Research Paper

ROE

ECNU Review of Education
2019, Vol. 2(3) 279–296

DOI: 10.1177/2096531119878964
journals.sagepub.com/home/roe

SAGE

ECNUP

Side Effects of Large-Scale Assessments in Education

Trina E. Emler, Yong Zhao, Jiayi Deng and Danqing Yin
University of Kansas

Yurou Wang
The University of Alabama

Abstract

Purpose: In line with a recent call for side effects research in education, this article aims to synthesize the major concerns that have been raised in the literature concerning large-scale assessments (LSAs) in education.

Design/Approach/Methods: The researchers endeavored to complete a deep review of the literature on LSAs to synthesize the reported side effects. The review was synthesized thematically to understand and report the consequences of the ongoing push for the use of LSA in education.

Findings: Thematic analysis indicated overarching side effects of LSA in education. We discuss why negative side effects exist and present evidence of the most commonly observed side effects of LSA in education, including distorting education, exacerbating inequity and injustice, demoralization of professionals, ethical corruption, and stifling of innovation in education.

Originality/Value: While concerns about the use and misuse of LSA in education are not new and have been discussed widely in the literature, rarely have they been discussed as inherent qualities and consequences of LSAs that can do harm to education.

Corresponding author:
Trina E. Emler, Department of Educational Leadership and Policy Studies, School of Education, University of Kansas, 1122 W Campus Road, Lawrence, KS 66049, USA.
Email: trina.emler@ku.edu

Keywords

Inequity, large-scale assessments, side effects

Date received: 16 April 2019; accepted: 6 September 2019

Recent decades saw a significant increase in the use of large-scale assessments (LSAs) in education. More LSAs have been developed and implemented. More students and teachers have been subject to the consequences of LSAs. More countries and areas have engaged with LSAs. As a result, the impact and influence of LSAs have become broader, deeper, and more consequential.

The growth of LSAs suggests their value and utility in education. LSAs can be powerful tools to hold educators accountable, efficient means to collect and provide evidence for policy making, and are perceived as objective and incorruptible instruments for sorting and selecting individuals for competitive opportunities such as prestigious colleges. Accompanying the growth in use and impact is the growth in criticism of LSAs. While LSA has never lacked criticism (Popham, 1999), its growth has attracted more (Lewis & Lingard, 2015). The criticism of LSA touches virtually all aspects of LSA. Some of the criticism concerns technical issues such as the design and implementation, which can be addressed with more sophisticated techniques and technologies. Others are more concerned with the uses of LSA and their negative consequences, which cannot be addressed with better testing techniques, more sophisticated statistical models, improved sampling methods, or more careful analytical methods. As such, these problems of LSAs cannot have a technical fix.

These problems are side effects—adverse, undesirable, yet unavoidable effects of effective treatments. LSAs in education, like any other educational policies and practices, are accompanied with inevitable side effects while being effective in delivering positive outcomes (Abu-Alhija, 2007; Karsten, Visscher, & De Jong, 2001; Zhao, 2017, 2018d). That is, even while LSAs can work well in solving certain problems, they can cause harm. And the harm cannot be avoided because the mechanism that causes the harm is the same as that which makes it valuable and effective. The purpose of this article is to discuss the side effects of LSAs in education.

Defining characteristics of LSAs

LSAs have been used for a variety of purposes. Chief among them are three: accountability, selection, and comparison. LSAs for accountability are assessments intended to hold educational professionals accountable for their students learning. The U.S., for example, has been using statewide LSAs to hold teachers and school leaders accountable for student achievement through legislations such as the *No Child Left Behind Act* (*NCLB*; *No Child Left Behind Act of 2001*, 2002). LSAs are also widely used for selection purposes of all sorts—from selecting for

educational opportunities such as college and graduate programs to selecting for occupational positions such as civil servants. The SAT and ACT in the U.S. are examples. A third purpose LSAs have been used to serve is comparisons—comparing education qualities among regions, states, and nations. The Trends in International Mathematics and Science Study (TIMSS) is a prime example of international comparisons.

Sometimes the purposes are intertwined, with one LSA serving multiple purposes. For example, the National College Entrance Exam in China, while primarily serving as a selection assessment, also serves the purpose of accountability. Quite often the performance of students on the National College Entrance Exam affects the reputation of schools and teachers. It is also directly linked to teachers' economic well-being because students' performance on the exam is associated with bonuses and other benefits of teachers.

Whatever the purpose, what defines LSA is its scale. Large scale in LSA means the number of people the assessment is given to. Only when an assessment is administered to a large number of people can it be called LSA. In other words, regardless of the purposes of an assessment, it is the number of people it affects that makes it an LSA. The nature of being large scale results in a number of defining characteristics of LSAs.

Uniformity

In order for it to be administered to a large population of students and maintain the necessary validity and reliability, LSAs must have uniformity in a number of aspects. An LSA must first have a uniform format in that it cannot be of different formats for different people who take the assessment. It must also have a uniform underlying construct or constructs in that it cannot be measuring different constructs for different people, although some items can be different. Furthermore, LSAs must be uniformly administered, uniformly scored, and uniformly reported. In other words, the procedures for administration, scoring, and reporting should be highly standardized.

High cost

LSAs are expensive. Regardless of who bears the cost of a particular LSA, it requires a lot of money to design, develop, administer, score, and report. For instance, the U.S. spends roughly US$1.7 billion on state assessments for K-12 students each year (Chingos, 2012). These assessments are paid for by taxpayers. The base cost for each participating country in the Programme for International Student Assessment (PISA) 2021 is 205,000 euros or roughly US$230,000 (PISA, n.d.). If a country wants support with preparation and implementation, it needs to pay additional 210,000 euros. And support with data analysis and reporting costs extra at 250,000 euros per country. The SAT costs $47.50 for each test-taker (SAT with essay cost US$64.50) in the 2017–2018 school year, and 2,136,539 students took the SAT in 2018 (College Board, 2019).

