

Testing like William the Conqueror

Abstract

The spread of academic testing for accountability purposes in multiple countries has obscured at least two historical purposes of academic testing: community ritual and management of the social structure. Testing for accountability is very different from the purpose of academic challenges one can identify in community “examinations” in 19th century North America, or exams' controlling access to the civil service in Imperial China. Rather than testing for ritual or access to mobility, the modern uses of testing are much closer to the state-building project of a tax census, such as the *Domesday Book*, the social engineering projects described in James Scott's *Seeing like a State* (1998) or the “mapping the world” project that David Nye described in *America as Second Creation* (2004). This paper will explore both the instrumental and cultural differences among testing as ritual, testing as mobility control, and testing as state-building.

Testing like William the Conqueror

Today, supporters and opponents of test-based accountability argue in the public international discourse of education reform that the debate over accountability is an explicit morality tale. Advocates of test-based accountability in the United States and elsewhere argue that such accountability is required for human-capital development and to satisfy equity concerns. In the testing-policy advocates' morality tale, opponents are often painted as self-interested "defenders of the status quo." In contrast, Finnish educator Pasi Sahlberg (2011) has argued that a global educational reform movement (or GERM) is an engulfing army that invades various countries and replaces alternative models of education as a public good. In Sahlberg's morality tale, Finland and some other outposts are lonely, brave opponents of an educational colonialism.

A morality-tale approach to accountability policy obscures important alternative perspectives on the discourse. This article focuses on both the instrumental and cultural uses of testing in a range of countries and periods. From a broader view, one can see three dominant uses of testing which present as either instrumental or cultural in orientation. However, though advocates of the three dominant uses have focused on either cultural or instrumental goals, in reality any use has both cultural and instrumental uses. For the purposes of this perspective, I broadly define "testing" as the formal assessment of examinee (commonly student) skills and achievement constructed and evaluated by non-examinees. In defining testing in this way, I borrow John Calhoun's (1973) description of intelligence as audience-defined: assessment of students can be formal or informal, conducted by outsiders or by students themselves (as

children who play competitive games can well attest). But if we consider testing as a formal outsider-driven process, then the formal testing administered by adults tells us a great deal about the purposes of schooling from adult, outsider perspectives.¹ Doing so also avoids restricting discussion to eras and educational regimes that have access to particular technologies such as standardized, paper-and-pencil tests.

The historical purposes of testing²

{Need an orienting/introducing paragraph here.}

Ritual. At least at a first approximation, testing serves one of three purposes: cultural ritual, status gatekeeping, or state building. An example of testing as cultural ritual is the end-of-session public examination in many early 19th century village (“district”) schools in the U.S. North. As explained by William Reese (2013), end-of-session exams were community events where adults witnessed both the teacher and the pupils reciting texts, answering questions, and otherwise engaged in ritualized performances. Long before Helen and Robert Lynd (1929) described Indiana small-town basketball as community glue, and long before Mary Metz (1991) explained how cultural “scripts” of high school imposed institutional isomorphism, American schools commonly followed a testing script that was focused on community and ritual. This common ritual was not universal: certainly, many male students beat teachers into a hasty retreat

1

¹ The complementary perspective, by children, would be well worth the effort, probably best conducted by anthropologists.

2

¹ Other uses of testing are not considered here. Testing for individual diagnostic purposes and testing for formative assessment (or similar instructional-guidance purposes) are recent phenomena--they may be able to be incorporated in the analysis here, but they are outside the scope of this paper.

from a town in the 19th century, but it was common enough to be a recognized element in American schooling in the early 19th century. This ritual lives on in the Scripps National Spelling Bee and similar competitions.

Gatekeeping. A second purpose of testing is as a status gatekeeper, filtering applicants for positions in a social structure. Mazzeo (2001) calls this role the guidance function of testing when it occurs inside school structures, though the same function can also exist as a filter lying between levels of formal schooling, such as university entrance examinations in a broad range of industrialized countries. For heuristic purposes, we can emphasize external relationships between testing and the broader social structure. Two prominent examples of this purpose are the civil service exam in Imperial China and the construction of IQ exams in the early 20th century in the United States. For hundreds of years, a formal examination served as a minimum gatekeeper for civil service status in China. The importance of the examination lasted through a number of regime changes, became a prominent part of Chinese social rigidity, and provoked the creation and maintenance of a system of private tutoring of applicants. In the United States, the importation and translation of so-called intelligence tests in the 20th century served as one tool by which urban public school systems divided students into curricular tracks or streams. Consistently associated with race, class, and gender divides, IQ-based tracking decisions limited educational opportunities for millions in the first half of the 20th century. In both imperial China and 20th century North America, a significant purpose of testing was to serve as a status gatekeeper.

