Education Policy Analysis Archives

Volume 8 Number 46 September 17, 2000 ISSN 1068-2341

A peer-reviewed scholarly electronic journal
Editor: Gene V Glass, College of Education
Arizona State University

Articles appearing in **EPAA** are abstracted in the *Current Index to Journals in Education* by the ERIC Clearinghouse on Assessment and Evaluation and are permanently archived in *Resources in Education*.

Should Achievement Tests be Used to Judge School Quality?

Scott C. Bauer
University of New Orleans

Abstract

This study provides empirical evidence to answer the question whether student scores on standardized achievement tests represent reasonable measures of instructional quality. Using a research protocol designed by Popham and the local study directors, individual test items from a nationally-marketed standardized achievement test were rated by educators and parents to determine the degree to which raters felt that the items reflect important content that is actually taught in schools, and the degree to which raters felt that students' answers to the questions would be likely to be unduly influenced by confounded causality. Three research questions are addressed: What percentage of test items are considered suspect by raters as indicators of school instructional quality? Do educators and parents of school-age children differ in their ratings of the appropriateness of test items? Do educators and parents feel that standardized achievement test scores should be used as an indicator of school instructional quality? Descriptive statistics show that on average, raters felt that the content reflected in test questions measured material

> that is important for students to know. However, for reading and language arts questions, between about 20% to 40% of the items were viewed as suspect in terms of the other criteria.

Introduction

Since publication of *A Nation at Risk* in 1983, issues associated with accountability have been at the forefront of educational reform in the United States. Kirst (1990) estimated that in the 1980's alone, 40 states created or amended their accountability systems. Stecher and Barron (1999) note that the number of states with a mandated student testing program rose from 29 in 1980 to 46 in 1992. Presidents Bush and Clinton both proposed the creation of a voluntary national test that would allow the reporting of student performance in relation to national standards (Carnevale & Kimmel, 1997).

The emergence of high-stakes accountability policies has intensified the debate over whether state-mandated assessment is a useful instrument for changing educational practice (Firestone, Mayrowetz, and Fairman, 1998; Ginsberg and Berry, 1998; Sheldon and Biddle, 1998). Proponents of high-stakes testing assume that poor performance in American schools results from a lack of attention to school performance. "To solve such problems, according to this view, we need to set high standards for students, assess students' performance with standardized tests, and reward or punish students, their teachers, and their schools, depending on whether those standards are met" (Sheldon and Biddle, 1998, p. 165):

> Forty-nine states and a number of urban districts have set standards for what students should know and be able to do at various points in their school careers. Half the states hold schools accountable and apply sanctions to those whose students fail to meet the standards. At least a third – with more soon to follow – require students to score at designated levels on tests to get promoted and/or graduate. (Wolk, 1998, p. 48)

A recent survey by the Council of Chief State School Officers (1998) shows that while the states are increasingly introducing less traditional performance measures like portfolios into their assessment programs, 31 states use norm-referenced tests to measure student achievement in language arts, reading and mathematics. Tests are generally a part of the accountability system because they are inexpensive and quick to implement, and they are considered socially accepted as indicators of student performance (Linn, 1999).

At the heart of the debate over the use of high-stakes testing policies as a reform is the assumption that introducing new assessments will result in changes in teacher behavior in the classroom. As Firestone, Mayrowetz and Fairman (1998) observed, there is in fact a good deal of evidence that testing changes patterns of teaching, "if only by promoting 'teaching to the test'" (p. 96). There is evidence that school-based performance and reward programs such as Kentucky's produces desired results (Kelley and Protsik, 1997), and research supports the notion that school leaders take high-stakes testing very seriously (Mitchell, 1995). However, research also suggests that high- stakes testing programs do not necessarily provide valid data on students and schools (Stecher & Barron, 1999), and these systems tend to produce a high level of stress for teachers and principals. Critics argue that high-stakes testing may encourage teachers to consider test scores as ends in themselves:

> Evidence...reveals various perils associated with rigid standards, narrow

> accountability, and tangible sanctions that can debase the motivations and performances of teachers and students. Teachers faced with reforms that stress such practices may become controlling, unresponsive to individual students, and alienated. Test- and sanction-focused students may lose intrinsic interest in subject matter, learn at only a superficial level, and fail to develop a desire for future learning. (Sheldon and Biddle, 1998, p. 164)