Assuming all students just took the SAT (without essay), the total cost was over 100 million dollars in the year of 2018. This cost was largely borne by the families of the test-takers.

Broad impact

LSAs affect a large number of people. Regardless of the purpose, LSAs by design affect the lives of many people—those who take the test, their teachers, their educational institutions, and possibly their families, to say the least. For instance, while over two million students took the SAT, about the same number of students took the ACT, the other college admissions assessment in the U.S., making the total number of college admissions test-takers over four million in 2018. The Chinese National College Entrance Exam has about 10 million test-takers annually. The 2015 round of PISA were administered to 540,000 students, representing about 29 million 15-year-olds in more than 70 countries (Organisation for Economic Co-operation and Development [OECD], 2016a). The test-takers are not the only ones affected by LSAs. An LSA can affect the lives of many people associated with the test-takers. College admissions assessment often directly affect the family of the test-takers and in some instances the teachers, school administrators, and even the local government officials because they are considered responsible for the results.

High stakes

All LSAs carry high stakes for some people. It is easy to understand that some LSAs carry high stakes because they are designed to have serious consequences on the test-taker. For example, college admissions assessments are meant to be used for making admissions decisions and thus carry a very high stake for the students. State accountability tests required by *NCLB* and *Every Student Succeeds Act* in the U.S. are high stakes because they have significant consequence on teachers and schools, although they have no direct consequences on students.

But even LSAs that are not meant to be high-stakes assessments can have significant consequences, perhaps not necessarily for the test-takers directly. For example, the National Assessment of Educational Progress (NAEP) in the U.S., unlike state accountability assessments, is designed to have no consequence on the test-takers or their teachers and schools. The National Literacy and Numeracy Assessment Program (NAPLAN) in Australia is also meant to carry low stakes. PISA is also a low-stakes LSA. But they all carry significant consequences, indirectly for the students who take the assessment, and for others who are associated with the students. NAEP results, for example, have often been used to influence educational policies, which in turn can affect students, educators, policy makers, and even assessment makers. PISA results are often interpreted as an indicator in global rankings of national competitiveness and national education quality, which can affect people's decisions in migration, investment, and/or conducting business. NAPLAN results

in Australia affect school reputation, which in turn affects housing prices and the fate of schools because people can make choices about where to send their children based on NAPLAN scores.

In some sense, there is no LSA that is low stakes. Everyone who decides to invest tremendous amount of resources in an LSA expects the assessment to have impact. And to have impact, the assessment needs to have consequences for someone.

The power of LSAs

These characteristics are inherent in all LSAs. If they are removed, LSAs are no longer large scale. Furthermore, they are the mechanism that makes LSAs effective and useful. Therefore, they cannot be altered. As such, the negative impact resulting from these characteristics cannot be removed with technical improvements.

The high stakes that they carry for a large number of people give LSAs the power to command attention of all involved in education. LSAs for accountability direct the attention of teachers and school leaders, who in turn direct attention of their students. LSAs for selection command the attention of students, their teachers, and their families. LSAs for comparisons hold the attention of political leaders and education system leaders, who hold the power to make significant policy changes, which then gets the attention of students and teachers.

Through the attention they can command, LSAs exert tremendous power over all aspects of education and having influence over educational practices and policies is often the intended, desirable outcome of LSAs. When performing well on an LSA becomes the ultimate goal, whatever it measures dictates what is taught and learned, directs how teachers and students spend their time, and affects where society and parents invest their resources. Therefore, if LSAs could measure *all* that is desirable in education and what matters in life, they would serve as an excellent tool to improve education.

However, it is impossible for LSAs to measure everything that matters in education and for individual success and societal prosperity for several reasons. First, there is much dispute over knowledge, skills, and other human qualities that make a person successful in life or a society prosperous in the future as human societies are constantly changing (Duckworth & Yeager, 2015; Florida, 2012; Goldin & Katz, 2008; Levin, 2012; Pink, 2006; Zhao, 2018b, 2019). Second, we do not have valid and reliable ways to assess many important human qualities such as creativity, entrepreneurship, and social–emotional well-being on a large scale (Duckworth & Yeager, 2015; Zhao, 2016a). Third, the uniform nature makes it difficult for one LSA to validly and reliably measure everything that matters because what matters can be in conflict or competing with each other (Zhao, 2018d). Finally, the high costs make it difficult to develop and administer unlimited number of LSAs frequently. For these reasons, LSAs have only been able to measure skills and

knowledge in a very limited number of areas such as mathematics, language, and science. What has been measured is typically limited to cognitive ability in these areas.

Side effects of LSAs in education

The combination of the power of LSAs in dictating education and the narrow scope of skills, knowledge, and human qualities they are able to or choose to measure are the source of their negative impact on education. The negative impact or side effects have been observed and reported in different ways in the literature. One of the most frequently used phrases that refer to these impacts is "unintended consequences." They are also referred to as "collateral damages" (Nichols & Berliner, 2007). This section summarizes some of the most troubling side effects of LSAs that have been observed and reported in the literature.

Distorting education

One of the most damaging side effects of LSAs is the distortion of education. The distortion happens on multiple levels. First, LSAs distort the purpose of education by misleading the public to believe that performance on LSAs accurately reflects the quality of education, albeit not necessarily intentionally. Very few would dispute that the purpose of education encompasses much more than mastering the skills to perform well on tests, let alone on tests that measure a very limited number of subjects (Dewey, 1975; Labaree, 1997; Sjøberg, 2015; Zhao, 2016c). However, LSAs have effectively instilled in the mind of the public that the purpose of education is primarily to prepare for tests and consequently test scores reflect the quality of education. As evidence, TIMSS and PISA scores have been equated with the quality of education in different countries (OECD, 2016b; Tucker, 2011; Zhao, 2016c), resulting in strong media and policy reactions around the world (Baird et al., 2016; Dillion, 2010; Figazzolo, 2009; Gruber, 2006; Sellar & Lingard, 2014; Sjøberg, 2015). The fact that schools and teachers are judged based on their students' test scores (*ESSA*, 2015; Hout & Elliott, 2011; *No Child Left Behind Act of 2001*, 2002) on LSAs is another piece of evidence, so is the fact that NAPLAN "has become a de-facto measure of schools" (Baker, 2019).

