey; consideration of submissions of evidence and a range of public engagement meetings across the province involving over 5000 people (Campbell et al. 2018).

The review did not oppose all large-scale assessment. There are legitimate roles for such assessment in both maintaining accountability and also providing system leaders with knowledge of how their students are performing. However, the authors of the review argued, "going forward, large-scale assessment data should not be used for individual student diagnostic or evaluative purposes and students should not be subject to excessive test preparation for a summative system-level snapshot" (p. 4). The report identified a strong consensus about the negative, indirect side effects of

EQAO assessments on students' learning and well-being. It concluded that inappropriate uses of EQAO and other large-scale assessments include "using large- scale provincial assessments for student diagnostic purposes, to infer evaluation of educators, and for ranking schools and school boards" (p. 8).

Among the report's recommendations were phasing out and ending of EQAO assessments before Grade 6 to avoid threats to well-being among young children; allowing flexibility over timing of assessment events for different students; and "vigilant attention to ensuring curriculum and assessment materials provide linguistically, culturally and geographically relevant items and materials", including indigenous ways of knowing. The report also recommended providing effective accommodations and modifications for students with special educational needs including appropriate exemptions; and mitigating unintended negative consequences especially by sticking to the original purpose of EQAO in providing a summative snapshot rather than also being used as a tool to promote improvements and interventions in particular schools and districts and to rank schools (pp. 11–13).

The report was presented in March 2018. In April, all recommendations were accepted, in principle, by the Premier. In May 2018, the Liberal Government was not re-elected and the new Progressive Conservative Government disbanded the team of advisors (which had included Campbell, Fullan and me), removed the report from its Ministry website, and, for the first time, appointed a new Chair of the Board of EQAO on a 6-figure salary rather than following previous practice of it being a voluntary position.

In November 2018, EQAO announced some modest adjustments to test practices in order to "modernize" the test environment, including broader student access to "headphones, calming white noise or music" and "encouraging the classroom environment to look as it would normally during an assessment", in terms of leaving up displays of student work and other materials that do not amount to test item prompts. Subsequently, other adjustments have been announced, such as greater accommodations for English Language Learners, and inclusion of Indigenous content, pointing to incremental implementations of other aspects of the advisors' review, but falling short of abolishing testing before Grade 6.

Meanwhile, as of February 2020, the Ontario Government is in dispute with the teacher unions over class sizes, mandated online learning and teacher salaries, and amid a series of rotating strikes, the Grade 9 assessment has been suspended by most school districts and there is wider talk among the unions of abolishing EQAO altogether. In March, 2020, all these discussions were suspended due to the COVID-19 pandemic.

In 2020, Michael Fullan's work has moved on from advocating government-driven, large-scale reform, focusing on literacy and mathematics aligned with census-based testing. Instead, he is concentrating on developing global networks of innovative schools committed to deep learning as expressed in qualities such as creativity, character, and citizenship (Fullan et al. 2018). With Dennis Shirley, my own work, meanwhile, draws on our research on the second period of Ontario educational reform from 2014, to outline a move in Ontario and elsewhere from an *Age of Achievement and Effort* in which narrow achievement goals and testing prevailed, to an *Age of Engagement, Identity and Wellbeing* in which broader and more inclusive

goals of learning and human development are more evident alongside declining commitments to top-down reform strategies, including large scale and especially high stakes testing (Shirley and Hargreaves, in press). In this respect, with regard to the wider field of educational change and assessment, while the work of Fullan and myself had undergone a considerable divergence in the first decade and a half of the twenty first century during the *Age of Achievement and Effort*, it has reached a new convergence in the latter part of this second decade.

Overall, what we have found from our intensive long-term research in Ontario across the two different ages of educational change is that mid-stakes assessments combined with strong support are to some degree contestable when their contribution to raising achievement results in foundational areas of achievement like literacy and mathematics, is set against the negative side effects reported here. But when educational goals for learning and well-being become broader and deeper, as many contemporary system goals have become in an *Age of Engagement, Identity and Wellbeing*, the evidence of this research on their use in Ontario, is that this 20-year old strategy of accountability and systemic improvement is now an educationally ineffective and systemically incoherent anachronism.