State building. A third purpose of testing is to serve as part of state capacity-building. As Reese (2013) explains, early advocates of standardized testing in American common schools thought it could be used to put pressure on schools to improve, either directly or indirectly (the

latter through competitive exams to enter secondary schools). More recently, advocates of test-based accountability have encouraged the development of additional state capacity to manage both the process of testing and the analysis, dissemination, and use of those test results to manage schools and school personnel. It is important to understand the parallels between test-based accountability and the types of state-building projects that James Scott (1998) describes in *Seeing like a State*. While Scott focuses on modern state-building and instrumental (and largely failed) utopianism, the activity of data collection is much broader and older. The construction of the *Domesday Book* on the orders of William the Conqueror is an example of the type of state-building data collection that predates both industrialization and the modern nation-state. Without a rigid definition of state-building, it is sufficient for now to note how accounting activities serve both central government functions such as tax collection and also provide an intimate connection between central governments and the everyday life of people.

The cultural uses of testing

While testing as ritual, testing as gatekeeper, and testing as state apparatus appear to be very different in character, in some ways one can describe them all as having both cultural uses and also instrumental uses. The cultural use of testing exists even when testing is not used for ritualistic purposes. That stems from the fact that instrumental functions are generally tied to particular discourses around education, and also the ways in which the experience of being tested generates common touchpoints for cultural artifacts. Certain school rituals may be bounded by time and place, and it may be a matter of some irony that more standardized, authority-driven, and common experiences can be the genesis or inspiration of more cultural expression than less-standardized school practices. But that is a feature of the ideological role of authority, not an accident. As David Nye observed (2004), the act of surveying townships in the young United

States was as much an assertion that the United States was outside history, remaking North America, as the surveying was a practical act of governance. Measuring and using power is not just a matter of “seeing like a state” (Scott, 1998) but defining and communicating like a state as well. Those acts invariably become part of the cultural history of examination, and this section describes three cultural uses of testing.

Affirmation of community. As described above, the use of testing for ritualistic purposes in schooling may take the form of public performances, in which case one cultural use of such testing is to affirm the worth and nature of the community. Consider the ritualistic school performances in the young United States. In enforcing a public performance that invited interaction, such public “examination” validated the connection between schooling and the local adults, the power of local adults over the schoolmaster, and the experience of pupils during the school session. Such validation could also exist with the discourses surrounding testing for gatekeeping or testing for state authority. In the case of gatekeeping, the discourse of testing and tracking has often involved an affirmation of community by exclusion: with admissions tests for selective secondary schools and universities, for example, the students within the school are identified as being both worthy of the school’s education and simultaneously a member of the select community of the school. More generally, the discourse of testing as gatekeeping across time and place has often revolved around meritocratic ideologies--if not the meritocracy of an entire society, at least the competence, merit, and virtue of those who pass the tests and become the members of the Chinese imperial bureaucracy, the graduates of English universities, or members of professions with limited access in the last century.

While the discourse of testing as gatekeeping emphasizes the community of successful examinees, the discourse of testing for state-building emphasizes the community of citizens--in

human-capital rhetoric, contributing adults or children presumed to be future contributing adults. One variant of this discourse of community is the argument for high-stakes testing for equity purposes: in this case, the role of testing is not to build citizens but to guarantee the rights bestowed by citizenship. One final role of citizenship tied to testing-as-state-building involves the citizen as taxpayer: what are taxes purchasing when they are spent on schooling? In the United States, this discourse began no later than the mid-1960s and became a prominent reason for the construction of state assessment systems in the 1970s (e.g., AUTHOR, DATE BLINDED).