Opponents of these measures conclude that they result in dumbing-down the curriculum (e.g., Corbett and Wilson, 1991), while others argue that they deny the reality of the situation faced by students, particularly those in urban districts, who are not well prepared to meet harsh standards (Wolk, 1998). Still others question whether policy is an effective instrument for shaping instructional practice at all (e.g., Cohen, 1995). Newmann, King and Rigdon argue that high-stakes accountability programs are doomed to failure because insufficient attention is paid to increasing schools' capacity for change, and Mayer (1998) raises the question of whether pursuing standards-based reform while leaving testing policy largely unchanged undermines reform. Wallace (2000, p. 66) concludes, "Provincial achievement exams create undue pressure on students, teachers, and schools. Even worse, the tests fail to assess what students will need to know in the next century."

Nevertheless, rating school performance based on the results of state testing programs has become an increasingly popular feature of state accountability programs (Watts, Gaines & Creech, 1998). The CCSSO survey referenced earlier indicates, in fact, that standardized achievement tests generally serve as summative indicators of elementary, middle, and high school performance, at least in part. For instance, in my home state of Louisiana, the new testing program is used to produce a school performance score that includes scores from the state's criterion-referenced test (60% of score), a nationally-marketed norm-referenced test (30% of score), and student attendance and dropout rates (10 percent of score). The school performance score will be used to establish 10-year goals, and schools will be held accountable for reaching two-year targets that represent progress toward these goals. A series of corrective actions are spelled out for schools that fail to meet their targets (Louisiana's School and District Accountability System, 1999).

At the 1998 Annual Meeting of the Mid-South Educational Research Association, W. James Popham raised the following question: Is it *appropriate* to use norm-referenced tests to evaluate instructional quality? Specifically, he challenged participants to consider whether norm-referenced tests measure knowledge that is taught and learned in schools. Popham then invited researchers to participate with him in a study to answer the question: Should student scores on standardized achievement tests be used to evaluate instructional quality in local schools?

In a subsequent paper, Popham (1999) laid out the basic argument that frames this study. While standardized achievement tests are useful tools to provide evidence about a specific students' mastery of knowledge and skills in certain content domains, "Employing standardized achievement tests to ascertain educational quality is like measuring temperature with a tablespoon" (p. 10). There are several difficulties with using aggregate measures from norm-referenced tests to judge the performance of a school. First, there is considerable diversity across states and school systems with regard to content standards, and therefore test developers produce "one-size-fits-all assessments" which do not adequately align with what's supposed to be taught in schools. Second, because norm- referenced tests must provide a mechanism to differentiate between students based on a relatively small number of test items, test

developers select "middle difficulty" items. As Popham put it,

> As a consequence of the quest for score variance in a standardized achievement test, items on which students perform well are often excluded. However, items on which students perform well often cover the content that, because of its importance, teachers stress. Thus the better the job that teachers do in teaching important knowledge and/or skills, the less likely it is that there will be items on a standardized achievement test measuring such knowledge and skills (p. 12).

Finally, scores on standardized achievement tests may not be attributable to the instructional quality of a school. Student performance may be caused by any number of factors, including what's taught in schools, a student's native intelligence, and out-of-school learning opportunities that are heavily influenced by a students' home environment. Popham terms this last issue the problem of "confounded causality."

Here we report the results of one of several local studies designed to provide empirical evidence to answer the question of whether student scores on standardized achievement tests represent reasonable measures of instructional quality. Using a research protocol designed by Popham and the local study directors, individual test items from a nationally-marketed standardized achievement test were rated by educators and parents to determine the degree to which raters felt that the items reflect important content that is actually taught in schools, and the degree to which raters felt that students' answers to the questions would be likely to be unduly influenced by confounded causality. Three research questions are addressed:

1. What percentage of test items are considered suspect by raters as indicators of school instructional quality?
2. Do educators and parents of school-age children differ in their ratings of the appropriateness of test items?
3. Do educators and parents feel that standardized achievement test scores should be used as an indicator of school instructional quality?

Methods

The investigation consisted of a series of three separate item-review studies designed to secure evidence regarding the appropriateness of using students' scores on standardized achievement tests as evidence of instructional quality. All sections of a nationally-marketed standardized achievement test was studied at the third grade level. The test covers mathematics, reading and language arts content areas. The test was secured by the local study director, who also took responsibility for security.