Second, when the purpose of education is distorted to be only preparing for tests, curriculum gets distorted. The distortion primarily comes in the form of curriculum narrowing. While many education systems prescribe a quite broad curriculum that includes much more than the limited number of subjects often assessed with LSAs, the assessed subjects are at the core and receive a lot more attention and resources, pushing the non-assessed ones to the periphery. What students should know and what students do not know are all highly controlled by the examinations (Kirkpatrick & Zang, 2011).

As evidence, in China, for instance, only subjects tested by the National College Entrance Exam are taken seriously (Yu & Suen, 2015; Zhao, 2014). And the U.S. has witnessed a trend of curriculum narrowing since the enactment of *NCLB* in the U.S. (Abu-Alhija, 2007; Albrecht & Joles, 2003; Berliner, 2011; Bowen & Rude, 2006; Klenowski, 2011; Klinger & Rogers, 2011; Popham, 2000; Tienken & Zhao, 2013; Towles-Reeves, Garrett, Burdette, & Burdge, 2006; Volante, 2005, 2006). In surveys of U.S. classrooms in which students were tested on two core subjects (reading and math), classrooms added between 75 and 150 min weekly to each of the two areas. If added to both, up to 300 min of instructional time was added to the tested disciplinary strands (Berliner, 2011). Of course, this added time has to come at the expense of other subject areas, diverse skills, and varied forms of thinking, including cuts to social studies, science, physical education, recess, art, and music (Berliner, 2011). In Australia, NAPLAN has resulted in a narrower curriculum for children (Dulfer, 2012) as well.

Third, LSAs distort education by distorting instruction, resulting in the phenomenon of teaching to the test. Teaching to the test refers to instruction that is solely focused on and often limited to what is to be included in the test. It also means spending class time teaching students test-taking skills. Teaching to the tests is a wide spread phenomenon in classrooms of education systems that have a long tradition of LSAs for selection such as China and Korea, where students are not only taught how to take tests but also spend an excessive amount of their time exhausting different ways a knowledge point can be assessed by repetitively doing model tests or going over tests from previous years (Zhao, 2009, 2015b). The effect of LSA on distorting instruction is clearly evidenced by the increase of instructional time devoted to preparing students for state accountability assessments in the U.S. since *NCLB* (Menken, 2006). There is also evidence that suggests a significant reduction of time spent in deep learning and other important skills such as creativity, problem-solving, organization of knowledge, self-monitoring skills, and the like (Abu-Alhija, 2007; Albrecht & Joles, 2003; Berliner, 2011; Chudowsky & Pellegrino, 2003; Klenowski, 2011; Klinger & Rogers, 2011; Popham, 2000; Towles-Reeves et al., 2006).

When education is reduced to test preparation, students are deprived of opportunities to develop skills and abilities that are more important for their own future and the future of the society (Zhao, 2012, 2014, 2015a). For instance, creativity, problem-solving, organization of knowledge, self-monitoring skills, entrepreneurship, and other skills and abilities are all considered of highest import for today's learners to be successful in society (Chudowsky & Pellegrino, 2003; Duckworth & Yeager, 2015; Trilling & Fadel, 2009). Moreover, the narrowed curriculum has a lasting negative impact on students' opportunities to explore and develop other talents and find their passions (Abu-Alhija, 2007; Albrecht & Joles, 2003; Berliner, 2011; Klenowski, 2011; Klinger & Rogers, 2011; Popham, 2000; Towles-Reeves et al., 2006). Volante (2005) notes the excluded and reduced pathways are "no less worthy in their own right, these excluded areas are an essential

component of a student's educational experience. Once more, skills in domains such as visual arts, music, and physical education are vital for a large sector of the workforce" (p. 2). Rather, by not allowing students to pursue diverse subjects and lines of thinking, the divergent thinking, creativity, self-esteem levels, engagement, and desire for lifelong learning are all diminished and harmed (Berliner, 2011; Volante, 2005, 2006; Zhao, 2016c).

Demoralization and other psychological damages

Another side effect of LSAs on education is demoralization of education professionals and students. Most educators enter the education system with the strong desire and motivation to support students and encourage lifelong learning. LSAs add unique stresses and factors that diminish teachers' and leaders' motivation. School leaders report struggling between having more responsibility for outcomes while having less autonomy and authority in determining best practices and teacher effectiveness (Goodwin, Cunningham, & Childress, 2003; Prytula, Noonan, & Hellsten, 2013). Teachers suffer under such a regime even more, noting pressure-related stresses, frustration over the inconvenience and irrelevance of testing, diminished relationships with peers and leadership, and overall feelings of regret, shame, and failure, resulting in general disillusionment (Abu-Alhija, 2007; Berliner, 2011; Prytula et al., 2013). Additionally, in the light of the accountability mind-set, teachers and leaders associate students' performance with their own self-worth and effectiveness, notably in both high- and low-stakes environments, creating instead a culture of fear and regression where school faculty and staff feel they must do whatever necessary to make the grade (Abu-Alhija, 2007; Berliner, 2011; Klenowski, 2011; Klinger & Rogers, 2011; Popham, 2000; Volante, 2005, 2006).

Students, too, can be demoralized and suffer psychological damages. LSAs in many countries, such as China and Korea, are always criticized for bringing huge burden and stress to students (Ho, 2012). In countries where the higher education utilizes LSAs as the gatekeeper to access the higher level schools and only top performers can get the opportunity to enroll in prestigious universities, students experience years of obsession and high-intensity preparing for the test. The pressure not only comes from the exam itself but also from all aspects of society: their parents, teachers, schools, media, and policy makers (Chung & Chea, 2016; Kirkpatrick & Zang, 2011; Kristof, 2011), as the score is the only source to judge the worthiness of a student. In China, the National College Entrance Exam has always been criticized for directly causing a toxic level of stress. Over 80% of Chinese students reported significant worries about examinations, which lead to headaches and abdominal pain for them (Hesketh et al., 2010). In the 2015 PISA, comparing to other Organisation for Economic Co-operation and Development countries, Korean students are the unhappiest because of the burden of tests (So & Kang, 2014).

