Education Policy Analysis Archives/Archivos Analíticos de Políticas Educativas
ISSN: 1068-2341
epaa@alperin.ca
Arizona State University
Estados Unidos

Schutz, Dick
Toward Educational Testing Reform: Inside Reading Achievement Tests
Education Policy Analysis Archives/Archivos Analíticos de Políticas Educativas, vol. 21, enero, 2013, pp. 1-29
Arizona State University
Arizona, Estados Unidos

COMMENTARY

education policy analysis archives

A peer-reviewed, independent, open access, multilingual journal

epaa aape

Arizona State University

Volume 21 Number 90 December 20th, 2013 ISSN 1068-2341

Toward Educational Testing Reform: Inside Reading Achievement Tests

Dick Schutz
3RsPlus, Inc.
United States

Citation: Schutz, D. (2013). Toward educational testing reform: Inside reading achievement tests. *Education Policy Analysis Archives, 21*(90). Retrieved [date], from http://epaa.asu.edu/ojs/article/view/1288

Abstract: The commentary (1) uses the U. S. National Assessment of Educational Progress (*NAEP*) as a prototype for examining standardized reading achievement tests at the item level, and (2) sketches an alternative based on an initiative underway in the United Kingdom.
Keywords: achievement testing; NAEP

Hacia una Reforma de los exámenes educativos: Un análisis de las pruebas de aprovechamiento de lectura
Resumen: Este comentario (1) usa la Evaluación Nacional de Progreso Educativo (*NAEP*) como prototipo para examinar las pruebas de rendimiento estandarizadas de lectura a nivel de los ítems individuales, y (2) esboza una alternativa basada en una iniciativa en marcha en el Reino Unido.
Palabras clave: Pruebas de rendimiento; NAEP

Para uma reforma de testes educacionais: uma análise da leitura de testes de desempenho.

Journal website: http://epaa.asu.edu/ojs/
Facebook: /EPAAA
Twitter: @epaa_aape

Manuscript received: 12/19/2012
Revisions received: 11/12/2013
Accepted: 11/12/2013

Resumo: Este comentário (1) usa a Avaliação Nacional do Progresso Educacional (*NAEP*) como um protótipo para analisar o desempenho de leitura padronizada evidência a nível de itens individuais, e (2) apresenta uma alternativa baseada em uma iniciativa lançada no Reino Unido.
Palavras-chave: probas de desempenho; NAEP

Introduction

This commentary picks up where two books end. The books together present a critique of prevailing achievement testing practice that has not been refuted, but the practices prevail. The first book, *Toward the Reform of Program Evaluation,* was published in 1980, just a few years before *A Nation at Risk* (National Commission on Excellence in Education, 1983) triggered the long-term crusade for educational "reform." The book summarizes the work of the Stanford Evaluation Consortium, a multi-disciplinary, multi-institution initiative that operated between 1973 to 1979, in which leading scholars under the leadership of Lee Cronbach exchanged views and arrived at consensus on educational policy issues?. Following the precedent of Martin Luther, the book outlines "95 Theses." A sense of the Consortium's goals can be gained from an extract of five theses:

- Accountability emphasizes looking back in order to provide praise and blame. Evaluation is better used to understand events and processes for the sake of guiding future activities.
- It is better for an evaluative inquiry to launch a small fleet of studies than to put all the resources into a single approach.
- Merit lies not in the form of inquiry but in relevance of information.
- External validity—that is, the validity of inferences that go beyond the data—is the crux. Increasing internal validity by elegant design often reduces relevance.
- The evaluator is an educator; success is to be judged by what others learn.

Obviously, the Consortium's theses did not have the same effect as Luther's; the book has been long out of print and has had no identifiable impact on testing practices to date. Although the term "reform" in educational parlance has been reduced to an honorific label attached to whatever nostrum is being proposed, the reforms formulated by the Consortium are as relevant today as they were earlier. What the Consortium vision lacks is workable methodology to effect its realization.

The second book that provides a take off for this commentary is Diane Ravitch's *The Death and Life of the Great American School System: How Testing and Choice Are Undermining Education* (2011). Ravitch makes a strong case that prevailing testing practices are undermining pre-collegiate education in the United States. But she, like others, does not look inside the tests; Ravitch's critique ends without offering an alternative scenario for providing more relevant and useful information to illuminate educational practice

This commentary focuses on standardized reading tests, since they are at the heart of achievement testing. It examines the tests at the item level and then offers alternative methodology consistent with the Evaluation Consortium's vision and absent the undesired effects that Ravitch analyzes. Many standardized reading achievement tests are administered, but the content of these tests is rarely examined. Some possible reasons for this include:

- The tests are shrouded in secrecy and are considered "confidential."
- Much is made of test security, teaching to the test, and cheating.
- Few adults have warm feelings about the tests that they were forced to take as students.

- The tests are for kids and do not make for interesting adult reading.
- A few sample test items are sufficient to convince one that "The tests look OK and there is no reason to look any further."
- The test results are typically reported as a single number. Beyond some sense of whether the number is high, low, or average, few people know or care about how the number was derived.
- The tests are held by authorities to be "valid and reliable" and "the best measures available." What more needs to be said?

Until the advent of the Internet, test items were just not widely available for general inspection. Currently, samples of the National Assessment of Educational Progress (*NAEP*) as well as state tests in the United States, national tests in Australia and the United Kingdom, and international tests are accessible to the public online. These tests contain a dizzying array of items, but the good news is that people interested in understanding reading assessments do not need to read all of them. A great deal of insight into all the tests (including those outside reading) can be gained by considering the *NAEP* tests—which have the greatest detailed information. The rest of the tests are "same song, different verses." I provide links for anyone caring to confirm my contention that all the tests sing the "same song" (see appendices) and I then propose an alternative orientation and methodology.