Participants

Participants were solicited from two sources. First, principals associated with the School Leadership Center of Greater New Orleans (SLC-GNO) were invited to put together teams of teachers and parents to host an item-rating session. Two principals were able to put together groups of ten and eleven raters. From these 21 participants, 10 were parents and 11 were educators. These rating sessions were held at the participant's schools after school hours. Additionally, nine teachers enrolled in a graduate level course dealing with testing and measurement at the University of New Orleans formed a

third group. This rating session was held on campus. In sum, then, 30 reviewers served as item raters, including two principals, 18 teachers, and 10 parents of elementary school children.

Procedures

Reviewers were provided with a description of the goals and procedures associated with the study prior to the actual rating session. In addition to signing a standard human subjects protocol outlining the responsibilities and risks associated with participation, reviewers signed a test- confidentiality form prior to their participation, and the item reviews were carried out under the scrutiny of the local director so that no security violations could occur. All test booklets were retained by the study director. Data were recorded on forms that do not reveal the specific test reviewed or any test questions.

Reviewers were asked to make their item-by-item judgments individually on summary rating sheets (see Exhibit 1 for a sample of the rating sheet), without group discussion, using a protocol that asked them to examine test items and judge their appropriateness in terms of five criteria:

1. IMPORT: Is the skill or knowledge measured by this item truly important for children to learn?
2. TAUGHT: Is the skill or knowledge measured by this item likely to be taught if teachers follow the prescribed curriculum?
3. SES: Is this item *free* of qualities (form or content) that will make the likelihood of a student's answering correctly be dominantly influenced by the student's socioeconomic status?
4. INHERITED CAPABILITIES: Is this item *free* of qualities (form or content) that will make the likelihood of a student's answering correctly be dominantly influenced by the student's inherited academic capabilities?
5. VALIDITY: Will a student's response to this item contribute to a valid inference about the student's status regarding whatever the test is supposed to be measuring?

Exhibit 1. Sample item rating sheet

Item	Import?	Taught?	SES?	IQ?	Validity?
1	Y ? N	Y ? N	Y ? N	Y ? N	Y ? N
2	Y ? N	Y ? N	Y ? N	Y ? N	Y ? N
3	Y ? N	Y ? N	Y ? N	Y ? N	Y ? N
4	Y ? N	Y ? N	Y ? N	Y ? N	Y ? N
5	Y ? N	Y ? N	Y ? N	Y ? N	Y ? N

Exhibit 2. The five item-review questions

1. IMPORT: Is the skill or knowledge measured by this item truly important for children to learn?

2. TAUGHT: Is the skill or knowledge measured by this item likely to be taught if teachers follow the prescribed curriculum?

3. SES: Is this item *free of qualities* (form or content) that will make the likelihood of a student's answering correctly be dependent on the student's socioeconomic status? **WOULD A STUDENT FROM A WELL-OFF HOME BE UNLIKELY TO GET THE ITEM CORRECT JUST BECAUSE HE OR SHE IS MORE "ADVANTAGED?"**

4. IQ: Is this item *free of qualities* (form or content) that will make the likelihood of a student's answering correctly be dependent on the student's inherited academic capabilities? **WOULD A STUDENT WITH GREATER NATIVE INTELLIGENCE (IQ) BE UNLIKELY TO GET THE ITEM CORRECT JUST BECAUSE OF THIS INBORN QUALITY?**

5. VALIDITY: Will a student's response to this item contribute to a valid conclusion about the student's ability relating to whatever the test is supposed to be measuring? **IS THIS ITEM A VALID MEASURE OF THE ABILITY THE TEST IS MEASURING IN THIS SECTION OF THE TEST?**

During an orientation phase, prior to item-review, the local study director practiced reviewing a selection of test items from a test-booklet's sample items and/or from a different test to clarify item-reviewers' understanding of the five item-review questions. During a pre-test of the procedure, it became clear that respondents may have difficulty with the questions related to SES, IQ, and validity, thus some clarifying language was added and a summary sheet was provided to raters which allowed them to access the definitions as they performed the ratings. (Exhibit 2 shows the summary sheet.)

Each rating session was held in the afternoon, and took approximately three hours. Because of the time of day and the considerable investment of time and energy, participants were provided with a light dinner after each rating session. They also participated in a short debriefing session, during which they answered questions about the methodology and their ability to sensibly rate the test items.