While stress may produce positive outcomes and improvement under some circumstances, unhealthy amounts of the pressures for students have become severe, being a detriment to both

students' physical and mental health. The pressures produced by negative exam-related experiences result in psychological stress and make students escape from the normal interaction with others, which are main factors leading to psychological problems and may have long-term impact in their whole life (Muthanna & Sang, 2015). Some students even start substance abuse of tobacco and alcohol to cope with such stress (Unger et al., 2001). It is reported that from 2013, over 90% suicide in elementary and middle school happened after a student had endured school-related stress (Deng, 2014). Exam-induced suicide has been reported in many countries and areas: China, Hong Kong China, Taiwan China, Korea, Singapore, Cambodia, Vietnam, and Japan (Cui, Cheng, Xu, Chen, & Wang, 2011). In addition, due to the high-stakes nature and very limited chances of the test, students contribute most of their time in study and even sacrifice the time for physical exercise. Staying in classroom to study instead of doing sports in physical education is a very common phenomenon in China, especially among students in the last year of high school, leading to severe obesity problem among these students.

Corruption and cheating

Another side effect LSAs have on education is moral corruption. The psychologist Donald Campbell coined Campbell's law: "The more any quantitative social indicator is used for social decision-making, the more subject it will be to corruption pressures and the more apt it will be to distort and corrupt the social processes it is intended to monitor" (Campbell, 1976, p. 49). As a quantitative indicator used for social decision-making, LSAs are not above the law. They lead to distortion and corruption.

One of the most obvious forms of corruption in education caused by LSAs is cheating on the tests. PISA has seen accusations of cheating by certain countries (Loveless, 2014; Sands, 2017). Cheating on the National College Entrance Exam and LSAs in China has been frequently reported (Kirkpatrick & Zang, 2011; Moore, 2013; Suen & Yu, 2006; Zhao, 2014). But cheating is not limited to China. Instead it happens in any country that LSAs are used to make consequential decisions. In the U.S., cheating on the SAT and ACT for college admissions has been going on for years, and in 2019, over 55 people were charged by federal prosecutors for various forms of cheating to gain entrance to elite colleges (Hassan, 2019).

Teachers and school leaders have engaged in unethical behaviors as well. Teachers have been found to use exact test prep materials (Towles-Reeves et al., 2006), use copies of tests or look-alike items for preparation (Albrecht & Joles, 2003; Popham, 2000; Volante, 2005), give test answer keys ahead of the exam (Popham, 2000; Volante, 2006), and relax timing or testing procedures (Albrecht & Joles, 2003; Berliner, 2011). What is more, at both the teacher and the school level, there is a trend of reducing the impact low-performing students may have on classroom and school data by discouraging their participation. Students are sent on field trips, encouraged to stay home,

or belittled and treated poorly in hopes they will transfer or drop out, all in hopes of gaining a few more points (Abu-Alhija, 2007; Albrecht & Joles, 2003; Berliner, 2011; Volante, 2006). The most visible cheating scandal in the U.S. is what happened in Atlanta Public Schools. In 2008, the state of Georgia flagged potential cheating at 58 Atlanta schools and confirmed cheating in 44. A total of 178 educators were named as participants, and more than 80 confessed. Many of them went to jail after one of the longest trials in Georgia's history (McCray, 2018).

LSAs that are meant to ensure equal access, best practices, and the most effective education have rather succeeded in destroying the empathy and talents of those within the educational systems, compromising their beliefs, ethics, and positive impact and potential. This is actually to be expected. Berliner (2011) references Campbell's law, noting, "In high pressure situations people frequently do whatever they deem necessary to achieve their goals and keep their jobs or status . . . [it] corrupts individuals, and the indicator [the high stakes assessment] itself may quickly become invalid" (p. 289).

Exacerbating inequity and injustice

A further negative side effect of LSAs is that they tend to exacerbate inequity in education. Vast educational inequity and injustice already exist due to socioeconomic, racial, geographical, cultural, and historical reasons. As a result, for no fault of their own, children have drastically different educational opportunities in terms of quality. It has long been established that children's family background is the largest determinant of their educational achievement and future success (Berliner, 2006; Coleman et al., 1966; Crane, 1996; Hanushek, 2016; Lee & Burkam, 2002; Zhang & Zhao, 2014). LSAs add to the inequity and injustice in a number of ways.

First, using results on LSAs to award opportunities for social mobility such as admissions to college is biased against students from disadvantaged backgrounds because these students typically perform worse than their peers from advantaged backgrounds due to unequal educational opportunities. LSAs have always favored the advantaged students. SAT scores, for instance, have a strong positive correlation with family income (Dixon-Román, Everson, & McArdle, 2013; Zwick & Greif Green, 2007), meaning that children of wealthier families score higher than children of lower income families. As a result, children from disadvantaged families have less of a chance to attend colleges or attend colleges of high prestige, hampering their opportunity for upward social mobility and exacerbating the existing inequity and injustice disadvantaged children face.

But LSAs do not have to be used to make such important decisions. The argument for using LSAs to make such life-changing decisions such as college admissions is based on the false belief of meritocracy and that LSAs can accurately measure people's merit to award social resources and positions (Zhao, 2016b, 2018a, 2018b). However, the purported meritocracy does not exist (Lemann, 2000), and LSAs cannot accurately capture the merit that will make one successful in

the future. For instance, the two most important college admissions assessments, SAT and ACT, have been found to be a poor predicator of success in college and life (Hiss & Franks, 2014; Kobrin, Patterson, Shaw, Mattern, & Barbuti, 2008; Noble & Sawyer, 2002). In China, the top scoring students are often not the most successful in life (Zhao, 2009).

