

**A Critical Review of “Getting Tough? The Impact of High School
Graduation Exams”**

Richard P. Phelps

The highly-praised and influential study, “Getting Tough?,” was published in 2001.¹ Briefly, while controlling for a host of student, school, state, and educator background variables, the study regressed 1988 to 1992 student-level achievement score gains onto a dummy variable for the presence (or not) of a high school graduation test at the student’s school. The 1992-1988 difference in scores on the embedded cognitive test in a US Educational Department longitudinal survey comprised the gain scores. The study was praised for its methodology, controlling for multiple baseline variables which previous researchers allegedly had not, and by some opposed to high-stakes standardized testing for its finding of no achievement gains. Indeed, some characterized the work as so far superior in method it justified dismissing all previous work on the topic (see, for example, Phelps 2012b, pp. 232–235). For example:

¹ Jacob, B. A. (2001, Summer). Getting Tough? The Impact of High School Graduation Exams. *Educational Evaluation and Policy Analysis*, 23(2), 99–121.

“In fact, there is little evidence on the effects of exit exams on achievement. One exception is Jacob (2001).” (Reardon, 2008, p. 11)²

“...though one study that examined high school exit exams and that controlled for individual student characteristics (unlike most of the research on this topic) found no such relationship.” (Hamilton, 2003, p. 40)

“While several studies found a positive association between student achievement and minimum competency testing, *a recent study with better controls* for prior student achievement finds no effect (Jacob 2001).” (Jacob, 2001, p.7)

Moreover, the article was timely, its appearance in print coincident with congressional consideration of the No Child Left Behind Act (2001) and its new federal mandate requiring annual testing in seven grades and three subjects in all U.S. public schools. The article also served as the foundation for a string of ensuing studies nominally showing that graduation exams bore few benefits and outsized costs (e.g., in dropout rates). Graduation exam opponents would employ

² Reardon was wrong on this. Just some of the relevant pre-2008 studies of the effects of minimum-competency or exit exams and the problems with a single passing score include those of Alvarez, Moreno, & Patrinos (2007); Grodsky & Kalogrides (2006); Audette (2005); Orlich (2003); StandardsWork (2003); Meisels, et al. (2003); Braun (2003); Rosenshine (2003); Tighe, Wang, & Foley (2002); Carnoy & Loeb (2002); Baumert & Demmrich (2001); Rosenblatt & Offer (2001); Phelps (2001); Toenjes, Dworkin, Lorence, & Hill (2000); Wenglinsky (2000); Massachusetts Finance Office (2000); DeMars (2000); Bishop (1999, 2000, 2001, & 2004); Grissmer & Flanagan(1998); Strauss, Bowes, Marks, & Plesko (1998); Frederiksen (1994); Ritchie & Thorkildsen (1994); Chao-Qun & Hui (1993); Potter & Wall (1992); Jacobson (1992); Rodgers, et al. (1991); Morris (1991); Winfield (1990); Ligon, Johnstone, Brightman, Davis, et al. (1990); Winfield (1987); Koffler (1987); Losack (1987); Marshall (1987); Hembree (1987); Mangino, Battaille, Washington, & Rumbaut (1986); Michigan Department of Education (1984); Ketchie (1984); Serow (1982); Indiana Education Department (1982); Brunton (1982); Paramore, et al. (1980); Ogden (1979); Down(2) (1979); Wellisch (1978); and Findley (1978).

these critical studies as evidence to political effect.³ From a high number of more than thirty states around the turn of the millennium, graduation tests are now administered in only seven or eight states.

“Getting Tough” relied on two data sources: the U.S. Education Department’s Base Year 1988 National Educational Longitudinal Study (three waves: 1988, 1990, & 1992), and a table listing state high school exit examinations assembled by the North Central Regional Educational Laboratory (NCREL) and the Council of Chief State School Officers (CCSSO) (Bond & King, 1995).