Analysis

Response sheets were collected and numbered after each session. The number of items rated yes, no, or with a question mark (not sure) were tallied for each content area of the test, and the number of no and "not sure" (question mark) ratings were entered into an SPSS 9.0 for Windows system file. To address the question of what percentage of test items raters considered suspect as indicators of school instructional quality, the mean percentages of items rated "no" or "not sure" were computed for each of the rating criteria and for each content area of the test. Descriptive statistics related to the raters' judgments of items in each content area of the test and for each of the criteria are presented. Additionally, a summary statistic indicating the mean percentage of items rated as suspect on at least one criterion was computed. For purposes of discussion, the percentage of items rated as either a "no" or "not sure" are combined; given the high-stakes involved in the state accountability programs, if raters cannot determine if an item meets the criteria used in this study, we will consider it suspect. The full breakdown of ratings are presented in the Appendices.

To see if educators and parents of school-age children differ in their ratings of the

appropriateness of test items, analysis of variance was computed to test whether the mean ratings are statistically significant. Eta-squared is also reported; Stevens (1996) recommends that to interpret the effect size, an eta-squared of .01 should be treated as a small effect, .06 a medium effect, and .14 a large effect.

To address whether educators and parents feel that standardized achievement test scores should be used as an indicator of school instructional quality, the frequency distribution is reported for a summary question which asked respondents to answer yes, no, or "not sure" in regard to this question. Chi-square was computed to see if there is a statistically significant association between the answer to this summary question and group membership.

As a final portion of the study, answers to questions posed during debriefing sessions were analyzed to determine whether raters felt confident in their ability to assess test items on these criteria. In an exploratory study such as this, rater's sense of their ability to render reliable judgments in terms of these criteria is an important question. These data may shed some light on whether the methodology provides a valid assessment of the usefulness of the test to judge school quality.

Results

Table 1 displays the mean percentage of test items rated as suspect by respondents. As mentioned earlier, the percentage reflects the number of items rated as either a "no" or "not sure" on each of the five criteria for each content area of the test. Overall, the mean percentage of items rated as suspect varies widely; only 2% of the items were rated as suspect in importance for math procedures, whereas 41% of the vocabulary items were rated as suspect because the likelihood seemed great that student's answering correctly would be dependent on the student's inherited academic capabilities (IQ). Overall, raters felt that the items dealing with reading and language arts were more often suspect as indicators of school quality, especially in terms of the likelihood that students' answering these items correctly would be unduly influenced by native intelligence (IQ) or socio-economic status (SES). Raters were somewhat more comfortable with measures relating to mathematics problem-solving and reasoning, and considerably more comfortable with the items measuring mathematics procedures.

Table 1
Mean percentage of items rated as "suspect"
for each content area

Content Area	Important?	Taught?	SES?	IQ?	Valid?
Vocabulary	11%	26%	38%	41%	26%
Reading comprehension	14	26	38	40	26
Grammar & language	8	24	37	38	21
Math problem solving & reasoning	11	24	19	33	21
Math procedures	2	7	11	22	10

Viewing the data in Table 1 in terms of criteria instead of content area, one sees that from among the various criteria used to rate test items, raters judged the test items more likely to be suspect in terms of SES and IQ. That is, from among the five possible reasons a test item might be inappropriate to assess school quality, raters felt the greatest threat to validity was the likelihood that a student might answer the item correctly because of socio- economic advantage or because of native intelligence rather than because of what he or she learned in school. In fact, for the reading and language arts content areas, between 30 and 40% of the items were rated as suspect in these regards. Considerably fewer items were rated as suspect because they were deemed unimportant for students to know, and for most content areas between 20 and 30% of the items were deemed unacceptable because raters felt that the material was not a part of the standard curriculum at that grade level.

The above-mentioned data show the mean percentage of items rated as suspect on each of the five criteria; a final summary statistic was computed to show the mean percentage of items in each section of the test that was rated as suspect on *at least one* of the five criteria. Table 2 shows that for all areas of the test, approximately 50% of the items were deemed inappropriate as indicators of instructional quality on at least one criterion. The table also shows that the range of ratings is considerable – for most areas, at least one rater felt that nearly all of the items were alright as indicators of instructional quality on all criteria, and at least one rater felt that all items were suspect on at least one of the five criteria.