Second, LSAs worsen the inequity and injustice in education because their results are used to drive policies and practices intended to help but in reality harm students. And the harm to disadvantaged children is more severe. For instance, results of LSAs have been the primary fuel that powered the "achievement gap mania" in the U.S. (Hess, 2011). LSAs have consistently revealed large gaps in scores among different groups of students. The gaps are primarily a result of socioeconomic and racial inequality and other factors that schools and teachers cannot control (Berliner, 2006; Ladson-Billings, 2006), but the U.S. government decided to use policies to hold schools and teachers accountable for closing the gaps (*No Child Left Behind Act of 2001*, 2002). And LSAs are used as a tool for measuring progress and hold teachers and schools to account. The result has been disastrous as discussed earlier in this article (Nichols & Berliner, 2007; Ravitch, 2010; Zhao, 2018d). The policies prescribed to close the gap did not work to close the gap in test scores but resulted in narrowed curriculum, distorted instruction, and demoralized educators. But the damage has been worse for disadvantaged children: curriculum narrowing, instructional distortion, demoralization, and corruption scandals happened more frequently in schools of disadvantaged students. The reductions in time spent in non-tested areas are statistically greater for those schools serving the poor as opposed to those schools serving the wealthy, exacerbating social segregation and inequities rather than reducing achievement gaps (Berliner, 2011). In other words, the efforts to close the achievement gap have widened the opportunity gap, creating more inequity and injustice (Tienken & Zhao, 2013).

Third, LSAs exacerbate inequity and injustice by damaging confidence and self-efficacy of disadvantaged students. Because of the narrowness of what they measure, LSAs rarely capture the strengths of students from disadvantaged background, but they often tell these students that they are no good as evidenced by their low scores. In other words, the poor students are constantly told that they are not good at anything and need remediation. These students are deprived of opportunities to discover their strengths because they must receive remediation on the tested subjects so they can score better on LSAs. As a result, poor children lose confidence, interest, and self-efficacy. They develop a strong sense of "learned helplessness" (Green, 1982; Weisz, 1981).

Stifling innovation in education

Stifling innovation in education is yet another side effect of LSAs. It is widely recognized that today's education is outdated. Education needs to change in order to meet the challenges of the transformation brought about by technology. Innovation is thus a necessity for bettering

education for the future (Barber, Donnelly, & Rizvi, 2012; Goldin & Katz, 2008; Wagner, 2008, 2012; Zhao, 2012).

LSAs stifle innovation in at least two ways. First, LSAs give policy makers and educators a rearview mirror, directing people to look backward (Zhao, 2016c, 2018c). LSAs are frequently used to identify effective educational policies and practices and encourage others to emulate these policies and practices. Many so-called evidence-based effective educational practices are identified based on the evidence of outcomes of LSAs.

Based on TIMSS and PISA, for example, policies and practices in countries and areas like Singapore, Finland, Korea, Shanghai, and other high-achieving education systems have been recognized as worthy of emulating by others (Barber & Mourshed, 2007; Jensen, 2012; Schleicher, 2018; Tucker, 2011). As a result, England decided to borrow math teaching from Shanghai (Wang, 2019). The testing regime popular in China, Korea, and Singapore was suggested for emulating in the U.S. (Tucker, 2011). The approach to cultivating teachers in Finland, Singapore, and Korea is recommended for copying by other systems (Barber & Mourshed, 2007).

While LSAs indeed provide evidence of success, the success is not necessarily in all aspects of education. Moreover, the evidence of success is achieved by policies and practices in the past. For example, whatever led to the top performance on PISA in Shanghai and Finland took place in the past. Hence encouraging others to emulate their policies and practices is encouraging others to look backward instead of forward. It is a race to the past (Zhao, 2015b, 2016c, 2018c). Unless other countries or areas would like to be like Shanghai or Finland in 2009 or 2001 when they were the top PISA performers in 2030 or 2040, copying their past does not make much sense.

LSAs stifle innovation because they make people believe that answers to the future already exist. Moreover, they further stifle innovation through homogenization. As people are convinced that effective educational policies and practices already exist and they only need to adopt them in their own contexts, they focus on looking for and adopting the policies and practices. As a result, there has been a global trend toward homogenizing policies and practices in education (Schleicher, 2018). Homogenization squeezes out possibilities for creativity and innovation (Zhao & Gearin, 2016).

So what: A final word

The thesis of this article is that LSAs have negative side effects. These side effects are inherent and thus cannot be removed through technical improvements. The side effects are real and have been observed. However, the side effects do not necessarily take away from the positive effects of LSAs. Thus, it is not our intention to argue for banning the use of LSAs in education. Rather, we raise the issue of side effects as a caution.

We want to caution policy makers not to overuse LSAs in education because of these side effects, just like doctors caution their patients to not overdose on certain medicine. We want to caution educators, students, and the public not to blindly believe results of LSAs so as to mediate the side effects on education. We want to caution advocates of LSAs that what they advocate can cause harm to education. Most importantly, we want to caution everyone in education not to use LSAs to measure everything. This is especially important at a time when new constructs such as creativity, entrepreneurial thinking, social emotional well-being, critical thinking, and others are beginning to be recognized as important educational outcomes, and it is tempting to develop measures for them. Not everything that counts can be counted!

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

References

Abu-Alhija, F. N. (2007). Large-scale testing: Benefits and pitfalls. *Studies in Educational Evaluation*, *33*, 50–68.

Albrecht, S. F., & Joles, C. (2003). Accountability and access to opportunity: Mutually exclusive tenets under a high-stakes testing mandate. *Preventing School Failure*, *47*, 86–91.

Baird, J.-A., Johnson, S., Hopfenbeck, T. N., Isaacs, T., Sprague, T., Stobart, G., & Yu, G. (2016). On the supranational spell of PISA in policy. *Educational Research*, *58*, 121–138.