References

American Educational Research Association. (2000). *Position statement on high stakes testing*. Washington, DC: AERA. http://www.aera.net/About-AERA/AERA-Rules-Policies/Association-Policies/Position-Statement-on-High-Stakes-Testing.

American Statistical Association. (2014). ASA statement on using value-added models for educational assessment. Alexandria, VA: ASA. https://www.amstat.org/asa/files/pdfs/POL-ASAVAM-Statement.pdf.

Baker, M., & Foote, M. (2006). Changing spaces: Urban school interrelationships and the impact of standards-based reform. *Educational Administration Quarterly, 42*(1), 90–123.

Baker, B., Oluwole, J., & Green, P., III. (2013). The legal consequences of mandating high stakes decisions based on low quality information: Teacher evaluation in the Race-to-the-Top era. *Education Policy Analysis Archives, 21*(5), 1.

Barber, M. (2009). From system effectiveness to system improvement. In A. Hargreaves & M. Fullan (Eds.), *Change wars* (pp. 71–94). Bloomington, IN: Solution Tree.

Biggs, J. (2003). *Teaching for quality learning at university—What the student does* (2nd ed.). Buckingham: SRHE/Open University Press.

Bird, S., Cox, D., Farewell, V., Goldstein, H., Holt, T., & Smith, P. (2005). Performance indicators: Good, bad and ugly. *Journal of the Royal Statistical Society: Series A, 168*(Part 1), 1.

Board of Education. (1911). *Report of the consultative committee on examinations in secondary schools, (Cd 6004)*. London: HMSO.

Booher-Jennings, J. (2005). Below the bubble: "Educational triage" and the Texas accountability system. *American Educational Research Journal, 42*(2), 231–268.

Bray, M. (2006). Private supplementary tutoring: Comparative perspectives on patterns and implications. *Compare, 36*(4), 515–530.

Braun, H. (2015). The value in value added depends on the ecology. *Educational Researcher*, *44*(2), 127–131. https://doi.org/10.3102/0013189X15576341.

California Department of Education. (2016). *Our journey together on the California way*. Retrieved from https://www.cde.ca.gov/eo/in/bp/bp2intro.aspevery.

Campbell, D. T. (1976). *Assessing the impact of planned social change*. Kalamazoo, MI: Evaluation Center, College of Education, Western Michigan University.

Campbell, C., Clinton, J., Fullan, M., Hargreaves, A., James, C., & Longboat, K. D. (2018). *Ontario: A learning province*. Toronto: Queens Printer, Ontario Ministry of Education.