Table 2
Mean percentage of items deemed suspect
on *at least* one criterion

Content area	Mean %	High	Low
Vocabulary	57%	100%	15%
Reading comprehension	52%	100%	3%
Grammar & language	55%	100%	13%
Math problem solving & reasoning	48%	100%	3%
Math procedures	46%	100%	0%

To address the question of whether educators and parents rated the test items differently, analyses of variance were computed to test the null hypothesis that the mean percentages do not differ between the two groups of respondents. These data, presented in Table 3, show that the only statistically significant differences between the mean percentage of items rated as suspect by parents and educators exist for the criteria dealing with whether the content measured by the test item is taught in the regular school curriculum (taught). Parents consistently felt that a greater percentage of the items on the test covered material that would not be a part of the standard curriculum. An examination of eta-squared shows that for most of the content areas, the effect size of the difference in means for this criterion (taught) is large (eta^2 for vocabulary=.16, for reading comprehension=.16, for math problem-solving=.19) or moderate (eta^2 for

grammar and language=.10, for math procedures=.11).

Table 3
Mean ratings by respondent group

Vocabulary					
Respondent	IMP	TAU	SES	IQ	VAL
Educator	.09	.20	.43	.44	.21
Parent	.14	.39	.29	.37	.35
F (1,28)	.72	5.40*	1.88	.47	2.89
Eta-squared	.03	.16	.06	.02	.09
Reading comprehension					
Educator	.11	.20	.43	.38	.21
Parent	.20	.39	.29	.46	.35
F (1, 28)	1.99	5.40*	1.88	.50	2.89
Eta-squared	.07	.16	.06	.02	.09
Grammar and language					
Educator	.07	.18	.38	.38	.22
Parent	.10	.35	.35	.37	.20
F (1, 28)	.69	2.95	.04	.02	.08
Eta-squared	.02	.10	.01	.00	.00
Math problem-solving & reasoning					
Educator	.11	.17	.19	.35	.18
Parent	.10	.37	.17	.29	.25
F (1, 28)	.01	6.36*	.04	.25	1.08
Eta-squared	.00	.19	.00	.01	.04
Math procedures					
Educator	.02	.05	.08	.24	.12
Parent	.01	.12	.16	.19	.07
F (1, 28)	.00	3.39	.59	.11	.85

Eta-squared	.00	.11	.02	.00	.03

* p<.05

Table 4 shows the results for the summary item that asked raters to judge whether they would recommend using standardized achievement tests as an indicator of instructional quality. Results show that approximately a quarter of the educators and 30% of the parents felt that standardized achievement tests ought to be used as an indicator of school quality, whereas about two-thirds of the educators and 40% of the parents felt that they should not. Another 30% of the parents and 11% of the educators were not sure, and one respondent left the question blank. The chi-square test of association showed that there is not a statistically significant association between the answer to this question and role [X^2 (2, n=29) = 2.11, p<.05].

Table 4
Should standardized tests be used to measure instructional quality?

	Yes	**Not sure**	**No**
Educator	5 (26%)	2 (11%)	12 (64%)
Parent	3 (30%)	3 (30%)	4 (40%)
Total	8 (28%)	5 (17%)	16 (55%)

The final data collected in this study had to do with the methodology itself. A formal debriefing was held after each item rating session. Respondents were asked a short series of questions in writing about their ability to rate test items and about the kinds of factors they felt influenced their ratings. Raters also discussed their experiences and any difficulties they perceived with the rating process. These data provide us with some sense of the threats to validity present in the ratings.

Respondents were asked to rate how easy they felt it was to make judgments about the test items, on a scale of 1 = "very easy" to 10 = "very difficult." On average, these data show that respondents felt that it was relatively easy to assess whether an item measured import material for students to know (2.1) and whether the item was likely to be taught as a part of the regular curriculum (2.9). Raters found it most difficult to rate whether an item would be more likely to be answered correctly because of a child's inherited capabilities (IQ) or socio-economic status (5.0 and 4.5, respectively). Respondents also found it relatively more difficult to judge whether an item was a valid measure of the skill it was intended to measure (4.7). Overall, then, on a ten-point scale raters found their job moderately easy (i.e., lower than the midpoint between very easy and very difficult), though some criteria were more difficult to apply than others.

