PHILOSOPHY OF EDUCATION *2011* | Robert Kunzman, editor

Standardized Quantitative Learning Assessments and High Stakes Testing: Throwing Learning Down the Assessment Drain

Matthew J. Hayden
Teachers College, Columbia University

The worldwide standardization of education is well underway and includes institutional structures, teacher training, curriculum development, and more importantly, specifically delineated outcomes of schooling and the assessments that measure them. I contend that the core of the standardized testing movement has a narrow, limited, and inaccurate working definition of learning and has accepted wholesale the assumptions inherent in statistical quantitative learning assessments (SQLA) outcome measurement, which make them unsuitable for use in high stakes testing (HST). I use Gert Biesta's work on evidence-based learning to frame my analysis of SQLA by thinking with Hans-Georg Gadamer about prejudice, foreknowledge, and interpretation to better understand these problems. I then examine how the methods used to analyze SQLA actually end up determining what "learning" is, and how using them in the context of HST may actually decrease learning even if scores increase.

"Quality" and "Accountability"

There has been much recent handwringing over the alleged "failure" of United States' schools, partly based on the average standardized test scores of U.S. students as compared to the average scores of students from around the world on the same or similar tests. One result has been increased, and abstract, demands for "quality" and "accountability" in schooling, that find some specificity in scientific and technological concepts from industry and a market-based paradigm wherein quality is determined by the satisfaction of the "customers" or "consumers."[1] Such a conception begs the question (*Who are the customers?* Students, parents, or the general public?) and easily leads to education becoming merely the delivery of goods (that is, grades and test scores) and services wherein the teacher is a "server" and operates at the customer's behest. The arguments about accountability are similarly vague and seem to imply teachers, administrators, and schools, but what about students, parents, and society? The market-based paradigm supported by the language of SQLA is not going to hold the "customer" accountable; the customer is always right. Thus the "service" is held accountable, focusing on instrumental accountability that has as its object instrumental education, not "rich" knowledge or competencies.[2] "Accountability" also finds comfort in its linguistic brethren *accounting* and the quantitative and numerical values of accountancy, betraying a cultish faith in quantitative analysis to support vague determinations of quality. This market-based paradigm and the "evidence" it produces ignores the historical and intellectual prejudices embedded in the practices of statistical analyses that constrict and restrict the parameters of inquiry, substantially predetermine outcomes, and limit what can be known from the results.[3]

The Baseless Claims of Evidence-Based Research

A fundamental problem with SQLA finds an analogous problem in evidence-based research in education, especially given the turn in education toward quantitative, data-driven methods of assessing student learning and teacher quality. Gert Biesta's examination of evidence-based research in education found that the indeterminate nature of "evidence" and the tendency of research to focus on bringing about predetermined ends are problematic.[4] In the former case, evidence can only illustrate what has been done and what the result was, but can have no direct bearing on the outcome of future efforts. In the latter case, one can at most rely on probabilities in attempting to bring about desired ends (such as, student proficiency in *x*). Thus, forming models for present and future efforts based on what *has worked* cannot account for the dynamism of one's environment and is unlikely to achieve similar results. This is particularly the case if one systematizes a model that then remains fixed — as is the case for SQLAs. It is the nature of standardization to resist flexibility. Such fixed models will fail to assess what they were designed to assess once the conditions that created the model change, and those conditions change constantly.[5] Since SQLA indicate a student's performance relative to a norm group, his or her relative position will not change quickly, making year-to-year HST ineffectual.[6] Further, SQLA (as evidence-gathering tools) are designed far away from and without the input of the actual lessons and learning conditions that have informed the learning that *has* taken place. So instead of measuring what students know — something that could be done with a localized assessment instrument — SQLA can only tell us what the students might know in relation to what the assessment decided, independently, that it would assess. We have no idea *why* or *how* the student might or might not have been able to select the "correct" response to each question, only that the student did or did not select the "correct" response. But in HST, the "why" is precisely what the SQLA will be used to determine (for example, *Why low scores?*). These results do not tell us that the student actually knows the answer or how she or he came to know it, or does not know the answer and why not. The answers to these questions are invaluable educational learning assessment tools. SQLA results in HST do not allow for such evidence to inform teaching practices because the evidence is used summatively rather than formatively. So, instead of SQLA being a tool for learning assessment, they actually stand in opposition to it.[7]