NELS-88 is a wonderful data source.⁴ Eighth-grade students agreed to participate for the more-than-a-decade duration of the several-wave project. With each of the first three waves, students took a cognitive test comprised of items borrowed from the National Assessment of Educational Progress (NAEP). In addition, complementary surveys were administered to teachers, parents, and school administrators. “Getting Tough?” makes good use of the wide reach of information provided by NELS-88’s multiple surveys.

One might argue that the author did all that he could in the way of statistical controls with the NELS-88 data base and its embedded cognitive tests, administered in Spring 1988 and Spring 1992. But, the key variable in his study comes not from the NELS, but rather from the NCREL/CCSSO table—the

³ See for example, the Bill & Melinda Gates Foundation funded 2014 report (Hyslop). The Gates Foundation has lobbied for dropping state standards-based exams in favor of the Common Core Initiative exams they favored.

⁴ See <https://nces.ed.gov/surveys/nels88/>

dummy variable for the existence of a graduation test (or not) in the student's jurisdiction.

Moreover, the author altered the NCREL/CCSSO information. First, he removed three test states, presumably because he received contradictory information from another or other sources. Second, he retained as graduation test states three others which NCREL/CCSSO identified as first implementing their graduation exam programs in the years after the author's key 1988–1992 window. Third, the structure of state testing programs simply did not conform well to the assumption of a pseudo-experimental binary condition: graduation exam or not, with all else equal. Fourth, confusing diversity existed even within the appellation "graduation exam."

Perhaps unknown to the author, coincident with the NCREL/CCSSO surveys in the 1988–1992 period were two other more detailed surveys of state testing programs. The first covering the 1990–1991 school year was administered by the General Accounting Office⁵ to all US states and a nationally-representative sample of over 600 public school districts and included a separate questionnaire for each and every systemwide test administered in that year (1993). The second covering the 1991–1992 school year was conducted by Ellen Pechman of Policy Studies Associates for the US Education Department (1992). I employ information from both for comparison with that presented in "Getting Tough?"

The three NCREL/CCSSO test states re-classified in "Getting Tough?" as non-test states? Michigan, Ohio, and Virginia.

⁵ The organization is now called the Government Accountability Office.

NCREL/CCSSO included Virginia’s Literacy Passport Test in its list of “state graduation testing programs.” Information from the GAO and USED surveys indicates that that Virginia test was administered between grades 6 and 8, with its passage required for entry to grade 9. That may explain why “Getting Tough?” dropped Virginia from its state list of high school exit exams.

If the stakes of the Literacy Passport Test motivated students to work harder or study more, and that spillover were registered on the NELS test, its effect may have been in grade 8 rather than grade 12—either just before or contemporaneously with “Getting Tough’s” NELS-88 pre-test. If Virginia’s testing program had any effect on Getting Tough’s gain scores, it might have been to lower them.

The GAO study, however, reveals that Virginia administered another, different test in grade 11 that apparently was “used to determine promotion, retention, or graduation.” It was an adapted hybrid of the Iowa Test of Educational Development (ITED) and Test of Achievement and Proficiency (TAP).

Similarly, “Getting Tough?” may have dropped Ohio from its list because NCREL/CCSSO identified its graduation exam as “Twelfth-Grade Proficiency Testing,” and no test by that name existed, apparently. Neither the GAO nor the USED survey listed such a test. But both included a “Ninth-Grade Proficiency Test” administered from grade 9 to grade 12—a high-stakes test, passage of which was required for graduation.

Conversely, “Getting Tough?” kept three other states—Georgia, New Jersey, and North Carolina—which the NCREL/CCSSO table published in November 1995 listed as not implementing graduation tests until after the author’s 1988–1992 window. For whatever reason, the author changed the “first affected class” dates for tests in those three states from the 1992–1993 school year to 1986 (Georgia), New Jersey (1981), and 1978 (North Carolina).