About the *NAEP* Reading Test

The distinctive feature of *NAEP* is the presumed continuity and uniformity in the tests, which make it possible to compare results across calendar years (generally from 1970 and particularly from 1992 to the present) and across grades (4, 8, and 12). The content of the test is allegedly maintained constant by the structure of a test framework. However, comparisons across grades and across years are not a function of the framework structure but of the statistical score scale—the numbers—used to report the test results. Both the test framework and the score scale warrant brief description.

The Test Framework

The *NAEP Reading Framework* is provided in separate documents for the years 2007, 2009, 2011, and 2013 (Table 1).[1] Actually, the framework described for 2007 prevailed from 1992-2007, and the frameworks presented for 2009-2013 are the same descriptions—that is, the 2009 Framework is essentially repeated for subsequent years. The modification of the framework in 2009, outlined in Table 1, was accompanied by no noticeable change in the test items. Much more verbiage is used to describe the frameworks, but the tabled categories are the only part of the description that is actually used to categorize test items.[2]

---

[1] http://www.nagb.org/publications/frameworks.htm

[2] *NAEP* operates under the auspices of the Congressionally-mandated National Assessment Governing Board. All aspects of *NAEP* are described in great detail at http://www.nagb.org/toolbar/sitemap.html. Wading through all of the detail on the National Governing Board website, however, is a daunting task. The website is largely a repository for documenting the Board's legislative basis and bureaucratic operations. The site ostensibly provides different sections for parents, educators, policymakers, business leaders, and media, but whatever category is chosen, the menus are identical: *NAEP* Data and Resources, Assessment Schedule, News, Frequently Asked Questions, and Nation's Report Card Site, each with a different "outreach" document added for parents, policy makers, and business leaders. *The Nation's Report Card* web site http://www.nationsreportcard.gov is actually a stand-alone site and provides all the working information

The *Reading Framework* reflects the perspective of test construction rather than the perspective of reading instruction. Although it is easy to classify a text passage as "literary" or "informational," the test items cannot be so classified, and the "cognitive target" of individual test items is difficult to classify. The *Framework* categories are not categories that any ordinary reader would use in approaching a text or that a teacher would use in organizing instruction.

Table 1
NAEP Reading Test Framework

1992-2007
Contexts for Reading
Reading for Literary Experience
Reading for Information
Reading to Perform a Task
Aspects of Reading
Forming a General Understanding
Developing Interpretation
Making Reader/Text Connections
Examining Content and Structure
2009-2013
Content Area
Literary
Informational
Cognitive Target
Locate/Recall
Integrate/Interpret
Critique/Evaluate

The Score Scale

NAEP results are reported in terms of scale scores, so the scale is fundamental. The scale is a single scale that incorporates all grade levels: 4, 8 and 12. (The scale runs from 00-500, but the bottom and top 100 points on the Scale have no entries.) Because a scale score alone has no meaning, results are commonly interpreted in terms of proficiency levels mapped by school grade. When the proficiency levels are mapped at a given grade, they appear reasonable. But when the grades are mapped on the single *NAEP* scale, as in Table 2, they are anything but reasonable. That is, one would expect both the grade and proficiency/expertise scale values to progress linearly. However, Grade 4 Advanced ranks higher than Grade 12 Basic; Grade 8 Advanced is higher than Grade 12 Proficient (Table 2). Another way of referencing the Scale is via the "item mapping" provided in *NAEP* documentation.[3]

about the tests and test results. Again, because of the magnitude and detail, navigating and "comprehending" all of the information included on the site is a formidable task. The glossary alone includes several hundred terms and involves grappling with such matters as balanced incomplete block design, logistic regression model, serpentine sorting—and much more.

[3] http://nces.ed.gov/nationsreportcard/itemmaps/index.asp

Table 2
NAEP Grade 4 – Grade 12 Proficiency Scale

Score	Grade	Scale
208	Grade 4	Basic
238	Grade 4	Proficient
243	Grade 8	Basic
265	Grade 12	Basic
268	Grade 4	Advanced
281	Grade 8	Proficient
302	Grade 12	Proficient
323	Grade 8	Advanced
346	Grade 12	Advanced

The "maps" are in terms of representative test items reported by *NAEP* as "marker items" that define given proficiency level scores on the 0-500 scale. Table 3 shows marker items for the proficiency levels by grade for the 2009 and 2011 administrations. For a given scale score, what the item requires of a student and the Cognitive Target of the item is shown.

It is clear that the scale values of the items are determined by something other than the *Framework* designations. For example, why "Make an inference to recognize a character trait" would mark "Below Basic" proficiency and "Interpret a story to infer a character trait with support from the text" would mark "Advanced" proficiency at Grade 4 in 2011 is not clear, apart from the substance of the items themselves. If the grade identification, scale score, and proficiency level of test items were removed, it would be impossible to rank the marker items on the scale.

Table 3
Marker Items for the NAEP Reading Scale

2011	
Grade 4	
Lowest Score 180	
Below Basic	
185	Make an inference to recognize a character trait—Integrate/Interpret
Basic 208	
236	Locate and recognize a relevant detail in an expository text—Locate/Recall
Proficient 238	
247	Recognize the main purpose of an expository text—Integrate/Interpret
Advanced 268	
320	Interpret a story to infer a character trait with support from the text—Integrate/Interpret
Highest Score 340	
Grade 8	
Lowest Score 200	
Below Basic	
230	Recognize an implicit main idea of a story—Integrate/Interpret
Basic 243	
276	Recognize the main purpose of an informative article—Integrate/Interpret
Proficient 281	
204	Recognize the main purpose of an informative article—Integrate/Interpret
Advanced 323	
338	Evaluate the effectiveness of the beginning of an article and justify with text support—Critique/Evaluate
Highest Score 370	
2009	
Grade 4	
Lowest Score 170	
Below Basic	
187	Make a simple inference to recognize description of character's feeling—Locate/Recall
Basic 208	
220	Use information across text to infer and recognize character trait—Integrate/Interpret
Proficient 238	
251	Provide cross-text comparison of two characters' feelings—Integrate/Interpret
Advanced 268	
309	Use specific information to describe and explain a process—Integrate/Interpret
Highest Score 340	
Grade 8	
Lowest Score 180	
Below Basic	
239	Recognize causal relationship between facts in article—Locate/Recall
Basic 243	
257	Use information from an article to provide and support an opinion—Critique/Evaluate
Proficient 281	
286	Recognize meaning of word describing character's action—Integrate/Interpret
Advanced 323	
336	Describe event and explain causal relation in narrative poem—Integrate/Interpret
Highest Score 370	