As evidence-based research into student learning that should inform teaching practices, SQLAs are premised on the idea that teachers do something — "they administer a treatment" — and then something happens as a result (that is, the students learn).[8] However, the "evidence" from a test says nothing about what is *supposed* to happen, only *what has* happened. Biesta notes that these assumptions may work in the field of medicine, but we should be skeptical in education. The model of evidence-based research misses the fact that education is not a single cause-and-effect process, but is instead "a process of *symbolic* or *symbolically mediated* interaction."[9] Therefore, what has gone on before the evidence gathering (the assessment) is not subject to simple, discrete, quantifiable, and linear only

10.47925/2011.177

cause-and-effect processes. The nontechnocratic fact of learning is that it is the result of interpretation on the part of students as they "make sense of what they are taught."[10] It is only through processes of mutual interpretation that education is possible, and through interpretation we can see that Gadamer's notion of prejudice begins to render dubious the efficacy of SQLA.

Gadamer and Prejudice

According to Gadamer, prejudice plays a vital role in our ability to learn by forming prejudgments. He leans heavily on Martin Heidegger, who uses the term "fore-meaning" to refer to the prior knowledge that forms our prejudices.

> every revision of the fore-projection is capable of projecting before itself a new projection of meaning; rival projects can emerge side by side until it becomes clearer what the unity of meaning is; interpretation begins with fore-conceptions that are replaced by more suitable ones. This constant process of new projection constitutes the movement of understanding and interpretation.[11]

Since we understand things in light of the things we already know, or in light of the range of knowable options presented to us by our own prior-understandings, our understanding is a "fore-conception of completeness" that cannot be understood independent of this prior-knowledge.[12] The knowledge that precedes the newly acquired information informs our understanding of it. This foreconception is a prejudice that is carried by the individual into the new learning moment.

> It [foreconception of completeness] states that only what really constitutes a unity of meaning is intelligible. So when we read a text we always assume its completeness, and only when this assumption proves mistaken — i.e., the text is not intelligible — do we begin to suspect the text and try to discover how it can be remedied.[13]

It is not only in discovering how the text can be remedied, however, that one learns. We often remedy the mistakes in our foreknowledge. *Foreconception of completeness* contains the assumption that whatever it is that is to be understood is something that is an intelligible and unified whole. Prejudices, however, may reveal gaps or sections of unintelligibility when we find that something is not a unified whole. We then seek to remedy these inconsistencies. However, in SQLA there is no opportunity to remedy the unintelligibility encountered, and this is reflected in deficient scores. SQLA do not tell us *what* is unintelligible, only that there are low scores.

Gadamer's prejudice is *foreknowledge*, not *foreopinion*. It is not based on a conception of what one desires to be, but rather what one presently knows to be. SQLA, however, are based on what knowledge the SQLA designer desires the students to know, and not what the designer *knows* the students know — impossible because the designer has no way of knowing what the students have been doing; she or he can only assume or predict. This makes SQLA even more dubious as a method of assessment of what students have learned. Biesta shows that for evidence-based data gathering to be remotely effective at telling us something about learning, it would have to be constitutive of a process of mutual interpretation, involving the various foreconceptions and prejudices present. However, SQLA are designed without any understanding of the prior-knowledge of those being assessed and are designed well outside of the interpretive processes of educational practices.[14] The

10.47925/2011.177

evaluation of SQLA are infused with prejudices that privilege the type of information SQLA produce as well as the type of statistical evaluation that is possible, all of which assumes that it is the type of information on which static and summative judgments of learning can be accurately made.