Judging from the information gathered by the GAO and USED, New Jersey did, indeed, administer a high-stakes examination in the high school years in the 1988–1992 period. But it was not the “Grade 11 High School Proficiency Exam,” which had no student stakes. Rather, New Jersey applied stakes to its “9th Grade Proficiency Exam.”

According to the GAO and the USED, Georgia and North Carolina did not administer graduation exams in the high school years during the 1988–1992 window. They did, however, administer tests in lower grades to which some stakes were attached.

If my memory serves me well, North Carolina had in earlier years administered a high-stakes test in several grades, including the 11th, but dropped it for the 11th in particular by 1990, under legal advice. They had been using the California Achievement Test, a commercially available nationally norm-referenced test. The federal courts’ *Debra P. v Turlington* legal decision a few years earlier had decided that such tests, not directly based on a state’s own curricular standards, violated students’ constitutional rights when used as a graduation requirement.

As for Georgia, the GAO test classification (for 1990–1991) seems at first glance that it might concur with Getting Tough’s but, in this case, the USED survey split the data into more detail. Similar to North Carolina, the state administered a nationally norm-referenced test at four grade levels but, at the high school level, only matrix-sampled.

There remain three more states apparently mis-classified in “Getting Tough?” The November 1995 NCREL/CCSSO table did not identify them as having student-level accountability exams in the high school years. But the GAO study did. In California, school districts picked their own high school exam, but the state required them to have one, with passage required for graduation.⁶ Officials in Indiana and Missouri both claimed continued use of high school exams as graduation requirements into and through 1990–1991. By the next year 1991–1992, according to the USED survey, those stakes had been dropped, even as the testing programs continued.

In sum, “Getting Tough?” apparently:

- mis-classified five states *with* high-school graduation tests—California, Indiana, Missouri, Ohio and Virginia—as *not* having them in the 1988–1992 period; and
- mis-classified three states *without* high school graduation tests—Georgia, New Jersey, and North Carolina—as having them

⁶ The district-choice model was also employed in Tennessee, which was classified a graduation test state in “Getting Tough?,” and Colorado, which was not.

That the three contemporary sources on state testing programs disagreed on some details suggests that it would have been prudent to consult them all. The GAO study was generally the most detailed of the three, and its survey responses from state and local district officials were required by federal law. But the USED study, though gathering only state-level information, was remarkably detailed and thorough; comprising an excellent data source as well.

Arguably the best feature of the NCREL/CCSSO survey, upon which the author relied, was its annual repetition, which could accumulate trends. Embedded in that same feature, however, was a drawback. In order to ease the burden of the annual, voluntary survey response request, state assessment directors were presented their data from the previous year and asked only to indicate changes. Thus, a non-response would not register as missing data but, rather, the previous year's data, which would be reliable only if the same program had continued with the exact same characteristics.

Tables 1–3 below summarize information from the GAO survey regarding test content, grade levels, and purposes. All statewide tests included in the survey collection for the school year 1990–1991 are included.

Table 1 includes the states that were counted as having “graduation exams” according to “Getting Tough?” Other states are included in Table 2. A glossary can be found below each table to identify acronyms.

Excerpted in Table 3 is the section of the GAO questionnaire of state testing officials relating to test purposes at the student and school levels.

Table 1. Exams classified as minimum competency exit exams in “Getting Tough?”