Baker, J. (2019, April 12). People are frightened of NAPLAN: Australia's testing dilemma. *The Sydney Morning Herald*. Retrieved from https://www.smh.com.au/education/people-are-frightened-of-naplan-australia-s-testing-dilemma-20190411-p51dcu.html

Barber, M., Donnelly, K., & Rizvi, S. (2012). *Oceans of innovation: The Atlantic, the Pacific, global leadership and the future of education*. Retrieved from http://www.ippr.org/images/media/files/publication/2012/09/oceans-of-innovation_Aug2012_9543.pdf

Barber, M., & Mourshed, M. (2007). *How the world's best-performing school systems come out on top*. Retrieved from https://www.mckinsey.com/industries/social-sector/our-insights/how-the-worlds-best-performing-school-systems-come-out-on-top

Berliner, D. C. (2006). Our impoverished view of educational reform. *Teachers College Record*, *108*, 949–995.

Berliner, D. C. (2011). Rational responses to high stakes testing: The case of curriculum narrowing and the harm that follows. *Cambridge Journal of Education*, *41*, 287–302.

Bowen, S. K., & Rude, H. A. (2006). Assessment and students with disabilities: Issues and challenges with educational reform. *Rural Special Education Quarterly*, *25*, 24–30.

Campbell, D. T. (1976). *Assessing the impact of planned social change*. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.170.6988&rep=rep1&type=pdf

Chingos, M. M. (2012). *Strength in numbers: State spending on K-12 assessment systems*. Retrieved from https://www.brookings.edu/wp-content/uploads/2016/06/11_assessment_chingos_final_new.pdf

Chudowsky, N., & Pellegrino, J. W. (2003). Large-scale assessments that support learning: What will it take? *Theory into Practice*, *42*, 75–83.

Chung, P. J., & Chea, H. (2016). South Korea's accountability policy system and national achievement test. In W. C. Smith (Ed.), *The global testing culture: Shaping education policy, perceptions, and practice* (249–260). Oxford, England: Symposium Books.

Coleman, J. S., Campbell, E. Q., Hobson, C. J., McPartland, F., Mood, A. M., Weinfeld, F. D., & York, R. L. (1966). *Equality of educational opportunity*. Washington, DC: U.S. Government Printing Office.

College Board. (2019). SAT results: Class of 2018. Retrieved from https://reports.collegeboard.org/sat-suite-program-results/class-2018-results

Crane, J. (1996). Effects of home environment, SES, and maternal test scores on mathematics achievement. *The Journal of Educational Research*, *89*, 305–314.

Cui, S., Cheng, Y., Xu, Z., Chen, D., & Wang, Y. (2011). Peer relationships and suicide ideation and attempts among Chinese adolescents. *Child: Care, Health and Development*, *37*, 692–702.

Deng, C. (2014). China's cutthroat school system leads to teen suicides. Retrieved from https://blogs.wsj.com/chinarealtime/2014/05/15/chinas-cutthroat-school-system-leads-to-teen-suicides/

Dewey, J. (1975). *Democracy and education: An introduction to the philosophy of education*. New York, NY: Free Press.

Dillion, S. (2010, December 7). Top test scores from Shanghai stun educators. *New York Times*. Retrieved from http://www.nytimes.com/2010/12/07/education/07education.html?pagewanted=1&_r=2

Dixon-Román, E. J., Everson, H. T., & McArdle, J. J. (2013). Race, poverty and SAT scores: Modeling the influences of family income on black and white high school students' SAT performance. *Teachers College Record*, *115*, 1–33.

Duckworth, A. L., & Yeager, D. S. (2015). Measurement matters: Assessing personal qualities other than cognitive ability for educational purposes. *Educational Researcher*, *44*, 237–251.

Dulfer, N. (2012, November 25). Testing the test: NAPLAN makes for stressed kids and a narrow curriculum. Retrieved from http://theconversation.com/testing-the-test-naplan-makes-for-stressed-kids-and-a-narrow-curriculum-10965

Every Student Succeeds Act. (2015). Public Law No. 114-95, Congress.

Figazzolo, L. (2009). Impact of PISA 2006 on the education policy debate. Retrieved from http://download.ei-ie.org/docs/IRISDocuments/ResearchWebsiteDocuments/2009-00036-01-E.pdf

Florida, R. (2012). *The rise of the creative class: Revisited* (2nd ed.). New York, NY: Basic Books.

Goldin, C., & Katz, L. F. (2008). *The race between education and technology*. Cambridge, MA: Harvard University Press.

Goodwin, R. H., Cunningham, M. L., & Childress, R. (2003). The changing role of the secondary principal. *NASSP Bulletin*, *87*, 26–42.

Green, L. (1982). A learned helplessness analysis of problems confronting the black community. In S. Turner & R. Jones (Eds.), *Behavior modification in black populations* (pp. 73–93). New York, NY: Springer US.

Gruber, K. H. (2006). *The German "PISA-shock": Some aspects of the extraordinary impact of the OECD's PISA study on the German education system.* Paper presented at the Cross-national attraction in education: Accounts from England and Germany. Oxford, England: Symposium Books.

Hanushek, E. A. (2016). What matters for student achievement: Updating Coleman on the influence of families and schools. *Education Next*, *16*, 18–26.

Hassan, A. (2019, March 15). Exam-taking stand-ins and answers on pencils. *The New York Times*. Retrieved from https://www.nytimes.com/2019/03/15/us/college-scams-admissions.html?rref=collection%2Fnewseventcollection%2Fcollege-admissions-scandal&action=click&contentCollection=us&region=stream&module=stream_unit&version=latest&contentPlacement=7&pgtype=collection

Hesketh, T., Zhen, Y., Lu, L., Dong, Z., Jun, Y., & Xing, Z. (2010). Stress and psychosomatic symptoms in Chinese school children: Cross-sectional survey. *Archives of Disease in Childhood*, *95*, 136–140.