Biesta also notes that evidence-based research can only tell us what worked, not what will work. SQLA can only tell us what the students may not know, and not what they do know. The kind of information gathered by SQLA is premised on an intention, but has no way of showing what actually occurred in the learning of the student. The classroom teacher, the one with whom the student has engaged in mutual interpretation for learning based on foreconceptions and foreknowledge, does not design the HST SQLA. Therefore, the typical SQLA instrument can have no reliable means of testing *what has been learned*, though it may test a type of knowledge or be one way to determine what the student does not know (in relation to what is included in the assessment).[15] The "score" a student gets on a SQLA can only indicate, in numerical form, what the student scored on the test, not what the student learned in the course.[16] The score is merely correlative to our inference and ought not be used as sufficiently causal for the inference nor the terminal effect of a presumed cause (such as the teacher, lessons, or curriculum). Further numbers crunching does not change this fact; statistical significance does not automatically mean educational significance.[17] The inherent prejudice in this system is that there is no relevance to whether or not the questions have a direct connection to the mutual interpretation that has gone into the learning. The SQLA does not "care" nor inquire — it only collects and scores.

The Pre-determination of Statistical Analysis

The starting point of SQLA is one in which the nature of what constitutes learning is defined by the constraints of the measurement used. SQLA are analyzed by quantitative, numerical, and statistical analysis tools. Even when the HST consists of qualitative evaluation (such as an essay graded by a human), those results are converted into a quantitative symbol to be further morphed into a statistic that will then be used to offer an explanation of some kind. Quantitative analysis requires specific analysis tools. Chi-square, p-values, regressions, ANOVA, t-tests, among others, are used for specific types of data, are efficacious only in those contexts, and are often selected prior to the design of the assessment instrument — one cannot use an analysis tool for data that cannot be made quantitatively intelligible by it, so the design of the collection instrument is dependent on the analysis to be used. This means that when the analysis tool is chosen prior to the assessment design, the nature of the results is predetermined, and once the analysis tool is selected, we already know what kind of evidence of knowledge will be found: the type that can be inferred from that type of data and analyzed by that tool.[18] Therefore, any information produced by the test that is not intelligible to the statistical analysis tool used is ignored and not included in the assessment results. This means that *what will be assessed as "learning"* is determined by the analysis tool. No deeply philosophical understanding of what learning is informs the assessment; instead the statistical analysis tool makes this determination.

10.47925/2011.177

Furthermore, the results produced are taken to be or represent products of knowledge and thus serve to define and reify what learning is. However, as Biesta shows, the results are only an indication of what this specific assessment and students produced, not what will be produced or what students will "know" in the future. Further, as comparative tests of relative achievement, these standardized tests require a wide score-spread to be reliable and valid. However, test items that provide this are more often answered correctly by students with higher socio-economic status and higher innate academic potential, thus highlighting skills the school had nothing to do with and rendering a measure of school quality unreliable.[19] The acknowledgement of these problems calls for educational processes and assessments that are constitutive of the mutual interpretation of the subject and learning that take place. In the face of this, all standardized educational practices, let alone SQLA, seem less efficacious.[20]

SQLA do not accurately measure learning because their use precludes an accurate understanding of what learning is.[21] They fail to take into consideration the prejudicial and interpretive nature of learning, as well as the giving and taking of an assessment. When taking a SQLA, each student must first determine what the assessment says or asks, and then what it *means* (and thus an extremely important interpretive task that is not taken into account in the design of the SQLA).[22] *What the question means* is itself a pregnant situation, but the educative information to be gleaned from its examination is discarded by the analysis of SQLA prior to the exam's administration. There is very little control the school or teacher can have in guaranteeing the transmission and replication of identical interpretations given the complexity of individual experiences and historical embeddedness. The only reliable consistency is the nature of the results of the assessment (numerical), but not what they indicate or mean. Learning has been defined in SQLA by first choosing the measurement tools preferred or the kind of results or data that count as knowledge, rather than by defining what learning is and then choosing the best methods by which to assess it. Thus, learning becomes a narrow and rigid concept that is of very little value in the broad, comprehensive manner in which SQLA have been implemented.