State	Name in “Getting Tough”	Grades administered, Subject areas (GAO)	Purposes (GAO) – Student, School levels only	Other tests administered (grade levels) (student, school level purposes) (GAO)
AL	High School Basic Skills Exit Exam	3, 6, 9, 11, 12	1. Student Accountability	Stanford (4, 8) (3)
FL	High School Competency Test	10–PG Reading, Writing, Math	1. Student Accountability	Subject Area Exams (10, 11) (n/a) district choice NRT (4, 7) (n/a)
GA-1	Basic Skills Test-1	3, 6, 8 Reading, Writing, Math	1. Student Accountability 3. Special Prgm Screening	TAP (9) (3) Iowa (2, 4, 7) (3)
GA-2	Basic Skills Test-2	10 Reading, Writing, Math	n/a	
HI	Test of Essential Competencies	10, 11, 12 Full Battery + Civics	1. Student Accountability	Stanford (3, 6, 8, 10) (3, 5)
LA	Graduation Exit Exam	10, 11	1. Student Accountability	CAT (4, 6, 8) (1) “LEAP” (3, 5, 7) (1, 5)

MD	Functional Testing Program	9–12 Reading, Writing, Math, Civics	1. Student Accountability 5. School Accountability	School Performance Assessment (3, 5) (5)
MS	Functional Literacy Exam	11 Reading, Writing, Math	1. Student Accountability	Basic Skills Assessment (3, 5, 8) (2, 3) Subject Area Tests (8, 9) (2) Stanford (4, 6, 8) (2, 3) CAT + state component (3, 6, 8) (1, 2, 3) Early Warning Test (8) (3) Grade 9 Proficiency Test (9) (1, 5) Direct Writing Assessment (4, 6) (1, 2, 3, 5) CTBS (3, 5, 8) (1, 2, 3, 5)
NC	HS Graduation Test	n/a	n/a	
NJ	Grade 11 HS Proficiency	11 Reading, Writing, Math	n/a	
NM	HS Competency Exam	10, 11, 12 Full Battery	1. Student Accountability 2. Grouping or Placement 3. Special Prgm Screening 5. School Accountability	
NV	HS Proficiency Program	11, PG	1. Student Accountability 2. Grouping or Placement	CTBS (3, 6, 9) (2) Writing Test (11) (2)

NY	Regents Competency Tests	9–12	1. Student Accountability 5. School Accountability	Pupil Evaluation Program (3, 5, 6) (1) Pupil Evaluation Tests (4, 6, 8) (n/a) Regents’ Exam (8–12) (1, 5) Preliminary Competency Tests (8, 9) (5) Stanford (4, 5, 7, 9, 11) (1, 3, 5)
SC	Basic Skills Assessment Program	1, 2, 3, 6, 8, 12–PG Readiness, Reading, Writing, Math, Science	1. Student Accountability 3. Special Prgm Screening 5. School Accountability	
TN	Proficiency Tests (1)	9–12 Math, Aptitude	1. Student Accountability	Iowa (2–8) (1)
TX	Assessment of Academic Skills	3, 5, 7, 9, 11, 12 Reading, ELA, Math	1. Student Accountability 5. School Accountability	

* According to the GAO data collection, the TN Proficiency Tests were district managed

Glossary

CAT: California Achievement Test

CTBS: Comprehensive Test of Basic Skills

ELA: English Language Arts

Full Battery: core subjects, typically: Reading, ELA, Math, Science, Social Science or History

Iowa: Iowa Test of Basic Skills or Iowa Test of Educational Development

NRT: norm-referenced test

PG: post-graduate

Stanford: Stanford Achievement Test

TAP: Test of Achievement and Proficiency

Table 2. States with testing programs identified as NOT having a minimum competency exit exam in “Getting Tough”

State	Had Test with Stakes? (GAO)	Grades (Subjects) (GAO)	Purposes (GAO) – Student, School levels	Other test(s) administered (grade levels)
AK				Iowa (4, 6, 8)
AR-1	YES, “Minimum Performance Test”	3, 6, 8 (Full Battery, Civics)	1. Student Accountability 5. School Accountability	
AR-2	YES, MAT	4, 7, 10 (Full Battery)	2. Grouping or Placement 3. Special Prgm Screening 5. School Accountability	
AZ				Iowa, TAP (2–11)
CA	YES, district chose test used	10, 11 Reading, ELA, Math	1. Student Accountability	district choice (4–6) district choice (7–9)
CO	YES, district chose test used	district choice	5. School Accountability	International Assessment of Academic Progress (4, 8)