Normalization. A second cultural purpose of testing has been to establish normative boundaries, beyond the definition of a community. The setting of cultural norms is a messy, complicated process, and it may help to think of testing as a common experience that both prepares and can be used for setting expectations. For ritual exams, those expectations may have been minimal (can one recite a poem?), but they need not be. In contrast with the normalization of performance standards provided by ritual examinations, testing for gatekeeping has frequently generated a discourse around the normalization of the social structure, either the existing structure or a desired meritocratic structure of the future. As Dreeben (1968) argued, one historical role of schools has been to strip away the veneer of self-esteem provided by students' families, teaching students that their personal worth is to be judged by a narrow range of competencies. Kliebard (1986) describes an ideology of "social efficiency" tied to IQ testing and tracking in the early 20th century in the United States. But not all testing for gatekeeping purposes has been tied to justifying the existing social structure. Lemann (2000) argued that the growing use of the SAT for college admissions in the United States several decades later served the interests of elite university administrators who wanted to generate a national student market

they could identify as talented. It is important to note that in Lemann's view, the vision of elite university administrators was *not* one of broad access to elite education but equal opportunity, and a reconfirmation of the status of elite universities for a modern economy.

In addition, the discourse around testing for accountability has used and fed into norms about schooling. In a classic case of policy feedback, former Florida Governor Jeb Bush pushed for a policy in 1999 where the state labeled each local public school with a letter grade, A through F, akin to the North American letter grades assigned pupils in school subjects. These so-called grades have been based largely on student performance on mandated state tests, and the relative generosity of A and B grades applied to primary schools encouraged principals of those schools to publicly declare themselves as an "A school" or "B school" using roadside signs. Education reporters have casually used the grades as adjectives in like manner, including "failing school" as a term applied to schools with a state label of F. In generating this discourse, the A-F labeling policy normalizes test-based accountability as part of the "real school" cultural script (Metz, 1991). In doing so, the A-F labeling policy is more politically robust than an alternative labeling scheme would have been.³ In recent years, many other states and some cities have tried to adopt Florida's policy. Because the other jurisdictions have been less generous with the higher labels, it is uncertain whether they will be as politically robust.

Cultural referent. A third manner in which testing has cultural uses is when the experience of testing has become a touchpoint for cultural expression. The Scripps National Spelling Bee in the U.S. became the subject of the documentary *Spellbound* (Blitz, 2002). Often

³ Since the Republican party in Florida has held onto the governor's seat and both houses of the legislature since 1999, that political stability has also ensured continuity of the labeling policy.

enough, spelling contests among adults (not children) have been the fodder for fiction, as in the “Spelling-School” in Edward Eggleston’s *Hoosier Schoolmaster* (1871) or spelling contests at “Literaries” in Laura Ingalls Wilder’s fictionalized stories from her childhood. Currently, the Groot Dictee der Nederlandse Taal is a televised dictation challenge for adults in the Low Countries that has run since 1991. In children’s literature, examination rituals are often transformed: The “Spelling Bee” is a character (not an event) in Norton Juster’s child fantasy *The Phantom Tollbooth* (1961), and English-style ‘O’ and ‘A’ level exams became O.W.L. (Ordinary Wizarding Level) and N.E.W.T. (Nastily Exhausting Wizard Test) in J.K. Rowling’s Harry Potter series. The suspense involved in examinations have worked their way into fiction when the tests have a gatekeeping or state-building function: in the movie *Stand and Deliver* (1988), the climax comes when math teacher Jaime Escalante’s students retake an Advanced Placement exam after they were accused of cheating because their scores were *too* good in comparison with stereotyped expectations of Latino students.⁴ In contrast with the serious treatment of AP exams in *Stand and Deliver*, Wu Ching-tzu’s eighteenth-century novel *The Scholars* satirized the imperial Chinese civil service exam, poking fun at perennial candidates standing for the exam, the network of private tutors benefitting from the existence of the exam, and the elaborate warping of social prestige conferred by exam passage (see, e.g., Miyazaki, 1976). Ching-tzu’s novel is a long form of satire that was common in the later imperial period; P’u Sung-ling had written a number of shorter satires of the examination system around 1700 (Elman, 2000).

4

¹ They passed the second time, too.