Inflexible, standardized models carry with them embedded prejudices about a specific type of knowledge that was, at one time, expected to be present, as well as an embedded prejudice about what the model itself is capable of assessing. In the case of SQLA the fore-conception is that the test is intelligible to the students, is intelligible regarding the subject area, and will produce intelligible results. The results are taken as such, understood to be legitimate, and thus legitimize the assessment and its foreconceptions, creating a tidy, closed system that knows the nature of what it is going to find before it looks. It is tempting to suggest that the breakdown in efficacy is then in the interpretation of the results, but the breakdown occurred the moment SQLA was deemed sufficient for determining what learning is.

The Paradox of High Stakes Testing

Compounding the problem of the inefficacy of SQLA is their use as tools of evaluations of schools and teachers, not only students. Most U.S. states and districts

10.47925/2011.177

that utilize SQLA do so as part of high stakes testing (HST).[23] These are supposed to be achievement tests in which the futures of schools, administrators, teachers, and students are determined.[24] High scores mean schools stay open, administrators and teachers keep their jobs, and students move on to the next grade or are placed in "advanced" curriculum tracks. Low scores may often mean the exact opposite in each case. How do we know that a student has learned based on the score of a test? A test may meet the standards of reliability and validity (insofar as we can statistically measure that a test tests what it tests), but that does not necessarily mean that we know those scores reliably and validly represent learning.

Using the self-determination theory (SDT) of motivation, Richard Ryan and Netta Weinstein examined the effects of high stakes testing on the motivation of students.[25] SDT is a theory that focuses on intrinsic and extrinsic motivation. A key component of SDT is the social context of motivation of which there are three main areas: *autonomy, competence,* and *relatedness*. For intrinsic self-motivation, one must exist in a context in which one feels autonomous, competent, and positively connected to one's environment (classmates, co-workers, and community). Ryan and Weinstein found clear effects on student motivation dependent on the characteristics of the assessments and their purposes. Assessments used for informational purposes, such as formative assessments used by teachers to improve their teaching, have a positive impact on student motivation. Assessments that are "controlling," those that put pressure on students to achieve particular outcomes by offering rewards or punishments (that is, grades or passing to the next grade), have a negative influence on student motivation and lead to the exertion of the least amount of effort required to achieve the outcomes desired with an attendant *decrease* in performance (which holds regardless of how well students do on the tests). Additionally, assessments that induce students to feel helpless or incompetent (such as tests that are too challenging or provide negative feedback) also discourage effort and result in decreased motivation.

HST *can* tell us how students perform on these tests under a certain kind of pressure, but that is not what they are intended to discover. Since most of these tests require a specific reading skill level, even a mathematics assessment will be a reading literacy assessment, but we cannot be certain of its mathematics measurement if a student's reading literacy prevents her from accurately reading the question. What makes HST more dubious is that the reward is merely the opportunity to take more high stakes tests the next year, whereas the sanction is to be excluded from most educational tracks and opportunities. The penalties include social stigma and the limiting of one's educational, and therefore economic, opportunities; penalties that seem more destructive than the rewards are constructive.[26] Further, those who are rewarded carry with them decreased motivation and an instrumental approach to their education, thus increasing the potential for diminished "returns" across the board for all students.

HST has the same demotivating factors for teachers, especially if used to assess teacher effectiveness. Because the tests are high stakes and teacher jobs or the life of the school may be in jeopardy, teachers are compelled to organize their instruction

10.47925/2011.177

around and for them, depriving them of autonomy in their classroom and subject area, autonomy that is necessary to ensure the competence that the HST demands, and consequently depriving them of the context of autonomy required for intrinsically motivated action. Second, because HST are designed by someone not only external to the teacher but almost always external to the school, there is an implied perception of low teaching competence, which deprives the teacher of another requirement for intrinsic motivation; affirmation and exercise of competence. Third, since outcomes are measured rather than the behaviors that lead to them, teachers are disconnected from the process of developing the whole student, alienating them from the student, which leads to the perception that the student and his or her performance on the HST is an instrumental means to some nonstudent-centered end (teacher keeps job, school stays open). This is not a teaching environment we should endeavor to create.