CT	YES, "CT Mastery Test"	Reading, Math 4, 6, 8 Reading, ELA, Math	3. Special Prgm Screening	district choice
DE				Stanford (1, 4, 6, 9)
ID				Direct Writing Assessment (8, 11)
IL				Iowa, TAP (6, 8, 11) "IGAP" (3, 6, 8, 11)
IN	YES, "ISTEP"	1-3, 6, 8, 9, 11 Full Battery	1. Student Accountability 3. Special Prgm Screening 5. School Accountability	
KS				Math Pilot Assessment (3, 7, 10)
MA				"MEAP" (4, 8, 12) Basic Skills Test (3, 6, 9)
ME				"Maine Educational Assessment" (4, 8, 11)

MI				“Essential Skills Test” (4, 7, 10)
MN				“ELO Assessment” (4, 8, 11) Science Test (6, 9, 11)
MO	YES, Missouri Mastery Test	2–10 Full Battery + Civics	1. Student Accountability 2. Grouping or Placement 3. Special Prgm Screening	
MT				district choice among 8 NRTs
ND				CTBS & TCS (3, 6, 8, 11)
NH				CAT (4, 8, 10)
OH	YES, 9 th Grade Proficiency Tests	9–12 (Reading, Writing, Math, Civics)	1. Student Accountability	
OK-1	YES, MAT Writing	7, 10 (Writing)	2. Grouping or Placement 3. Special Prgm Screening	

OK-2	YES, Iowa + TAP	3, 5, 7, 9, 11 (Full Battery)	3. Special Prgm Screening	
OR				“Oregon Assessment” (3, 5, 8, 11)
PA	YES “TELLS”	3, 5, 8 (Reading, Math)	5. School Accountability	Writing Test (6, 9)
RI	YES, MAT	3, 6, 8, 10 (Reading, ELA, Math)	2. Grouping or Placement 3. Special Prgm Screening	Writing Test (3, 6)
SD	YES, Stanford + OLSAT	4, 8, 11 (Full Battery)	3. Special Prgm Screening 5. School Accountability	Ohio Vocational Interest Survey (n/a)
UT	YES, Stanford	5, 8, 11 (Full Battery)	2. Grouping or Placement 3. Special Prgm Screening 5. School Accountability	
VA-1	YES, “Literacy Passport”	6 (Reading, Writing, Math)	1. Student Accountability 3. Special Prgm Screening 5.	

			School Accountability	
VA-2	YES, Iowa + TAP	1, 4, 6, 7, 8, 11 (Full Battery)	1. Student Accountability	
WI				3 rd Grade Reading Test (3)
WV-1	YES, “STEP”	1–4 (Reading, Math, Science)	1. Student Accountability 5. School Accountability	Writing Assessment (8, 10)
WV-2	YES, CTBS	3, 6, 9, 11 (Full Battery)	3. Special Prgm Screening	

Glossary

CAT: California Achievement Test

CTBS: Comprehensive Test of Basic Skills

ELA: English Language Arts

Full Battery: core subjects, typically: Reading, ELA, Math, Science, Social Science or History

Iowa: Iowa Test of Basic Skills or Iowa Test of Educational Development

MAT: Metropolitan Achievement Test

NRT: norm-referenced test

OLSAT: Otis-Lennon School Ability Test

Stanford: Stanford Achievement Test

TAP: Test of Achievement and Proficiency

Table 3. Exam “purposes” in GAO state questionnaire [only student and school levels shown]

To what purpose was the test used? (*Check all that apply.*)

1. Student-level accountability. Assessment used to determine promotion, retention, or graduation.
2. Student grouping or placement. Assessment used to assign students to academic groups within their class.
3. Student screening for special programs. Assessment used to determine eligibility for special programs.
4. Individual student evaluation.
5. School-building-level accountability. Results are used to determine principal's retention, promotion, or bonus, or cash awards to, honors for, status of, or budget of the school.
6. School-building-level evaluation.
7. School-building-level curriculum appraisal and improvement.