Relative Difficulty/Readability of the Text Passages

An examinee's reading performance varies based on the difficulty of the texts the person is required to read. The standard means of assessing the difficulty encountered in reading a given text is via a "readability formula." One would expect a reading test to be structured in terms of texts of regularly increasing difficulty/readability.

An analysis of the sample text passages provided by *NAEP* for the two most recent years is shown in Table 4. The table shows the grade equivalent for each passage (e.g. For the "Daisy" passage in the Grade 4 test, 2.9 is Grade 2, 9 Months. "Daddy" is a Grade 6, 9 Months passage.). I used *ReadabilityStudio 2009*[4] software to obtain the difficulty/readability levels listed below. The readability values vary both within tests and across years. How then is the difficulty of the tests and the consistency of the scale maintained? Answer: By manipulating the stems (the "questions") of the test items (both multiple choice and constructed response) and the foils/distracters (choice options) of the multiple-choice questions and the complexity of the constructed response items. This practice is also the foundation for the racial/ethnic "Achievement Gaps," as explained in the following section.

Table 4
Difficulty/Readability of NAEP Reading Passages

Grade 4	Grade 8	Grade 12
	2011	
2.9 Daisy	7.0 Marian	6.8 E. B. White
6.2 Daddy		10.6 Women Vote
9.0 Tech Trash		
	2009	
9.0 Nutting	9.0 Alligator Poem	8.1 Open Window
6.4 Buzz	9.7 Alien	12.2 Rental Agreement
	10.9 Home on the Range	

[4] http://oleandersolutions.com/readabilitystudio.html

Inside Racial/Ethnic "Achievement Gaps"

The *NAEP* results that have garnered the greatest attention over the years are the "gaps" in the performance of Black and Hispanic Students when compared with White (and Asian)[5] students. The *NAEP Question Tool* permits investigation of the gaps at the item level. The percentage of students that responded to each of the four response options in multiple choice items and for the scoring categories of constructed response items is disaggregated by White, Black, and Hispanic students. Finding the "gaps" using the *Question Tool* requires several steps that are not immediately transparent due to the massive data involved. Appendix 1 at the end of the commentary provides directions for navigating with the *Tool*.

Table 5 shows the results for the multiple choice items accompanying one text passage in 2011 at Grade 4 and another set of items at Grade 8. The text passages and the items are presented in Appendix 2 as well as the statistics for constructed response scoring categories. To read Table 4, read across for each item. That is, for the first Grade 4 item, 3% Hispanic students chose option 1, compared with 2% of White students; 5% Hispanic and 3% White chose option 2; and 39% Hispanic vs. 34% White chose option 4. These small differences among the racial groups in their tendency to select the distracters/foils provide fissures for the larger differences in selecting the keyed option 3 (preceded by an asterisk): 52% Hispanic vs. 60% White.

It is obvious from inspection that item difficulty is a function of the wording of the test item not the substance of the text passage; the keyed response is always the most popular choice of each racial group, but the wording of the item, not the difficulty of the text passage is determinative. Appendix 2 indicates that the more remote the item and the choices are from the substance of the text per se, the greater the difficulty.

The largest differences in the array are the differences among items, and these differences hold consistently for racial groups. That is, what is a difficult item for Black students and for Hispanic students is regularly also difficult for White students—just not quite as difficult for White students. The popularity of distracters also differs, and the relative popularity also holds across racial groups. The "gap" in scale acores is produced by "fissures" in the choice of distracters among racial groups, which compound to "cracks" in the choice of keyed responses that magnify to "gaps" when the items are pooled into scaled scores.

The *NAEP Question Tool* provides access to the same information shown in Table 5 and Appendix 2 for passages and items from 1992 through 2011. The Question Tool also enables tracing "gaps" for Gender (Male, Female), Socioeconomic Status (School Lunch Program Eligibility), and School Location (City, Suburb, Town, Rural). The patterns evident in Table 5 and Appendix 2 are evident wherever one traces them with the *Tool*.

Although item statistics aren't publicly available for tests other than *NAEP*, national, international, and U.S. state sample tests are available online. A complete list is available in Appendix 3. Wherever one chooses to look, it is evident that the tests were cut from the same cloth.

[5] Other racial/ethnic categories are included in the *NAEP* database, but are not included here for reasons of simplicity.

Table 5
Ethnic "Gaps" in NAEP Test Item Difficulty (2011 Sample Items)

Grade 4—*Tough as Daisy*

Response Option	A			B			C			D		
Ethnic Group	W	B	H	W	B	H	W	B	H	W	B	H
Item No.												
1	2	4	3	3	6	5	*60	53	52	34	38	39
2	4	8	7	21	27	25	3	8	7	*72	57	60
3	*93	84	86	3	6	5	4	3	2	2	4	5
8	3	8	7	28	34	34	3	7	9	*65	50	48
9	2	6	4	*82	69	70	2	5	5	12	18	20
11	4	12	9	16	19	22	*51	36	34	29	34	35

Grade 8—*Daddy Day Care*

1	4	9	9	18	20	16	*69	59	61	8	12	13
2	4	8	7	21	27	25	3	8	7	*72	57	60
4	*48	42	44	7	10	17	18	18	16	26	21	22
6	2	5	4	9	17	19	*80	62	65	7	10	10
8	7	14	10	6	15	11	*84	63	71	4	7	8
10	14	22	21	*57	38	40	20	26	27	8	14	11