Hess, F. M. (2011, Fall). Our achievement-gap mania. *National Affairs*, *9*, 113–129.

Hiss, W. C., & Franks, V. W. (2014). *Defining promise: Optional standardized testing policies in American college and university admissions*. Retrieved from http://www.nacacnet.org/research/research-data/nacac-research/Documents/DefiningPromise.pdf

Ho, S. E. (2012). *Asia-Pacific education system review series no. 5: Student learning assessment*. Hong Kong, China: UNESCO.

Hout, M., & Elliott, S. W. (Eds.). (2011). *Incentives and test-based accountability in education*. Washington, DC: National Academies Press.

Jensen, B. (2012). *Catching up: Learning from the best school systems in East Asia*. Retrieved from http://www.grattan.edu.au/publications/129_report_learning_from_the_best_main.pdf

Karsten, S., Visscher, A., & De Jong, T. (2001). Another side to the coin: The unintended effects of the publication of school performance data in England and France. *Comparative Education*, *37*, 231–242.

Kirkpatrick, R., & Zang, Y. (2011). The negative influences of exam-oriented education on Chinese high school students: Backwash from classroom to child. *Language testing in Asia*, *1*, 36.

Klenowski, V. (2011). Assessment for learning in the accountability era: Queensland, Australia. *Studies in Educational Evaluation*, *37*, 78–83.

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

Kobrin, J. L., Patterson, B. F., Shaw, E. J., Mattern, K. D., & Barbuti, S. M. (2008). *Validity of the SAT® for predicting first-year college grade point average*. Retrieved from http://research.collegeboard.org/sites/default/files/publications/2012/7/researchreport-2008-5-validity-sat-predicting-first-year-college-grade-point-average.pdf

Kristof, N. (2011). China's winning schools? *The New York Times*. Retrieved from https://www.nytimes.com/2011/01/16/opinion/16kristof.html

Labaree, D. F. (1997). Public goods, private goods: The American struggle over educational goals. *American Educational Research Journal*, *34*, 39–81.

Ladson-Billings, G. (2006). From the achievement gap to the education debt: Understanding achievement in U.S. schools. *Educational Researcher*, *35*, 3–12.

Lee, V. E., & Burkam, D. T. (2002). *Inequality at the starting gate: Social background differences in achievement as children begin school*. Washington, DC: Economic Policy Institute.

Lemann, N. (2000). *The big test: The secret history of the American meritocracy*. New York, NY: Farrar, Straus and Giroux.

Levin, H. M. (2012). More than just test scores. *Prospects: The Quarterly Review of Comparative Education*, *42*, 269–284.

Lewis, S., & Lingard, B. (2015). The multiple effects of international large-scale assessment on education policy and research. *Discourse: Studies in the Cultural Politics of Education*, *36*, 621–637.

Loveless, T. (2014). *PISA's China problem continues: A response to Schleicher, Zhang, and Tucker*. Retrieved from http://www.brookings.edu/research/papers/2014/01/08-shanghai-pisa-loveless

McCray, V. (2018). Altered test scores years ago altered lives, stained Atlanta schools. *The Atlanta Journal-Constitution*. Retrieved from https://www.ajc.com/news/local-education/altered-test-scores-altered-lives-stained-atlanta-schools/nFHhI3jPSQ7MjIS9dRuCNM/

Menken, K. (2006). Teaching to the test: How no child left behind impacts language policy, curriculum, and instruction for English language learners. *Bilingual Research Journal*, *30*, 521–546.

Moore, M. (2013, June 20). Riot after Chinese teachers try to stop pupils cheating. *The Telegraph*. Retrieved from http://www.telegraph.co.uk/news/worldnews/asia/china/10132391/Riot-after-Chinese-teachers-try-to-stop-pupils-cheating.html

Muthanna, A., & Sang, G. (2015). Undergraduate Chinese students' perspectives on Gaokao examination: Strengths, weaknesses, and implications. *International Journal of Research Studies in Education*, *4*. doi: 10.5861/ijrse.2015.1224

Nichols, S. L., & Berliner, D. C. (2007). *Collateral damage: How high-stakes testing corrupts America's schools*. Cambridge, MA: Harvard Education Press.

No Child Left Behind Act of 2001. (2002). Public Law No. 107–110, Congress.

Noble, J., & Sawyer, R. (2002). *Predicting different levels of academic success in college using high school GPA and ACT composit score*. Retrieved from http://www.valees.org/documents/ACT_grades_predictors_of_success.pdf

Organisation for Economic Co-operation and Development. (2016a). PISA 2015: Results in focus. *PISA*. Retrieved from https://www.oecd.org/pisa/pisa-2015-results-in-focus.pdf

Organisation for Economic Co-operation and Development. (2016b). *PISA 2015 results (Volume I): Excellence and equity in education*. Paris: OECD Publishing.

Pink, D. H. (2006). *A whole new mind: Why right-brainers will rule the future*. New York, NY: Riverhead.

PISA. (n.d.). How to join PISA. Retrieved from http://www.oecd.org/pisa/aboutpisa/howtojoinpisa.htm

Popham, W. J. (1999). Where large scale educational assessment is heading and why it shouldn't. *Educational Measurement: Issues and Practice*, *18*, 13–17.

Popham, W. J. (2000). Big change questions: "Should large-scale assessement be used for accountability?" – Answer: Depends on the assessment, silly!. *Journal of Educational Change*, *1*, 283–289.

Prytula, M., Noonan, B., & Hellsten, L. (2013). Toward instructional leadership: Principals' perceptions of large-scale assessment in schools. *Canadian Journal of Educational Administration and Policy*, *140*, 1–30.

Ravitch, D. (2010). *The death and life of the great American school system: How testing and choice are undermining education*. New York, NY: Basic Books.

Sands, G. (2017, January 4). Are the PISA education results rigged? Retrieved from https://www.forbes.com/sites/realspin/2017/01/04/are-the-pisa-education-results-rigged/-792d5f4d1561

Schleicher, A. (2018). *World class: How to build a 21st-century school system*. Paris, France: OECD.