Campbell, C., Zeichner, K., Lieberman, A., & Osmond-Johnson, P. (2017). *Empowered educators in Canada: How high-performing systems shape teaching quality*. San Francisco: Jossey Bass.
Chiu, P. (2018). Changes to controversial Hong Kong primary school assessments set to be announced, sources say, South China Morning Post, March 15. https://www.scmp.com/news/hong-kong/education/article/2137395/changes-controversial-hong-kong-primary-school-assessments.
Daly, A. (2009). Rigid response in an age of accountability: The potential of leadership and trust. *Educational Administration Quarterly, 45*(2), 168–216.
Darling Hammond, L. (2004). Standards, accountability, and school reform. *Teachers College Record, 106*(6), 1047–1085.
Earl, L., & Torrance, N. (2000). Embedding accountability and improvement into large-scale assessment: What difference does it make? *Peabody Journal of Education, 75*(4), 114–141.
Elmore, R. (Ed.). (2004). *School reform from the inside out: Policy, practice and performance*. Cambridge, MA: Harvard University Press.
Falk, J., & Drayton, B. (2004). State testing and inquiry-based science: Are they complementary or competing reforms? *Journal of Educational Change, 5*(4), 345–387.
Fuhrman, S., & Elmore, R. F. (Eds.). (2004). *Redesigning accountability systems for education*. New York: Teachers College Press.
Fullan, M. (2000). The return of large-scale educational reform. *Journal of Educational Change, 1*(1), 5–28.
Fullan, M. (2010). The role of the district in tri level reform. In P. Peterson, E. Baker, & B. McGaw (Eds.), *International encyclopedia of education* (Vol. 6, pp. 295–302). Oxford: Elsevier.
Fullan, M. (2011). *Choosing the wrong drivers for whole system reform, Seminar Series 204*. Melbourne: Centre for Strategic Education.
Fullan, M. (2012). *Lead the change: Q&A with Michael Fullan, AERA Educational Change Special Interest Group, Issue 16*. https://michaelfullan.ca/wp-content/uploads/2016/06/13514675730.pdf.
Fullan, M., & Quinn, J. (2015). *Coherence: The right drivers in action for schools, districts, and systems*. Thousand Oaks, CA: Corwin.
Fullan, M., Quinn, J., & McEachen, J. (2018). *Deep learning: Engage the world to change the world*. Thousand Oaks, CA: Corwin.
Goldstein, H. (2001). Using pupil performance data for judging schools and teachers: Scope and limitations. *British Educational Research Journal, 27*(4), 433–442.
Hargreaves, A. (Ed.). (1997). *Rethinking educational change with heart and mind, 1997 ASCD Yearbook*. Alexandria: Association for Supervision and Curriculum Development.
Hargreaves, A. (2003). *Teaching in the Knowledge Society*. Teachers' College Press.
Hargreaves, A. (2019). Teacher collaboration: 30 years of research on its nature, forms, limitations and effects. *Teachers and Teaching*, *25*(5), 603–621. https://doi.org/10.1080/13540602.2019.1639499.
Hargreaves, A., Braun, H. I., Hughes, M., Chapman, L., Lam, K., Lee, Y., et al. (2012) *Leading for all: A research report on the development, design, implementation and impact of Ontario's "Essential for some, good for all" initiative*. Ontario: Council of Directors of Education.
Hargreaves, A., Earl, L., Moore, S., & Manning, S. (2001). *Learning to change: Teaching beyond subjects and standards*. San-Francisco: Jossey-Bass.
Hargreaves, A., & Fink, D. (2005). Why Ontario does not measure up. *Toronto Star*, October 25, p. A25.
Hargreaves, A., & Goodson, I. (2006). Educational change over time: The sustainability and non-sustainability of three decades of secondary school change and continuity. *Educational Administration Quarterly*, *42*(1), 3–41.
Hargreaves, A., & Harris, A. (2011). *Performing beyond expectations*. UK National College for School Leadership.
Hargreaves, A., Lieberman, A., Fullan, M., & Hopkins, D. (Eds.). (1998). *The international handbook of educational change*. Dordrecht: Kluwer Publications.
Hargreaves, A., & O'Connor, M. T. (2018). *Collaborative Professionalism*. Corwin.
Hargreaves, A., & Shirley, D. (2009). *The fourth way*. Corwin Press.
Hargreaves, A., & Shirley, D. (2012). *The global fourth way: The quest for educational excellence*. Corwin.
Hargreaves, A., & Shirley, D. (2018a). What's wrong with well-being? *Educational Leadership*, *76*(2), 58–63.
Hargreaves, A., & Shirley, D. (2018b). Well-being and success: Opposites that need to attract, Education Canada, Winter, November 29. https://www.edcan.ca/articles/well-being-and-success/.

Hargreaves, A., & Shirley, D. (2020). Leading from the middle: its nature, origins and importance. *Journal of Professional Capital and Community*, *5*(1), 92–114. https://doi.org/10.1108/JPCC-06-2019-0013.

Hargreaves, A., Shirley, D., Wangia, S., Bacon, C., & D'Angelo, M. (2018). *Leading from the Middle*. Ontario: Council of Directors of Education.

Hill, P. W., & Crévola, C. A. (1999). Key features of a whole-school, design approach to literacy teaching in schools. *Australian Journal of Learning Difficulties, 4*(3), 5–11.

Kohn, A. (2015). *Schooling beyond measure and other unorthodox essays about education*. Portsmouth, NH: Heinemann.

Koretz, D. M. (2017). *The testing charade: Pretending to make schools better*. Chicago: University of Chicago Press.

Kotter, J. P. (1996). *Leading change*. Boston, MA: Harvard Business Press.