About an Alternative

Yes, there is an alternative orientation and methodology. I propose the following set of premises derived from logic and filtered experience analogous to the way the Stanford Evaluation Consortium theses were derived:

- Acquiring expertise in any matter other than "school" gets easier rather than more difficult.
- Instruction can build on the assets that children bring to school. It cannot eliminate deficits.
- Although children differ, with very few exceptions, children age 4, can speak in whole sentences and participate in everyday conversation. This communication capability constitutes the minimal prerequisites to begin teaching written language communication capability.
- Printed words are oral language written down, governed by an Alphabetic Code consisting of 26 letters and 40-some sounds (phonemes).
- One English Alphabetic Code fits all. The Code is very tolerant to dialect differences, but intolerant to spelling differences.
- When a child has been taught how to handle the Alphabetic Code, the child can interact with text with the same level of understanding were the text read to the child.
- A child with this communication expertise is an independent reader, enabled to expand the capability to other academic accomplishments.
- With the Internet, Information (Gleick, 2011) (on anything and everything) is available ad lib.
- Expertise in **searching** and **filtering are** the prerequisites for intelligent interaction with the Internet.

These nine propositions pertain specifically to the reform of reading testing, but they also provide a general prototype for guiding, monitoring, and confirming instructional accomplishments—toward reform of educational testing consistent with the aspirations of the Stanford Consortium.

The propositions are not hypothetical speculation. The government of the United Kingdom is currently operationalizing the propositions. The U.K. began administering an Alphabetic Code (Phonics) Screening Check to all children completing Year/Grade 1, starting in June 2012. The test consists of 20 real words and 20 pseudo-words, administered to children individually by regular classroom teachers in less than 10 minutes. Each pseudo-word is accompanied by a drawing of an "alien," and the children are told that the word is the alien's name, to give meaning to the unobtrusive exercise. (A sample form of the test is shown in Table 6.)

Table 6
Sample Screening Check

Section 1	Section 2
tox	voo
bim	jound
vap	terg
ulf	fape
geck	snemp
chorm	blurst
tord	spron
thazz	stroft
blan	day
steck	slide
hild	newt
quemp	phone
shin	blank
gang	trains
week	strap
chill	scribe
grit	rusty
start	finger
best	starling
hooks	dentist

Although unobtrusive for children and teachers, the test is carefully structured with items arranged to reflect increasing complexity of the Alphabetic Code. Psychometrically, the Alphabetic Code (Phonics) Check is a Guttman-like scale, analogous to the Snellen Eye Chart used to test visual acuity in issuing drivers' licenses. That is, any capable reader can read all 40-items without difficulty. Any deviation from that reflects a reading capability flaw. However, just as 20-40 vision rather than 20-20 is set as "pass" for a driver's license, a score of 32 has been set as "pass" for Year/Grade 1, with the expectation that full reading capability will be achieved by the end of the following Year/Grade 2.

The important consideration in comparing the Alphabetic Code (Phonics) Check with conventional standardized reading achievement tests is that the results do not rely on the comparative performance of other students. The measure is a test of the instruction a student has received, not of the student's deficits when compared with other children.[6]

The Check was pilot-tested in June 2011 in 300 schools.[7] The Check was administered for the first time to all Year/Grade students in June 2012, but only very preliminary results have as yet been reported.[8] The first information concerning the status of the UK Government's commitment

[6] The Specifications for the Check can be accessed at http://media.education.gov.uk/assets/files/pdf/a/phonics_assessment_framework_web_ready_final.pdf.
[7] Information about the pilot testing is available at http://media.education.gov.uk/assets/files/phonics%20screening%20check%202011%20pilot%20technical%20report.pdf
[8] http://www.education.gov.uk/rsgateway/DB/SFR/s001086/sfr21-2012.pdf

to teach all children how to read by the end of Year/Grade 2 will become available after the administration of the Check to Year/Grade 2 students who did not pass the Check in Year/Grade 1 However, the Pilot Study and 2012 results provide several relevant findings:

- The modal score in 2012 was 40, *the highest possible score on the Check*. This result would be impossible on a standardized reading test due to the way those tests are constructed. But it is the result we would expect on this assessment if all children have been taught how to read.
- 58% of Year/Grade 1 students met or exceeded the pass score of 32 in 2012. This percentage is up from 31% who passed the Check in 2011, but it still leaves a challenge for Year/Grade 2 reading instruction.
- The Pilot Study indicated that 75% of children tested were in schools that report, "We encourage children to use a range of cueing systems, such as context or picture cues, as well as phonics" with only 25% of schools reporting, "We always encourage children to use phonics." Although the UK Government has mandated phonics instruction, mandating does not insure implementation.
- English-language Learners did as well on the Check as native-English speaking children. There was no "bilingual gap" in the results.
- There was a "poverty gap" between students eligible for free school meals (44% pass) and other students (61% pass). However, there was no correlation between the percentage of free meal-eligible students in the 138 Local Educational Agencies reporting results on the 2012 Check and the percentage of students in the local educational agency (LEA) passing the Check (r=.028, calculated from reported Tabled data). The "gap" appears a function of the instruction the schools are delivering rather than from the inexorable effect of poverty.