Stakes

The title of the article, “Getting Tough?” implies that the exams the author identified as graduation exams *should be* “tough,” which one presumes means high-stakes: one either passes the exam or one does not graduate.

Yet, “Getting Tough?” included no control for how many chances students got to pass, which could range from a few to infinity. In Hawaii students were administered the state’s “Test of Essential Competencies” “until students pass” (Pechman, Table B-1, p. 4). Apparently, Hawaiians wished to administer a graduation exam, but never had any intention of “getting tough” with it. In Florida, Nevada, and South Carolina students could keep trying to pass the graduation test even after they had completed their coursework and left school.

“Getting Tough” calculates gain scores for the period 1988–1992. The students in the study took the “pre-test” in Spring 1988 and the “post-test” in Spring 1992. “Getting Tough” assumes that the effect of a “tough” “graduation test” should reveal itself within that window of time.

But, a “graduation” exam covering 10th to 12th-grade subject matter, and administered just once a year under highly secure conditions in grade 12 cannot be considered equivalent to an exam based on 6th and 7th grade curricula, offered multiple times between grades 8 and 11, with no test security protocols, and a non-test alternative path to a diploma waiting in grade 12 for those who haven’t yet passed.

Then, there is the probable confounding effect of “medium stakes,” represented in the GAO survey by the test purposes “student grouping or

placement,” “student screening for special programs,” or “school-building-level evaluation.” With medium stakes, consequences may apply conditionally, to some students and not others, or for some programs and not others, and consequences may fall far short of denying a student a diploma or grade-level advancement. Still, there are consequences and, thus, behavioral incentives.

Further degrading the validity of the presumed isolation of a graduation test effect, twenty states claimed a combination of purposes involving both high and medium stakes for one or more tests.

The Timing of Stakes

According to the GAO and USED studies, some states attached stakes to tests administered in grade 9. Others began the first of several administrations of their high school exit exam in grade 9, giving students a chance to pass it already then, and then again in grades 10, 11, 12, or, in some states, even after grade 12. When most of the motivational effect of stakes was spent already in grade 9, or only kicked in after grade 12, one would expect it to be weak in the Spring semester of grade 12.

More significantly, some states administered other high stakes tests in grade 8, representing a motivational incentive at the same time, or just before, the administration of the author’s pre-test, during the Spring 1988 NELS survey administration. Apparently, eighth grade has long been a favorite grade level for states to administer important tests (Yeh, p. 13). In all, ten states administered high-stakes tests in grade 8. In these states, an eighth-grade motivational boost

would serve to elevate pre-test scores and thus dampen any grade 8 to grade 12 achievement gains. Six of these ten states were classified by the author as states with high school exit exams.

The senior slump and test-taking motivation

Post-testing for “Getting Tough?” took place in Spring 1992 with the administration of the NELS embedded test. The moment when any achievement gain would be registered was the last semester of the students’ senior year. The NELS test had no stakes for them, it was completely up to them whether or not they exerted any effort at all, just enough to complete an obligation, the requisite amount to show their best, or something in between.

One might argue that at least the 8th-grade NELS and 12th-grade NELS test are equivalent in their lack of direct personal consequences for the student test takers. But, studies of test-taking effort on no-stakes tests show effort declining as students age, with the lowest effort found among high school seniors. Studies of the “senior slump” among US students have pinpointed the last high school semester, when the “Getting Tough?” post-test was administered, as the nadir in cognitive effort.⁷

Stakes for whom?