Sellar, S., & Lingard, B. (2014). The OECD and the expansion of PISA: New global modes of governance in education. *British Educational Research Journal*, *40*, 917–936.

Sjøberg, S. (2015). OECD, PISA, and globalization: The influence of the international assessment regime. In C. H. Tienken & C. A. Mullen (Eds.), *Education policy perils. Tackling the tough issues* (pp. 102–133). New York, NY: Routledge.

So, K., & Kang, J. (2014). Curriculum reform in Korea: Issues and challenges for twenty-first century learning. *The Asia-Pacific Education Researcher*, *23*, 795–803.

Suen, H. K., & Yu, L. (2006). Chronic consequences of high-stakes testing? Lessons from the Chinese civil service exam. *Comparative Education Review*, *50*, 46–65.

Tienken, C. H., & Zhao, Y. (2013). How common standards and standardized testing widen the opportunity gap. In P. L. Carter & K. G. Welner (Eds.), *Closing the opportunity gap: What America must do to give every child an even chance* (pp. 113–122). New York, NY: Oxford University Press.

Towles-Reeves, E., Garrett, B., Burdette, P. J., & Burdge, M. (2006). Validation of large-scale alternate assessment systems and their influence on instruction—What are the consequences? *Assessment for Effective Intervention*, *31*, 45–57.

Trilling, B., & Fadel, C. (2009). *21st century skills: Learning for life in our times*. Hoboken, NJ: John Wiley & Sons.

Tucker, M. (Ed.). (2011). *Surpassing Shanghai: An agenda for American education built on the world's leading systems*. Boston, MA: Harvard Education Press.

Unger, J. B., Li, Y., Johnson, C. A., Gong, J., Chen, X., Li, C., . . . Lo, A. T. (2001). Stressful life events among adolescents in Wuhan, China: Associations with smoking, alcohol use, and depressive symptoms. *International Journal of Behavioral Medicine*, *8*, 1–18.

Volante, L. (2005). Accountability, student assessment, and the need for a comprehensive approach. *International Electronic Journal for Leadership in Learning*, *9*, 1–8.

Volante, L. (2006). An alternative vision for large-scale assessment in Canada. *Journal of Teaching and Learning*, *4*, 1–14.

Wagner, T. (2008). *The global achievement gap: Why even our best schools don't teach the new survival skills our children need—And what we can do about it*. New York, NY: Basic Books.

Wagner, T. (2012). *Creating innovators: The making of young people who will change the world*. New York, NY: Scribner.

Wang, M. (2019, March 5). Could do better: Shanghai-style maths teaching in UK. *The Telegraph*. Retrieved from https://www.telegraph.co.uk/news/world/china-watch/culture/shanghai-maths-teaching/

Weisz, J. R. (1981). Learned helplessness in Black and White children identified by their schools as retarded and nonretarded: Performance deterioration in response to failure. *Developmental Psychology*, *17*, 499–508.

Yu, L., & Suen, H. K. (2015). Historical and contemporary exam-driven education fever in China. *KEDI Journal of Educational Policy*, *2*, 17.

Zhang, G., & Zhao, Y. (2014). Achievement gap in China. In J. V. Clark (Ed.), *Closing the achievement gap from an international perpsective: Transforming STEM for effective education* (pp. 217–228). New York, NY: Springer.

Zhao, Y. (2009). *Catching up or leading the way: American education in the age of globalization*. Alexandria, VA: ASCD.

Zhao, Y. (2012). *World class learners: Educating creative and entrepreneurial students*. Thousand Oaks, CA: Corwin.

Zhao, Y. (2014). *Who's afraid of the big bad dragon: Why China has the best (and worst) education system in the world*. San Francisco, CA: Jossey-Bass.

Zhao, Y. (2015a). A world at risk: An imperative for a paradigm shift to cultivate 21st century learners. *Society*, *52*, 129–135.

Zhao, Y. (2015b). *Lessons that matter: What we should learn from Asian school systems*. Retrieved from http://www.mitchellinstitute.org.au/reports/lessons-that-matter-what-should-we-learn-from-asias-school-systems/

Zhao, Y. (2016a). *Counting what counts: Reframing education outcomes*. Bloomington, IN: Solution Tree Press.

Zhao, Y. (2016b). From deficiency to strength: Shifting the mindset about education inequality. *Journal of Social Issues*, *72*, 716–735.

Zhao, Y. (2016c). Who's afraid of PISA: The fallacy of international assessments of system performance. In A. Harris & M. S. Jones (Eds.), *Leading futures* (pp. 7–21). Thousand Oaks, CA: SAGE.

Zhao, Y. (2017). What works can hurt: Side effects in education. *Journal of Educational Change*, *18*, 1–19.

Zhao, Y. (2018a). Personalizable education for greatness. *Kappa Delta Pi Record*, *54*, 109–115.

Zhao, Y. (2018b). *Reach for greatness: Personalizable education for all children*. Thousand Oaks, CA: Corwin.

Zhao, Y. (2018c). Shifting the education paradigm: Why international borrowing is no longer sufficient for improving education in China. *ECNU Review of Education*, *1*, 76–106.

Zhao, Y. (2018d). *What works may hurt: Side effects in education*. New York, NY: Teachers College Press.

Zhao, Y. (2019). The rise of the useless: The case for talent diversity. *Journal of Science Education and Technology*, *28*, 62–68.

Zhao, Y., & Gearin, B. (2016). Squeezed out: The threat of global homogenization of education to creativity. In D. Ambrose & R. J. Sternberg (Eds.), *Creative intelligence in the 21st century* (pp. 121–138). Rotterdam, Netherlands: Sense.

Zwick, R., & Greif Green, J. (2007). New perspectives on the correlation of SAT scores, high school grades, and socioeconomic factors. *Journal of Educational Measurement*, *44*, 23–45.