The 2012 results to date have been reported only at national and LEA levels. However, shedding light on the instructional determinants of the results requires information at the school and classroom-within-school levels. It would be a straightforward matter to obtain this information from schools and teachers by questionnaire as was done in the 2011 Pilot Study and to disaggregate it by screening check performance in the same way that the conventional demographic categories of interest are commonly analyzed. For starters, matters such as the following could be investigated:

<u>School Level</u>

- The U.K. Government initiated a "matching funds" program for schools to obtain training for Year/Grade 1 teachers and instructional materials for classes. The extent and nature of participation of the school in this program can be investigated. Similarly,
- The degree to which all primary grade classes in the school use the same reading instructional program or whether this is left to the discretion of individual teachers
- Head Teacher/Principal's beliefs regarding reading instruction
- Nature of Head Teacher/Principal's supervision of teachers' reading instruction

<u>Class Level</u>

- "Matching funds" training received
- Instructional program(s) used in reading instruction

- Beliefs regarding reading instruction
- Extent of parent involvement in reading instruction

The analysis of this information need not involve complex statistical analysis. What one is looking for is "inter-occular significance"—effects that hit you between the eyes. Without such transparently reliable effects, there is nothing that can be operationalized to improve instruction.

Although the Alphabetic Code Test is being used as a screening check at Year/Grade 1 in the UK, the test can be used at any time to determine whether a student has been taught how to use the alphabetic code to understand written English in the same way that spoken English is understood. If a child can do this, no further instruction in reading per se is necessary. If the test shows that this capability has not yet been achieved, attention is directed to the instruction that will achieve this accomplishment rather than to the child's deficits.

The test constitutes a measure that cannot be taught to directly; that is, students could practice alternate forms of the measure until exhaustion without making any progress. What one has to teach is how to handle the alphabetic code to pass the test. That is a complicated but not intractable instructional task.

The alternative orientation I have sketched moves toward realizing the reform envisioned in the theses of the Stanford Evaluation Consortium. The orientation:

- is forward-looking rather than backward-looking
- relies on testing "small" changes in instruction to produce "large" reliable instructional effects
- is unobtrusive
- emphasizes external validity
- judges success by transparent operational improvement of instruction

Whether at the classroom, school, LEA, or national level, the reform provides grounded information regarding instructional status. The thing about schooling, though, is that a fresh cohort of students enters the system each year. "Fixes" can be made during their instruction that have a logical chance of improving the status. The population of students available is sufficiently large to confirm with randomly drawn sub-samples that the fixes reliably work as intended. Moreover, last year's students live on in the system as next year's students. "Fixes" in their instruction can also be implemented and tested via natural experimentation. Per the conclusion of the Stanford Evaluation Consortium: "The evaluator is an educator; success is to be judged by what others learn."

References

Cronbach, L. J., Ambron, S. R., Dornbusch, S. M., Hess, R. D., Hornik, R. T., Phillips, D. C., Walker, D. F. & Wainer, S. S. (1980). *Toward Reform of Program Evaluation.* San Francisco: Jossey-Bass.

Gleick, J. (2011). The *Information: A History, A Theory, A Flood.* New York: Pantheon.

National Commission on Excellence in Education. (1983). A nation at risk: The imperative for educational reform. An open letter to the American people. A report to the nation and the secretary of education. Washington, D.C.: Distributed by ERIC Clearinghouse.

Ravitch, D. (2011). The *Death and Life of the Great American School System: How Testing and Choice Are Undermining Education.* New York: Basic Books.

Appendix 1
Navigating the
***NAEP Questions Tool* to Access Reading Test Items**

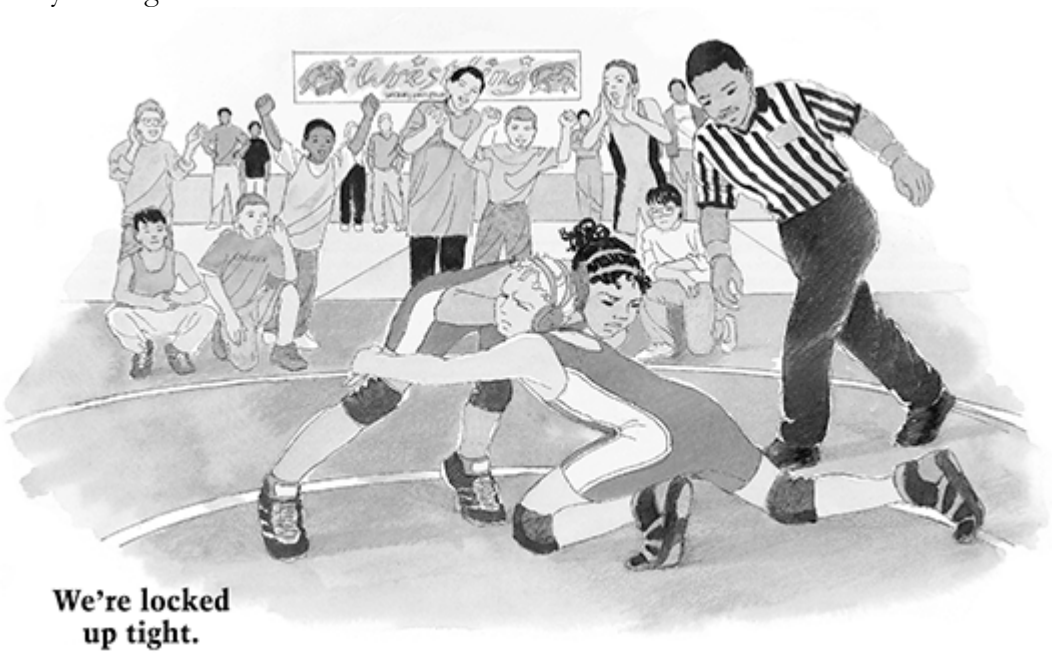
- Go to http://nces.ed.gov/nationsreportcard/itmrlsx/search.aspx?subject=reading
- From the left panel select *Year, Type, and Difficulty* of interest. From the right panel, click on an item of interest.
- Click on the green *View Item Detail* on the lower left under the list of items
- Click on *Show Reading Passage* to view the text
- Click on *National Data*
- Click on *More Data* in the lower right hand corner
- The blue menu allows you to examine item responses by:
    - Gender
    - Race/Ethnicity—White, Black, Hispanic, Asian/Pacific Islander, American Indian/Alaska Native, Two or More Races
    - School Lunch Eligibility
    - Type of Location—City, Suburb, Town, Rural