Conclusion

We should be concerned about what happens when education is defined by forces that seek market-based analogies and statistical ease over cognitive, intellectual, and emotional growth, and that once the discourse is framed in favor of the former, it becomes increasingly difficult to illustrate and explain that SQLA are not solely capable of determining whether or not, and to what degree, learning has taken place. This discourse needs to be reframed and reclaimed by educators and educational theorists who need to adequately define learning in a way that can, first, be understood as meaningful by the public, and second, be assessed in a way that suits the goals of the learning as defined. In other words, what learning *is* and consists of should determine how it is assessed rather than using the assessment to define learning.

---

1. W.A. Hart, "The Qualitymongers," *Journal of Philosophy of Education* 31, no. 2 (1997): 295–308.

2. See Andrew Davis, "Is there a Future for Assessment and Accountability?" *Journal of Philosophy of Education* 32, no. 1 (1998): 145–52, for a discussion about how instrumental education to provide workers for industry is flawed, and even so, testing of competencies is not a good way to achieve that goal.

3. Few customers would order such "goods and services" as disappointment, challenge, and pain even though classical and contemporary research identify these as necessary components in learning. Mintz shows that pain is an essential component to student learning and gains made provide a foundation for highly motivated future learning efforts. See Avi Mintz, "The Labor of Learning: A Study of the Role of Pain in Education" (PhD diss., Columbia University, 2008). Jonas states we must alleviate the suffering "only if assisting the individual will have the net effect of making her more autonomous" and result in a positive influence on learning. See Mark E. Jonas, "When Teachers Must Let Education Hurt: Rousseau and Nietzsche on Compassion and the Educational Value of Suffering," *Journal of Philosophy of Education* 44, no. 1 (2010): 56. See also Jean-Jacques Rousseau, *Emile, or On Education*, 2d ed., trans. Allan Bloom (New York: Basic Books, 1979).

4. Gert Biesta, "Why 'What Works' Won't Work: Evidence-Based Practice and the Democratic Deficit in Educational Research," *Educational Theory* 57, no. 1 (2007): 1–22.

5. Andrew Davis, "The Need for a Philosophical Treatment of Assessment," *Journal of the Philosophy of Education* 32, no. 1 (1998): 1–18. Davis says that individual schools cannot be held constant in relation to each other, nor can they be held constant with themselves. Measuring a student over four years in one school assumes that conditions are static. However, the school in one year will be a different place four years later, and thus one cannot assume the different data sets are reliably commensurable.

10.47925/2011.177

6. W. James Popham, "Ten 'Must-Know' Facts About Educational Testing," National Parent-Teacher Association, 2001, http://www.pta.org/2553.htm; and W. James Popham, *The Truth About Testing: An Educator's Call to Action* (Alexandria, Va.: ASCD, 2001).

7. Edward L. Deci, Richard Koestner, and Richard M. Ryan, "The Undermining Effect Is a Reality After All — Extrinsic Rewards, Task Interest, and Self-Determination: Reply to Eisenberger, Pierce, and Cameron (1999) and Lepper, Henderlong, and Gingras (1999)," *Psychological Bulletin* 125, no. 6 (1999): 692–700; and Richard M. Ryan and Netta Weinstein, "Undermining Quality Teaching and Learning: A Self-Determination Theory Perspective On High-Stakes Testing," *Theory and Research in Education* 7, no. 2 (2009): 224–33. SQLA in HST environments decrease student *motivation* to learn, which then inhibits a student's *ability* to learn. Decreasing motivation levels will lead to decreased performance on subsequent high stakes SQLA. See also Andrew Davis, "High Stakes Testing and the Structure of the Mind: A Reply to Randall Curren," *Journal of Philosophy and Education* 40, no. 1 (2006); also John White, "Thinking about Assessment," *Journal of the Philosophy of Education* 33, no. 2 (1999): 201–11.