In all of these ways, testing as an activity and examination systems as social phenomena are generators or objects of cultural interpretation. The definition of community by either inclusion or exclusion is a cultural use of testing that is inscriptive in nature, delineating boundaries and qualities at a local level. More imposing is the use of testing for the creation or reformation of social norms. Definition and norming are both generative activities, while using test experiences as a springboard for cultural creation is the use of testing as an object of expression, of arguments about the tests and test systems themselves. It is important to note the ways in which apparently “functional” purposes of testing can still become part of cultural expression, either in the discourse surrounding the function or as an object of commentary and social criticism.

Instrumental uses of testing

If one side of examinations is the cultural use of testing, then the other side is a functional use of testing, one associated more with blunt application of instrumental power than with discourse. The division between the two is somewhat arbitrary--many would see the normative purpose described above as instrumental. The distinction is still helpful to illustrate the broader point that testing is a “dual-use” activity in terms of both application of power and interaction with cultural expression. Again, this is true whether one is discussing the apparently-instrumental purposes of gatekeeping or state authority, on the one hand, or the “softer” use of testing for ritual purposes, on the other. To illustrate these dynamics, this section discusses four instrumental uses of testing: sorting at the micro or macro level, intervening in the lives of a school, enforcing a curriculum, and enforcing a language.

Sorting. One function of testing is to sort. The sorting can be for purposes of social prestige, the type of informal ranking that is loosely defined but important at a local level, or for

formal access (or denial of access) to opportunities in jobs or further education. Examinations for ritual purposes have also served as a performative delineation of social status or basis for judgment of prestige by others. The gatekeeping purpose of testing explicitly sorts examinees as individuals. Beyond the sorting of individuals, the gatekeeping role of testing also sorts opportunities for society more broadly. Testing is thus an implicit method to shape the social structure, including social classes. European countries with examination-based secondary tracking systems have done so implicitly for several decades, approximately one half-century after North American schools expanded tracking along with the expansion of high schools around the turn of the twentieth century. Examination-based entry into professions is a more explicit form of shaping social structure at least as far as advantaged and prestigious occupations are concerned, whether the imperial civil service or more modern-era professions such as medicine and law. Whether justified in softer terms such as “guidance” (see Mazzeo, 2001), in firmer terms of implied fixed ability, or in the slippery definition of deviance (e.g., Johnson, 1968), gatekeeping assessment systems in a range of times and places have become part of the repertoire for generating and regenerating a social structure.

While one may think of test-based sorting in terms of its effects on examinees, one important legacy of sorting students is the dual use of such mechanisms, with the capacity to sort and rank their teachers as well. Even at the level of informal sorting of social prestige, this type of function can be levied as much against a teacher as students, as in a music teacher’s ensemble as an indicator of her or his value among other music educators. But even before the recent development of formal test-based accountability systems, testing has been used as a form of sorting for teachers. In England, Payment by Results long predated the post-WW2 interest in so-called merit pay (e.g., Rapple, 1994). Such inclination to judge and sort teachers predated

Payment by Results: as early high schools in North America often used blindly-graded entrance exams, those results became one potential filter by which local grammar-school principals were judged if they applied to teach at Philadelphia Central High School (Labaree, 1988).

Intervening. A second instrumental use of testing has been in deciding to intervene in local school matters, whether at the level of the child, the classroom, or the school. While teachers have based informal instructional judgments on classroom assessments of performance for many decades in different societies, one finds it difficult to draw a clear distinction between the use of such judgment as intervention in contrast with sorting. In Joseph Lancaster's monitorial school model in the early 19th century, frequent testing would decide on the role of pupils within a class as well as placement in recitation groups (e.g., Kaestle, 1973). Lancaster's use of assessment looks much closer to sorting than deliberative instructional adjustment. Closer to the mark is the practice of retaining or promoting students between curriculum levels, using examinations to do so. In the United States, such actions could be seen beginning in late nineteenth-century cities with the evolution of age-grading. While the research on grade retention is mixed at best, there is a plausible interventionist argument to make about retention in grade, or (less common) advancement in grades and tracks or streams in the middle of school years.⁵

As with sorting, the use of testing for interventions extends beyond the examinees to schools and employees. One source of such pressures has been the school system itself as an organization, where test-based accountability and so-called league tables can create political pressure on school administrators to make significant changes in schools even without explicit

⁵ As indicated earlier, I will not be discussing more recent research on the use of formative assessment, in part because of that recency and in addition because it is uncertain how extensive such practices are (e.g., AUTHOR, DATE BLINDED).