Lee, J., & Kang, C. (2019). A litmus test of school accountability policy effects in Korea: cross-validating high-stakes test results for academic excellence and equity. *Asia Pacific Journal of Education*. https://doi.org/10.1080/02188791.2019.1598851.

Meyer, A., Rose, D. H., & Gordon, D. (2013). *Universal design for learning: theory and practice*. Wakefield, MA: Cast Incorporated.

Newton, P. (2008). Contextualising the comparability of examination standards. In P. Newton, J. Baird, H. Goldstein, H. Patrick, & P. Tymms (Eds.), *Techniques for monitoring the comparability of examination standards* (p. 22). London: Qualifications and Curriculum Authority.

Ng, P. T. (2017). *Learning from Singapore: The power of paradoxes*. New York, NY: Routledge.

O'Grady, K., Deussing, M.-A., Scerbina, T., Fung, K., & Muhe, N. (2016). *Measuring up: Canadian results of the OECD PISA study*. Ottawa: Canada, Council of Ministers of Education.

Ontario Ministry of Education. (2014). *Achieving excellence: A renewed vision for education in Ontario*. Retrieved from http://www.edu.gov.on.ca/eng/about/great.html.

Organization for Economic Cooperation and Development. (2011). *Strong performers and successful reformers in education: Lessons from PISA for the United States*. Paris: OECD.

Organization for Economic Cooperation and Development. (2017). *PISA 2015 results (Volume III): Students' well-being*. Paris: Author. https://doi.org/10.1787/9789264273856-en.

Owston, R., Wideman, H., Thumlert, K., & Malhotra, T. (2016). *Transforming learning everywhere: A study of the second year of implementation*. Toronto, Ontario: York University.

Ravitch, D. (2011). *The death and life of the great American school system: How testing and choice are undermining education*. New York: Basic Books.

Reedy, D. (2018). *Independent review of the Scottish National Standardised Assessments at Primary 1*. Edinburgh: Scottish Government.

Rothstein, R. (2014). *A strong precedent for a better accountability system*. Washington, DC: Economic Policy Institute, February 14. https://www.epi.org/blog/strong-precedent-accountability-system/.

Royal Commission on Learning. (1995). *For the love of learning*. Toronto: Queen's Printer for Ontario.

Sahlberg, P. (2011). The fourth way of Finland. *Journal of Educational Change, 12*(2), 173–185.

Schofield, J. (2001). Saving our schools, MacLeans. http://archive.macleans.ca/article/2001/5/14/saving-our-schools.

Severson, K. (2011). Systematic cheating is found in Atlanta's school system. *New York Times*. Retrieved from http://www.nytimes.com/2011/07/06/education/06atlanta.html.

Shirley, D. & Hargreaves, A. (in press). *Learn to be: Engagement, identity and wellbeing in education.* Bloomington, IN., Solution Tree.

Shirley, D., Hargreaves, A., & Washington, S. (2020). The sustainability and non-sustainability of teachers' and leaders' wellbeing. *Teaching and Teacher Education*. https://doi.org/10.1016/j.tate.2019.102987.

Smith, M. S., & O'Day, J. (1990). Systemic school reform. *Journal of Education Policy, 5*(5), 233–267.

Torlakson, T. (2015). *A blueprint for great schools: Version 2.0*. Sacramento, CA: California Department of Education.

Tucker, M. (2019). *Leading high-performance school systems: Learning from the world's best*. Alexandria, VA: ASCD.

Tucker, M. S. (2011). *Standing on the shoulders of giants: An American agenda for education reform*. Washington, DC: National Center on Education and the Economy (NCEE).

Tymms, P. (2004). Are standards rising in English primary schools? *British Educational Research Journal, 30*(4), 477–494.

US Department of Education. (2015). *Every Student Succeeds Act*. Washington: DC, Author.

Weber, M. (1968). *Economy and society: An outline of interpretive sociology*. New York: Bedminster Press.
Wolfe, R., Childs, R., & Elgie, S. (2004). *Final report on the external evaluation of EQAO's assessment processes*. Toronto: OISE, The University of Toronto.
Zhao, Y. (2018). *What works may hurt—Side effects in education*. New York: Teachers College Press.