8. Biesta, "Why 'What Works' Won't Work," 7.

9. Ibid., 8.

10. Ibid. See also Gert Biesta, *Beyond Learning: Democratic Education for a Human Future* (Boulder, Co.: Paradigm Publishers, 2006).

11. Hans-Georg Gadamer, *Truth and Method*, 2d rev. ed., trans. Joel Weinsheimer and Donald G. Marshall (New York: Continuum, 2005), 269.

12. Ibid., 294.

13. Ibid.

14. Access to prior test scores does not constitute access to prior knowledge. It allows for the comparison of test scores, but not of the foreconceptions present at the time of learning.

15. Davis says this is not a skepticism about whether or not we can know other minds, but instead that the "conception of knowledge and of psychological states seemingly required by the demands of high stakes testing is flawed." See Davis, "High Stakes Testing," 2.

16. Or, as Popham states, "The skills and knowledge children possess can't be seen" (Popham, "Ten 'Must-Know' Facts.") Teachers and evaluators can only make inferences as indicated by the numbers; the numbers themselves do not tell us something definite or about the learning.

17. Michael Scriven, "Philosophical Inquiry Methods in Education," in *Complementary Methods for Research in Education*, ed. Richard M. Jaeger (Washington, D.C.: American Educational Research Association, 1988), 134.

18. This will vary depending on the analysis tool used, but in almost all cases it is relevant only in relation to other scores, not in relation to the student's learning given the knowledge possessed prior to the "treatment" or lessons of the course.

19. Dan Goldhaber, "The Mystery of Good Teaching," *Education Next* 2, no. 1 (2002): 50–55. Goldhaber found that only 21 percent of the variation in student test scores could be attributed to school effects (including 8.5 percent attributed to teachers), but 60 percent of the variation was attributable to nonschool effects such as individual and family characteristics. Findings indicate that 97 percent of the teacher contribution to student learning could not be isolated or identified. See W. James Popham, *America's "Failing" Schools: How Parents and Teachers Can Cope With No Child Left Behind* (New York: Routledge, 2004), 59–60 for discussion of score spread and influence of innate spatial abilities.

20. See Daniel Tröhler, "Harmonizing the Educational Globe: World Polity, Cultural Features, and the Challenges to Educational Research," *Studies in Philosophy and Education* 29, no. 1 (2010): 5–17, for a discussion about how the Finnish cultural of collaborative education emerges as the significant factor in Finland's success in out-performing other countries on the Program for International Assessment or PISA (standardized) assessments.

21. This does not include the "degrees" of learning offered by Davis: "rich" learning as opposed to the learning of "thin" knowledge. "It is not that we are bad at assessment, although we may be. It is rather that that rich or proper knowledge lacks the definitive and specific character which would be required if its presence were to be detectable by standard educational assessment devices." Davis, "The Need for a Philosophical Treatment of Assessment," 17.

10.47925/2011.177

22. Gadamer, *Truth and Method*, 294.

23. Common examples of HSTs are tests used to meet No Child Left Behind (NCLB) requirements.

24. See W. James Popham, "Stopping the Mismeasurement of Educational Quality," *The School Administrator* 57, no. 11 (2000): 12–15; and see also Popham, "The Truth about Testing," for explanation of the origins of achievement tests, the distinction between them and *aptitude* tests, and analysis of their usefulness as accountability measurements.

25. Ryan and Weinstein, "Undermining Quality Teaching and Learning." This work assumes a causal relationship between motivation and learning.

26. In most HST implementation, penalties are implemented twice as often as rewards. B. Miner, "Testing: Full Speed Ahead," in *Failing our Kids: Why the Testing Craze Won't Fix our Schools*, eds. K. Swope and B. Miner (Milwaukee, Wisc.: Rethinking Schools, 2000).

I would like to thank Megan Laverty and Michael Schapira for their questions, comments, and suggestions on early drafts of this essay.

10.47925/2011.177