Respondents also answered open-ended questions that probed into the kinds of factors that they felt might threaten their ability to render reliable judgments about the test items. These answers show that most of the parents felt at least a bit unsure about what was in the regular or "official" curriculum, thus they were not sure about the reliability of their judgments on the criterion labeled "taught." One respondent pointed

out that SES and IQ were tough to assess because these relate to a subjective assessment of the fairness of an item, and several other respondents noted that SES was likely influenced by their own socio-economic status. That is, they questioned whether relatively well-off parents or teachers could render a valid judgment on this criterion. Some teachers questioned whether their beliefs about teaching would "get in the way" of their ability to rate the items, and several raters simply said that they found it tough – "speculative" – to assess the degree to which a students' answer on a test item would relate more to native intelligence than knowledge gained in school.

Summary and Conclusions

The purpose of this study was to attempt to amass credible evidence concerning whether student scores on standardized achievement tests should be used to evaluate instructional quality in local schools. Using a framework developed by Popham (1999) and a research protocol collaboratively devised by Popham and local study directors, educators and parents of school-age children rated all items contained on a commercially-marketed standardized achievement test that covered third grade content in reading, language arts, and mathematics. Descriptive statistics show that on average, raters felt that the content reflected in test questions measured material that is important for students to know. However, for reading and language arts questions, between about 20% to 40% of the items were viewed as suspect in terms of the other criteria. Raters saw fewer problems with questions dealing with mathematics problem-solving and reasoning, and they felt the fewest problems existed with questions on mathematical procedures. Overall, though, raters felt that about half of all items they appraised were suspect on at least one of the criteria used to assess the test. Educators and parents did not differ statistically on their ratings on most criteria, though about two-thirds of the educators felt that tests should not be used to judge instructional quality whereas only 40% of the parents felt this way. The range of ratings across respondents was considerable for all content areas and for each of the rating criteria; some respondents saw very few problems with any questions, while others felt that the vast majority of items were suspect on at least one criterion.

This study was prompted by the realization that while standardized achievement tests are useful tools to provide evidence about students' mastery of knowledge and skills in tested content domains, it does not logically follow that they should be useful as indicators of school performance. As reflected in the rating scheme used in this study, student performance on standardized tests may be caused by any number of factors, including what's taught in schools, a student's native intelligence, and out-of-school learning opportunities that are heavily influenced by a students' home environment.

The question that follows, then, is whether this confounded causality poses a problem in terms of using standardized test scores as measures of instructional quality. In a critique of Popham's argument regarding confounded causality, Schmoker (2000) argues that it does not. What happens in classrooms can "significantly mitigate and even overcome environmental and genetic factors" (p. 64), and standardized tests give schools focus and empower teachers by providing specific data on students' needs. "Standardized test results have provided the essential focus and urgency for schools to improve and refine instructional programs in reading, writing, and math practices" (p. 64).

This argument misses the point. There is no question that norm-referenced tests are exceedingly valuable in their intended purpose: to identify knowledge and skills that individual students need to improve, thus providing professional educators with essential data with which they can craft programs and practices. It does not follow, however, that

using aggregate average scores on standardized tests serves as a good indicator of *school quality*. To say that norm-referenced tests can help teachers identify areas in need of attention does not rely on an assumption that school programs alone caused a deficiency; instead, as Schmoker observed, this relied on the belief that schools can do something to overcome the deficiency regardless of cause.

The notion that aggregate scores on standardized tests should serve as an indicator of school quality relies on an assumption of causality. The underlying logic is that the scores are *predominantly* caused by something the school does or has some control over. For this assumption to hold, at a minimum we must be willing to believe that student performance on standardized tests is related to school quality, that the tests measure the skills and abilities stressed in school programs, and that there are no antecedent factors that might otherwise explain aggregate student performance on the tests. If the data presented here are credible, the soundness of this assumption must be questioned. On average about half of the items on the rated test suffer from "confounded causality" on at least one of these criteria.

The question of whether the data presented here are, in fact, "credible," deserves attention. The data collected from debriefing presented earlier barely scratch the surface of the potential threats to validity. Perhaps the biggest issues stem from the fact that the study was purposefully constructed to include both educators and parents. The fact that parents felt less knowledgeable about what *should be* in the regular school curriculum may have resulted in an exaggeration of the percentage of items that were deemed suspect on this criterion. Additionally, some respondents felt it difficult to judge whether items might be unduly influenced by a students' native intelligence (where do you draw the line between native intelligence and knowledge learned in school?) and some felt that their own social standing made it hard for them to determine if a students' socio-economic background would greatly influence the likelihood of answering a test item correctly.