⁷ See, for example, Wainer, H. (1993, Spring). Measurement Problems. *Journal of Educational Measurement* 30(1), pp. 12–13; or The National Commission on the High School Senior Year. (2001). *The Lost Opportunity of Senior Year*. Washington, DC: Author. <https://files.eric.ed.gov/fulltext/ED453604.pdf> or Venezia, A., Kirst, M. W., & Antonio, A. L. (2004). *Betraying the College Dream*. Palo Alto, CA: Stanford University Bridge Project.

In addition to identifying tests whose purposes included accountability for students, the GAO survey identified tests designated for *school* accountability, as in “Results are used to determine principal's retention, promotion, or bonus, or cash awards to, honors for, status of, or budget of the school.” One might surmise that these high or medium stakes for schools might incentive their leaders to encourage student achievement gains.

In nine states, school accountability paralleled student accountability, in five others—none of which were included in the “Getting Tough?” list of alleged graduation exams— it did not. Some tests charged with a school accountability purpose were included among the “Getting Tough?” graduation tests (N=5), but others were not (N=9). Ten states administered school accountability tests in grade 8,⁸ when their influence, if any, on grade-8-to-grade-12 gain scores would have been to dampen them.

Better controls

“Getting Tough?” was lauded for its alleged better-than-earlier-studies control for prior student achievement (or aptitude; the author employed the two terms interchangeably).

Like many multivariate regression studies, this one was meant to replicate an experiment, with the primary comparison made between 1988-to-1992 gain scores for students in schools with high school exit exams and those without. To the extent possible, all other mitigating factors are controlled.

⁸ AR, HI, LA, MO, NM, NY, PA, SC, SD, and UT

Yet, “Getting Tough?” rather conspicuously lacked controls for other key factors, such as the stakes of the “graduation exam” and other exams, the multiple purposes intended for each exam, the subject-areas or difficulty levels of the content, or any aspect of test administration, including the level of test security. Most of this information was available in the GAO and USED studies.

Multivariate analysis versus experiments

One might argue that multivariate regression clearly offers an authenticity advantage over experiments, particularly when the subject of study involves a large population. In general, as population size increases, the feasibility of a randomized controlled experiment (RCE) decreases. Moreover, an RCE is not even possible under the genuine circumstances of a high school exit examination that is legally required for all high school students within a jurisdiction. (Imagine a state legislature announcing that a randomly chosen half of its high-school seniors must pass a high-stakes exit examination in order to graduate, while their counterparts in the other half will face no such requirement.)

That said, a multivariate analysis comparing two conditions is only valid if those conditions can be meaningfully separated in the data from other confounding effects. Did “Getting Tough?” manage that? No.

Meanwhile, entirely unconsidered by policymakers in 2001–2003 was a century’s worth of focused experimental and qualitative research on each and every one of the aspects of graduation tests known to influence student

achievement gains (Phelps, 2012a). Also ignored: a large research literature on testing programs similar to high school graduation exams, such as those for occupational licensure.

Discussion

The author of “Getting Tough?” wrote: “These results suggest that policymakers would be well advised to rethink current graduation test policies.”

“Current graduation test polices”—apparently meaning current at the time “Getting Tough?” was published in 2001—differed substantially from their counterparts of 1988–1992, the time period covered in the study, however. 1988–1992 was an “in-between” period when states were still adjusting to the ramifications of the *Debra P. v Turlington* decision by dropping stakes attached to any norm-referenced tests administered at the high-school level that had not been hybridized to cover their state’s standards. Many states were just starting the process of writing matched content and performance standards or developing new *Debra P.*-compliant standards-based tests.

By 2001, this transition had been completed in most states, and the new “graduation test policies” regime bore little in common with that of a dozen years earlier.

The multivariate analysis in “Getting Tough?” should have had the advantage of authenticity—an analysis of a phenomenon studied in its actual context. But that should mean that the context is understood and specified in the analysis, not ignored as if it couldn’t matter.