Appendix 2
NAEP 2011 Sample Passages, Items, and Difficulty Values

Grade 4
Tough as Daisy
by David M. Simon

The sign on the YMCA door says *Wrestling Tournament Today.*
I enter the gym and take a deep breath. It smells like old sweat socks and the stuff they use to wash wrestling mats.
I love that smell. Weird, huh? Not to me.
I was raised around wrestling. My older brothers wrestle for the high-school team. My dad wrestled in college. So it was natural for me to want to wrestle. Except for one thing.
I'm a girl. I even have a girly name—Daisy.
My dad always says, "Pound for pound, no one's as tough as Daisy."
I see my family in the stands. I wave to them and smile, but I'm nervous.
Lots of boys are already on the mats, loosening up. I'm the only girl at the sign-up desk. Some of the boys point at me and laugh. We'll see about that.
Back in Ohio, people got used to seeing me wrestle. I kept showing up. I kept winning. They stopped pointing and started cheering.
Then we moved to California. Now I'm weird again.
The man says, "Name?"
"Daisy McGill."
"Have you wrestled before, honey?"
He didn't call any of the boys *honey.* "Yes, sir," I answer through clenched teeth. I hand him my registration form.
"OK," he says. "Climb on the scale." I weigh 70 pounds. He writes a number on the back of my hand. I head to the girls' locker room to change.
First match. The kid looks strong. That's OK. Boys with muscles always underestimate me.
I snap the chin strap on my headgear. The ref calls us to the middle of the mat. We shake hands.
The kid says, "I can't believe I have to wrestle a girl."
The whistle blows, and I hit him fast with a fireman's carry. He's on his back in three seconds. The ref's hand slaps the mat. Pinned. One match down.
The kid refuses to shake my hand. The ref raises my right arm. He tells me, "Beautiful takedown!"
There's a lot of whispering going on. I hear someone say, "Man, she pinned him fast. No girl is going to beat me."
My family cheers wildly. I feel good. It always takes one match for the butterflies in my stomach to settle.
They call my number for the next match.
People crowd around the mat to get a look at Bizarro Wrestler Girl. Sounds like a good name for a superhero!
This kid is tall and thin. He looks serious about winning.
The whistle blows. I shoot for his leg. He kicks back and snaps my head down. He spins around behind me and takes me down. Good. I love a challenge.
Final period of this match, and I'm down three to nothing. Time to make my move.

I escape for one point, then shoot a quick takedown. All tied up. Thirty seconds to go. He raises one leg and I take a chance. I reach around his head and knee. My hands close tight. I roll him onto his back.
The whistle blows. The ref holds up two fingers. I win by two points. Two matches down.
At least this kid shakes my hand. Some of the people watching even clap for me.
I'm in the finals for my weight class.
My brothers rub my arms and joke around with me. Dad says, "Just do your best, honey." It's OK when*he* calls me *honey*.
I head for the mat. The next kid I'm wrestling pinned both of his opponents. There's a huge crowd watching us. I can't tell if they want me to win or lose.
Doesn't matter to me.
We shake hands. "You're pretty good," he says. "Good luck."
"You, too," I say.
The whistle blows. He shoots, and I'm on my knees before I can blink. Wow, he's fast. I feel my heart hammering in my chest. Easy, Daisy.
I spin away. Escape. He misses an arm-drag, and I catch him flat-footed. Takedown.
After two periods we're all tied up.
We're both gulping for breath as the last period starts. My brothers are screaming, but they sound far away. The kid shoots for my legs. I flatten out. He has one leg hooked. I force my forearm across his face like a wedge. We're locked up tight.
I can see the clock ticking down. With ten seconds left, his arms relax. Just what I was waiting for. I push down and spin behind him for the win. Yes!
I hear cheering and realize it's for me. The kid says, "Nice match. But next time, I'm going to win."
He just might.
My dad wraps my sweaty body in a big bear hug. He says, "Pound for pound, no one's as tough as Daisy."
I guess today he's right.



We're locked up tight.

1. What is the main problem Daisy faces in this story
   A. She has to make new friends at school.
   B. She has to perform in front of huge crowds.
   C. She has to prove that she is a good wrestler.
   D. She has to wrestle strong boys.

	A	**B**	**C***	**D**
White	2	3	60	34
Black	4	6	53	38
Hispanic	3	5	52	39

2. These paragraphs are from the first part of the story:
   I enter the gym and take a deep breath. It
   Smells like old sweat socks and the stuff they use
   To wash wrestling mats.

   I love that smell. Weird, huh? Not to me.

   What do these paragraphs help show about Daisy?
   A. She needs to learn how to wrestle.
   B. She enjoys different sports.
   C. She does not listen to other people.
   D. She enjoys being a wrestler.

	A	**B**	**C**	**D***
White	2	5	4	88
Black	6	9	7	77
Hispanic	7	8	8	75

3. According to the story, why was it natural for Daisy to be interested in wrestling?
   A. Her father and her brothers wrestled.
   B. Her coach at school encouraged her to wrestle.
   C. She had seen wrestling matches on television.