mandates such as those embodied in No Child Left Behind in the United States. Another source of such pressures has been parents. As Labaree (1997) notes, the private uses of schools includes the role of schools in promoting economic advancement by students and their families. Preparing students for examinations is thus a frequent expectation of families with significant advantages where those examinations serve gatekeeper purposes. While preparation for exams can be private such as post-WW2 *juku* or cram schools in Japan, parents have also made judgments about public schools from the use of both entrance exams in postwar university admissions and also examination-based credit for college such as the United States Advanced Placement system. In recent decades, for example, wealthy or middle-class parents have held increasing expectations that high schools would offer Advanced Placement courses, thus preparing students for exams that can provide some college credit. In this context, parental pressure on schools to offer a certain curriculum can be seen as a form of intervention.

Enforcing a curriculum. A third instrumental use of testing has been in the enforcement of what examinees should learn. For example, for an early state commissioner of education in the U.S., Henry Barnard, entrance exams to high schools would "operate as a powerful and abiding stimulus to exertion throughout all the lower schools" even if only a minority of students attended high school (Barnard, 1865, p. 283). Though competitive admissions exams served a gatekeeping purpose on the surface, Barnard thought they could help with the control of primary and intermediate schools. The use of exams to enforce a minimal curriculum has become much more transparent in the past half-century, as central state curriculum planning has envisioned a tightly-linked connection between mandated curriculum, testing, and accountability. Professional licensure exams have encouraged the standardization not only of formal, large schooling, but also private test preparation--examinations for new lawyers generally prompt the type of

preparation and private tutoring that is as intense as law school classes. The putative purpose of such licensure exams is to guarantee a minimal level of professional knowledge for lawyers, medical personnel, those who work on electrical and plumbing systems, and so forth, and it is somewhat reassuring to know that a licensed electrician has to demonstrate at least a minimal knowledge of electrical regulatory code to acquire the right to wire a building.

It is important to understand that the enforcement of a curriculum is not just the domain of apparently-functional purposes for testing. In ritual examinations, the norms have involved both the standards of performance and the implied curriculum--in early federal America, recitation and other acts of memorization were validated by end-of-session public examination. To take a modern example, musical performance festivals for school ensembles are both a demonstration of performance (and implied leadership of music teachers) and also a confirmation of canonical material judged to be appropriate. The organization of such festivals has required securing rotating sites on school grounds and publicity of the results, and they have thus also been an affirmation of the role of music education in the curriculum in an era when arts education is unvalidated by state-mandated assessments, at least in the United States.

Enforcing a national language. At the intersection of curriculum and national identity is the role of schools in generating or enforcing language. In multiple contexts, the content and language of examinations has played a role in both the cultural authority of a language (e.g., spelling or dictation challenges) and the relationship between local communities and new nations. In some countries such as France or the United States, the presumed dominance of a desired single language is still an object of enforcement.⁶ In postcolonial nations, the

6

¹ This essay does not analyze the discursive role of cultural authority bodies such as L'Académie française for French or advocates such as Noah Webster, a language nationalist for the young United States.

management of multiple native and legacy colonial languages have been both a logistical and a cultural object of enforcement through schools. In countries as varied as Zambia and Papua New Guinea, the language of instruction and examination has been a matter of conscious policy, and the overlay of colonial examination systems, with the language of examinations, have on occasion become embedded in the internal status contests of the new nation.⁷

This classification of instrumental uses of testing is highly tentative, focusing on selected uses that stretch across multiple times, places, and explicit purposes of examinations. This analysis excludes formative assessment because of its recency of use and theorizing. The broader argument remains: the explicit dominant purpose of a testing regime does not eliminate either the cultural uses of testing or the instrumental uses. In the practices described here, the uses are generated from each type of testing. The discourse and instrumental uses of test-based accountability thus are not the unique or entirely new phenomenon that Sahlberg identified as his GERM. Rather, the history of examination for a range of purposes included both instrumental and cultural outlets. We may be experiencing a wave of testing that is organized as state control in a new way, but both the use of power and the cultural expressions surrounding test-based accountability is old. One can find schools in many countries asserting that they have accountability for the postindustrial global economy, but they remain close to testing like William the Conqueror.

Why has state-building dominated recently?