And, it could have been understood and specified. Most of the relevant information left out of “Getting Tough?”—specific values for other factors that tend to affect test performance or student achievement—was available from the three contemporary surveys, and the rest could have been obtained from a more detailed evidence-gathering effort.

The study could have been more insightful had it been done differently, perhaps with less emphasis on “more sophisticated” and “more rigorous” mathematical analysis, and more emphasis on understanding and specifying the context—how testing programs are organized, how tests are administered, the effective differences among the wide variety and forms of tests and how students respond differently to each, the legal context of testing in the late 1980s and early 1990s, and so on. The study could have incorporated the following:

- it is unreasonable to expect all tests that happen to be inconsistently labelled as “graduation tests” to have the same effect regardless of stakes,⁹ level of security,¹⁰ student effort,¹¹ and a variety of other factors correlated with student achievement gains;
- these other factors were essential to control; and

⁹ See, for example, Phelps, R. P. (2019). "Test Frequency, Stakes, and Feedback in Student Achievement: A Meta-Analysis". *Evaluation Review*. 43(3–4): 111–151. [doi:10.1177/0193841X19865628](https://doi.org/10.1177/0193841X19865628).

¹⁰ See, for example, Steger, D., Schroeders, U., & Gnambs, T. (2018). "[A Meta-Analysis of Test Scores in Proctored and Unproctored Ability Assessments](https://doi.org/10.1027/1015-5759/a000494)". *European Journal of Psychological Assessment*: 1–11. [doi:10.1027/1015-5759/a000494](https://doi.org/10.1027/1015-5759/a000494).

¹¹ See, for example, Finn B. (2015). "Measuring motivation in low-stakes assessments". Educational Testing Service. *Research Report RR-15-19*. <https://onlinelibrary.wiley.com/doi/full/10.1002/ets2.12067>

- those controls could have been included, as values for those variables were available from information sources that were known to experts in educational testing.

References

- Bond, L. A., & King, D. (1995). *State student assessment programs database*. Oakbrook, IL: Council of Chief State School Officers (CCSSO) and North Central Regional Educational Laboratory (NCREL).
- Debra P. v. Turlington*, 644 F.2d 397, 6775 (5th Cir. 1981).
- Hamilton, L. (2003, January 1). “Assessment as a Policy Tool,” Chapter 2 in *Review of Research in Education* 27(1), 25–68.
- Hyslop, A. (2014). *The Case Against Exit Exams*. Washington: New America Foundation.
- Jacob, B. A. (2001, Summer). Getting Tough? The Impact of High School Graduation Exams. *Educational Evaluation and Policy Analysis*, 23(2), 99–121. <https://www.jstor.org/stable/3594125>
- No Child Left Behind Act of 2001, P.L. 107-110, 20 U.S.C. § 6319 (2002).
- Pechman, E. M. (1992, July). *Use of Standardized and Alternative Tests in the States*. Prepared for the U.S. Department of Education, Office of Policy and Planning. Washington: Policy Studies Associates.
- Phelps, R. P. (2012a). The effect of testing on student achievement, 1910–2010. *International Journal of Testing*, 12(1), 21–43. <http://www.tandfonline.com/doi/abs/10.1080/15305058.2011.602920>
- Phelps, R. P. (2012b, Summer). Dismissive Reviews: Academe’s Memory Hole. *Academic Questions*, 25(2), New York: National Association of Scholars.
- Reardon, S. F., Arshan, N., Ateberry, A., & Kurlaender, M. (2008, September).

“High Stakes, No Effects: Effects of Failing the California High School Exit Exam,” Paper prepared for the International Sociological Association Forum of Sociology, Barcelona, Spain.

Yeh, J. P. (1978, June). *Test Use in Schools*. Los Angeles: UCLA, Center for the Study of Evaluation.

U.S. General Accounting Office. (1993). *Student Testing: Current Extent and Expenditures, With Cost Estimates for a National Examination*. PEMD 93-8, Washington, D.C.: Author.