D. Many of her friends were on the wrestling team.

	A*	B	C	D
White	93	3	4	2
Black	84	6	3	4
Hispanic	86	5	2	5

4. At the beginning of the story, when some of the boys point and laugh at daisy, she thinks, "We'll see about that." What does this tell you about Daisy? (5 lines)

	Unacceptable	Acceptable
White	31	69
Black	40	58
Hispanic	43	56

5. How did the people in Ohio feel about Daisy when she wrestled? Support your answer with information from the story. (7 lines)

	Little or No Comprehension	Partial Comprehension	Full Comprehension	Omitted
White	17	39	43	2
Black	22	36	39	2
Hispanic	21	38	39	2

6. According to the story, why was the move to California so difficult for Daisy? (7 lines)

	Little or No Comprehension	Partial Comprehension	Full Comprehension	Omitted
White	16	46	36	2
Black	28	45	23	4
Hispanic	28	45	25	2

7. In the story, Daisy's father describes her as "tough." What are two other ways to describe Daisy's character? Support your answer with information from the story. (5 lines)

	Unsatisfactory	**Partial**	**Essential**	**Extensive**	**Omitted**
White	26	31	25	14	3
Black	35	35	18	7	5
Hispanic	39	31	19	7	4

8. On page 3, Daisy says that she answered the man at the registration desk "through clenched teeth." This means that Daisy

A. had trouble speaking correctly.
B. was nervous about joining the team
C. had hurt her teeth while she was wrestling.
D. closed her teeth tightly when she spoke.

	A	**B**	**C**	**D***
White	3	28	3	65
Black	8	34	7	50
Hispanic	7	34	9	48

9. On page 3, Daisy says that boys with muscles always **underestimated** her. This means that the boys

A. think Daisy is not very smart.
B. think that they can beat Daisy.
C. feel sorry for Daisy.
D. make fun of Daisy.

	A	**B***	**C**	**D**
White	2	82	2	12
Black	6	69	5	18
Hispanic	4	70	5	20

10. How is the first boy that Daisy wrestles different from the last boy she wrestles? Support your answer with information from the story. (7 lines)

	Little	**Partial**	**Full**	**Omit**
White	15	54	28	3
Black	28	49	17	5
Hispanic	28	49	17	5

11. What is the main way the author shows us how Daisy feels?

A. He uses pictures to tell her story.
B. He tells us what other people say about her.
C. He tells us what she is thinking.
D. He describes the way she wrestles.

	A	**B**	**C***	**D**
White	4	16	51	29
Black	12	19	36	34
Hispanic	9	22	34	35

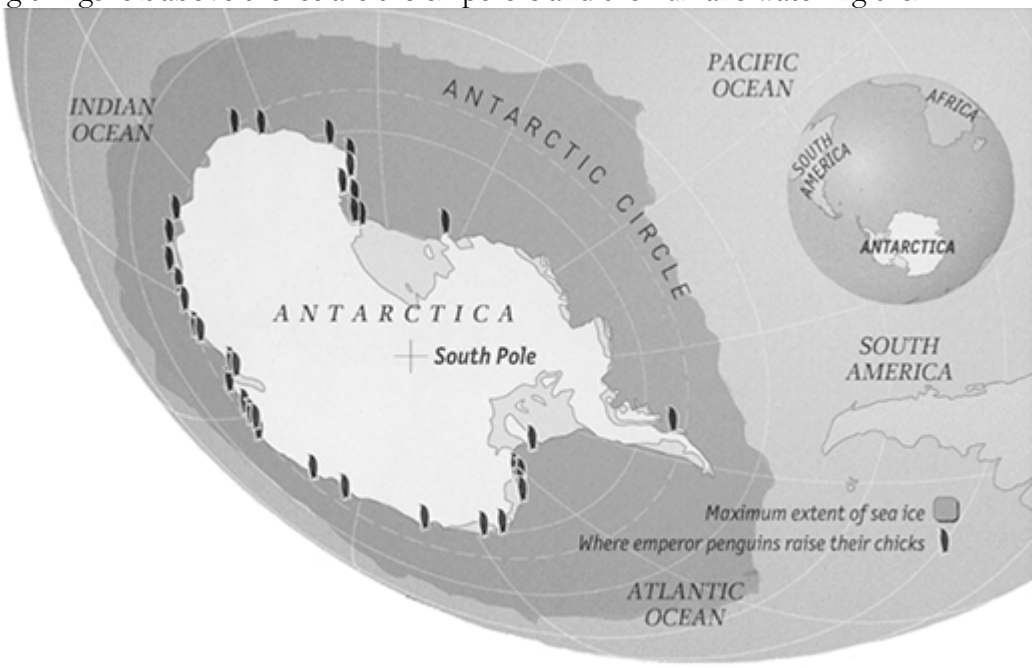
Grade 8
Daddy Day Care
Antarctica's ultimate stay-at-home dads
by Ruth Musgrave

When you think "tough," you may think of sharks, grizzly bears, or professional wrestlers, but you probably don't think of male penguins. Emperor penguins may not look it, but the males are tough enough to take on the deadly Antarctic winter and survive.

And they do it—without eating—while taking care of the eggs! When other animals head north in March to avoid the Antarctic winter, emperor penguins head south.

Antarctica is surrounded by a huge mass of sea ice in the winter. This ice floats on the ocean in the southernmost part of the Earth. Harsh and frigid, it is here where emperor penguins choose to mate and lay their eggs.

All the other animals, even other penguins, leave months before the Antarctic winter sets in. The only living things left above the ice are the emperors and the humans watching them.



© Martin S. Walz

Foothold for Family
At the breeding colony, all the males and females find mates. After courtship, the female lays one egg and gives it to her mate. Nesting in this barren, ice-covered world isn't a problem because emperors don't build nests. The male incubates the one-pound egg on his feet, covering it with a featherless fold of skin called a "brood patch."

Each male emperor penguin holds his egg throughout the brutal, Antarctic winter months of May and June. Nestled against a dad's warm, protective body, the softball-size egg remains untouched by the frozen world.

Meanwhile, the female travels to the sea to feed. She won't be back until just about the time the egg hatches—in about two months.

Warm-Up for Dads
The Antarctic weather wears on the male penguins with a viciousness that would seem unbearable to humans. Feathers, fat, and other adaptations are usually enough to keep adult penguins alive. But scientists who visit have to wear 22 pounds of clothing to stay warm!