Regardless of criterion, the rating process asked for a judgment, that is, the subjective assessment of an item's appropriateness. These are difficult conclusions to make. Yet, in terms of the message to policy-makers, this is precisely be the point. Aggregate average scores on standardized tests are at best a gross approximation of the instructional quality of a school, and any number of factors may have more to do with the production of this number than the quality of educational services delivered. We should be questioning what these numbers mean, especially considering the fact that in many states the numbers are being used to reward or punish school staff and students.

By design, policy makers have raised the stakes. As this analysis shows, though, when you get beneath the summary number and ask whether the test items that go into producing that number are sensible measures of knowledge and skills learned in school, the answer is far from clear. This would suggest, at a minimum, that policy-makers should consider eliminating or de-emphasizing their use of norm-referenced achievement tests as a barometer of how well a school is doing.

Note

Research presented here was supported by a grant from the School Leadership Center of Greater New Orleans. An earlier version of this work was presented at the Annual Meeting of the Mid-South Educational Research Association, Point Clear, AL, November 1999.

References

Carnevale, A. & Kimmel, E. (1997). *A national test: Balancing policy and technical issues*. Princeton, NJ: Educational Testing Service.

Cohen, D. (1995). What is the system in systemic reform? *Educational researcher, 24*, 11-17.

Corbett, H. and Wilson, B. (1991). *Testing, reform and rebellion.* Norwood, NJ: Ablex.

Council of Chief State School Officers. (1998). Key state education policies on K-12 education: Standards, graduation, assessment, teacher licensure, time and attendance - a 50- state report. Washington, DC: Author.

Firestone, W., Mayrowetz, D., and Fairman, J. (1998). Performance-based assessment and instructional change: The effects of testing in Maine and Maryland. *Educational evaluation and policy analysis, 20*, 95-113.

Ginsberg, R. and Berry, B. (1998). The capability for enhancing accountability. In R. MacPherson (Ed.), *The politics of accountability: Educative and international perspectives* (pp. 43-61). Newbury Park, CA: Corwin.

Kelley, C. and Protsik, C. (1997). Risk and reward: Perspectives on the implementation of Kentucky's school- based performance award program. *Educational Administration Quarterly, 33,* 474-505.

Kirst, M. (1990). Accountability: Implications for state and local policymakers. Washington, DC: OERI.

Linn, R. (1999). Standards-based accountability: Ten suggestions (CRESST Policy Brief). Los Angeles: National Center for Research on Evaluation, Standards and Student Testing.

Mayer, D. (1998). Do new teaching standards undermine performance on old tests? *Educational evaluation and policy analysis, 20,* 53-74.

Mitchell, K. (1995). Reforming and conforming: NASDC principals talk about the impact of accountability systems on school reform (Technical Report 143). Santa Monica, CA: Rand Corp.

Popham, W. J. (1999, March). Why standardized tests don't measure educational quality. *Educational Leadership, 57*, 8-15.

Schmoker, M. (2000, February). The results we want. *Educational Leadership, 58*, 62-65.

Sheldon K. and Biddle, B. (1998). Standards, accountability and school reform: Perils and pitfalls. *Teachers' College Record 100 (1)*: 164-180.

Stecher, B. & Barron, S. (1999). Test-Based Accountability: The perverse consequences of milepost testing. Paper presented at the Annual Meeting of the American Educational Research Association, Montreal, Canada, April 1999.

Stevens, J. (1996). *Applied multivariate statistics for the social sciences* (3rd ed.).

Mahwah, NJ: Lawrence Erlbaum Associates.

Wallace, D. (2000, February). Results, results, results? *Educational Leadership, 58*, 66-68.

Watts, J., Gaines, G. & Creech, J. (1998). Getting results: A fresh look at school accountability. Atlanta: Southern Regional Education Board.

Wolk, R. (1998). Education's high-stakes gamble. *Education Week, 18 (15)*: 48.

About the Author

Scott C. Bauer
348J Bicentennial Education Bldg.
University of New Orleans
New Orleans, LA 70148
504-280-6446 (voice)
504-280-6453 (fax)
Email: sbauer@uno.edu

Scott C. Bauer (Ph.D., Cornell University) is an assistant professor in the Educational Leadership, Counseling, and Foundations Department in the College of Education at the University of New Orleans, and Director of Research at the School Leadership Center of Greater New Orleans. His research and teaching focuses on the application of the principles of organizational behavior and development to the study of school leadership, organizational change and restructuring. Dr. Bauer's most recent work deals with designing and implementing site-based decision making systems in schools.