7

⁷ For an example of the practical considerations, see Delpit's (1995) description of education policy in newly-independent Papua New Guinea. Serpell (1993) uses the postcolonial experiences of Zambia to explore the social-cultural and psychological meaning of schooling, including language considerations.

One remaining substantive question is why the use of testing for state building has dominated recently--if Sahlberg's putative Global Educational Reform Movement is not as unique in some ways as Sahlberg claims, why *can* he plausibly claim it as a unique phenomenon? Certainly, there is a great quantity of test-based accountability currently, accountability activities organized by central states. The ubiquity of testing as a state activity is relatively new. As such, it is an example of institutional copycat behavior--what DiMaggio and Powell (1983) described as institutional isomorphism. This is a description, not an explanation.

One tentative explanation is that the modern use of testing for state-building has been amenable to broad coalition-building, with policies justified by the uses of schools for human capital development, for mobility access, for modernization of a poorer state, for paternal approaches to childhood and poverty, for social control, and sometimes even to address concerns about inequality. If the advantage of testing for state-building is its utility for coalition building, then the uses of testing for state-building should happen early and more quickly in societies where there is a potential desire or demand for state building -- societies with existing but messy central states. State-building through testing should be less of a clear strategy or need in societies with effective central states and more of a logical strategy for incomplete, aspiring states. Despite anti-state ideology in the United States, that description fits the circumstances surrounding the adoption of No Child Left Behind in the United States in 2001, where policymakers in both national parties had an interest in centralizing control of education and where states and the federal government were willing to "borrow authority" from each other in pursuit of mutual (state-building) ends (Manna, 2006).

References

Barnard, Henry. 1865. History of common schools in Connecticut. *American Journal of Education*, 15, 276-331.

Blitz, J. (2003). *Spellbound*. Documentary.

Delpit, L. D. (2006). *Other people's children: Cultural conflict in the classroom*. Boston: The New Press.

DiMaggio, P. J., & Powell, W. (1983). The iron cage revisited: Institutional isomorphism and collective rationality in organizational fields. *American Sociological Review*, 48, 147-60.

Dreeben, R. (1968). *On what is learned in school*. Reading, MA: Addison-Wesley.

Elman, B. A. (2000). *A cultural history of civil examinations in late imperial China*. Berkeley, CA: University of California Press.

Johnson, J. L. (1969). Special education and the inner city: A challenge for the future or another means for cooling the mark out? *The Journal of Special Education*. Retrieved from <http://psycnet.apa.org/?fa=main.doiLanding&uid=1970-15324-001>

Juster, N. (2011). *The phantom tollbooth*. New York: Random House Digital, Inc.

Kaestle, C. F. (1973). *Joseph Lancaster and the monitorial school movement: A documentary history*. New York: Teachers College Press.

Labaree, D. F. (1988). *The making of an American high school: The credentials market and the Central High School of Philadelphia, 1838–1939*. New Haven, CT: Yale University Press.

Labaree, D. F. (1997). *How to succeed in school without really learning: The credentials race in American education*. New Haven, CT: Yale University Press.

Lemann, N. (2000). *The big test: The secret history of the American meritocracy*. New York: Macmillan.

Lynd, R. S., & Lynd, H. M. (1929). *Middletown: A Study in American Culture*. New York: Harcourt Brace & World.

Manna, P. (2006). *School's in: Federalism and the national education agenda*. Washington, DC: Georgetown University Press.

Metz, M. H. (1989). Real school: A universal drama amid disparate experience. *Politics of Education Association Yearbook*, 4(5), 75–91.

Miyazaki, I. (1976). *China's examination hell: The civil service examinations of imperial China*. New Haven, CT: Yale University Press.

Nye, D. E. (2004). *America as second creation: Technology and narratives of new beginnings* (2nd ed.). Boston: The MIT Press.

Rapple, B. A. (1994). Payment by Results. *Education Policy Analysis Archives*, 2, 1.

Reese, W. J. (2013). *Testing wars in the public schools: A forgotten history*. Harvard University Press.

Sahlberg, P. (2011). *Finnish lessons: What can the world learn from educational change in Finland?* New York: Teachers College Press.

Scott, J. C. (1999). *Seeing like a state: How certain schemes to improve the human condition have failed* (New edition.). New Haven, CT: Yale University Press.