Appendix A
Descriptive statistics:
mean percentages, standard deviations
and range of all ratings

Skill area	Criteria	Rating	Mean	sd	high	low
Vocabulary	importance	not sure	.07	.11	.40	0
		no	.04	.06	.20	0
	taught	not sure	.21	.22	1.00	0
		no	.06	.09	.40	0
	SES	not sure	.14	.15	.75	0
		no	.24	.23	1.00	0
	IQ	not sure	.15	.12	.35	0
		no	.27	.24	1.00	0
	Validity	not sure	.12	.14	.45	0
		no	.14	.15	.75	0
Skill area	Criteria	Rating	Mean	sd	high	low

Reading comprehension	importance	not sure	.07	.12	.53	0
		no	.07	.10	.37	0
	taught	not sure	.16	.23	1.00	0
		no	.07	.11	.30	0
	SES	not sure	.08	.09	.30	0
		no	.20	.25	.93	0
	IQ	not sure	.13	.19	.93	0
		no	.28	.27	.97	0
	Validity	not sure	.12	.13	.57	0
		no	.15	.17	.77	0
Skill area	Criteria	Rating	Mean	sd	high	low
Grammar and Language	importance	not sure	.05	.07	.23	0
		no	.03	.05	.23	0
	taught	not sure	.17	.21	1.00	0
		no	.07	.16	.77	0
	SES	not sure	.13	.14	.57	0
		no	.24	.27	1.00	0
	IQ	not sure	.12	.15	.47	0
		no	.26	.32	1.00	0
	Validity	not sure	.10	.11	.40	0
		no	.12	.11	.40	0
Skill area	Criteria	Rating	Mean	sd	high	low
Math problem-solving and reasoning	importance	not sure	.06	.11	.50	0
		no	.05	.06	.27	0
	taught	not sure	.17	.22	1.00	0
		no	.07	.14	.57	0
	SES	not sure	.05	.07	.27	0
		no	.13	.25	.97	0
	IQ	not sure	.07	.08	.33	0
		no	.27	.30	.97	0
	Validity	not sure	.10	.10	.30	0
		no	.11	.14	.67	0
Skill area	Criteria	Rating	Mean	sd	high	low
Math Procedures	importance	not sure	.01	.03	.15	0
		no	.01	.03	.15	0
	taught	not sure	.05	.10	.40	0
		no	.02	.04	.15	0

	SES	not sure	.01	.02	.05	0
		no	.10	.27	1.00	0
	IQ	not sure	.03	.05	.20	0
		no	.20	.36	1.00	0
	Validity	not sure	.03	.05	.15	0
		no	.08	.15	.50	0

Appendix B
Descriptive statistics: mean percentages, standard deviations and range of combined ratings

	Criteria	Mean	sd	high	low
Vocabulary	importance	.11	.14	.50	0
	taught	.26	.23	1.00	0
	SES	.38	.27	1.00	0
	IQ	.41	.24	1.00	0
	validity	.26	.20	.80	0
Reading comprehension	importance	.14	.16	.57	0
	taught	.26	.23	1.00	0
	SES	.38	.27	1.00	0
	IQ	.40	.30	.97	0
	validity	.26	.20	.80	0
Grammar and Language	importance	.08	.10	.33	0
	taught	.24	.25	1.00	0
	SES	.37	.27	1.00	0
	IQ	.38	.32	1.00	0
	validity	.21	.18	.77	0
Math problem-solving and reasoning	importance	.11	.12	.50	0
	taught	.24	.22	1.00	0
	SES	.19	.24	.97	0
	IQ	.33	.30	1.00	0
	validity	.21	.17	.80	0
Math Procedures	importance	.02	.05	.20	0
	taught	.07	.10	.40	0
	SES	.11	.27	1.00	0
	IQ	.22	.38	1.00	0
	validity	.10	.16	.60	0

The World Wide Web address for the *Education Policy Analysis Archives* is **epaa.asu.edu**

General questions about appropriateness of topics or particular articles may be addressed to the Editor, Gene V Glass, glass@asu.edu or reach him at College of Education, Arizona State University, Tempe, AZ 85287-0211. (602-965-9644). The Commentary Editor is Casey D. Cobb: casey.cobb@unh.edu .

EPAA Editorial Board

EPAA Spanish Language Editorial Board