"The penguins make it look so easy," says Gerald Kooyman, a biologist who has made more than 30 research trips to Antarctica. "After watching them awhile you almost forget how remarkable they are—until the weather changes and the wind slices right through you!"

One of the impressive ways emperors stay toasty when temperatures plummet or the wind blasts is to "huddle." A huddle forms when hundreds, even thousands, of males crowd together. The birds move constantly, slowly rotating from the cold outside rings to the warm, wind-free center.

One scientist who spent an entire winter observing these amazing birds says it is staggering to see 10,000 penguins in a single quiet huddle. The temperature inside can be 77°F. Standing nearby when a huddle breaks up, observers can feel, smell, even see the heat. It's like a wall of steam. The penguins are packed in so tightly that when one comes out, the bird is square-shaped for a few moments from the pressure of the other birds.

All for One
Not only is it unbelievably cold while the emperor dad stands holding his egg all winter, it's also dark. Nevertheless, he keeps the egg warm, without stopping for anything, even food. He loses up to a half of his body weight before his mate comes back from feeding at sea in July. She takes over the egg, which then hatches. The male finally gets to go eat. When he gets back, the parents take turns holding the chick on their feet to keep it warm for the next eight weeks. At that point it's old enough to safely stand on the ice by itself.

A newly hatched chick stays warm by standing on top of a parent's feet.

Snack Time

These older chicks gather together in large groups while their parents feed at sea. When adults return with food for their young, they locate their chicks by their calls. Emperors may look alike, but they don't sound alike. Each individual has a unique call that is recognized by other penguins.
Looking like toddlers in overstuffed snowsuits, hungry chicks scurry to parents returning from sea. As they race toward the adults—and dinner—they chirp, letting their parents know "I'm over here!"

Older chicks gather together to stay warm while their parents find food.

Independence Day

By the time the chicks are finally ready to fend for themselves, it's December. This is summertime in the Antarctic. During the winter, the nearest open water could be 50 miles from the rookery. In summer, the ice that the chicks hatched on has begun to break up, so the chicks don't have far to go to the sea.

The chicks are on their own now. The adults leave to start the cycle again, so the young emperors must learn to swim and find food by themselves. Winter day care is over; it's time for summer independence!

1. What is the main purpose of the article?

A. To describe why older chicks stand together in groups.
B. To help people understand what winter in the Antarctic is really like.
C. To describe what male emperor penguins do to take care of their young.
D. To explain why emperor penguins travel south in the winter.

	A	**B**	**C***	**D**
White	4	18	69	8
Black	9	20	59	12
Hispanic	9	16	61	13

2. According to the article, what is the main way a male emperor penguin protects its eggs from the cold?

A. By growing extra feathers.
B. By gathering together with other penguins.
C. By building a nest for the egg in the snow.
D. By covering the egg with a flap of skin.

	A	**B**	**C**	**D***
White	4	21	3	72
Black	8	27	8	57
Hispanic	7	25	7	60

3. The article describes the male emperor penguins as "tough." Give two pieces of information from the article that show that male emperor penguins are tough. (7 lines)

	Little	**Partial**	**Full**	**Omitted**
White	11	30	58	2
Black	20	34	41	4
Hispanic	21	36	38	5

4. On page 4, the article says that male emperor penguins live in a **barren** world. This suggests that the penguins live in a place

A. where almost nothing grows.
B. few other penguins go.
C. there is a lot of danger,
D. It is dark most of the year.

	A*	**B**	**C**	**D**
White	48	7	18	26
Black	42	10	18	21
Hispanic	44	17	16	22

5. Explain how emperor penguins stay warm when they form huddles. (7 lines)

	Little	Partial	Full	Omitted
White	28	37	34	2
Black	54	28	15	4
Hispanic	56	26	14	3

6. On page 5, the article says that one scientist found it **staggering** to see 10,000 penguins in a single quiet huddle. This means that the scientist

A. thought that the penguins walked in a funny way.
B. doubted that penguins could survive in groups.
C. was amazed that so many penguins could gather in this way.
D. was confused because the penguins were so quiet.

	A	B	C*	D
White	2	9	80	7
Black	5	17	62	10
Hispanic	4	19	65	10

7. Describe the roles that male and female emperor penguins play in hatching and raising their young. Give information about the roles of both male and female penguins in your answers. (15 lines)

	Unsatisfactory	Partial	Essential	Extensive
White	9	17	61	8
Black	20	23	44	4
Hispanic	21	20	45	4

8. According to the article, how do adult emperor penguins returning from the sea find their own chicks to feed them?

A. They can smell their chicks.
B. The chicks wait in their nests.
C. Each chick sounds different.
D. Each chick looks different.

	A	**B**	**C***	**D**
White	7	6	84	4
Black	14	15	63	7
Hispanic	10	11	71	8

9. Why does the author include the map on page 3? (5 lines)

	Unacceptable	**Acceptable**	**Omitted**
White	17	81	2
Black	24	71	4
Hispanic	21	74	4

10. According to the article, why is summer in Antarctica a good time for the chicks to become Independent?

A. There are no animals around that could hurt the chicks.
B. The sea is not far away in the summer.
C. Both parents can be there to help their chicks.
D. It is easier to build nests in summer.

	A	**B***	**C**	**D**
White	14	57	20	8
Black	22	38	26	14
Hispanic	21	40	27	11

Appendix 3
Additional Testing Resources

Australia National Assessment Program, Literacy and Numeracy –Reading (Year 3, 5, 7, 9)
http://www.nap.edu.au/naplan/the-tests.html

United Kingdom Key Stage 2 (Year/Grade 6) Reading Test
http://www.emaths.co.uk/KS2SAT.htm#English

OECD Program for International Student Assessment (**PISA**)
http://pisa-sq.acer.edu.au/showQuestion.php?testId=2292&questionId=1

International Association for the Evaluation of Educational Achievement (IEA) Progress in International Reading Literacy Study (PIRLS)
http://nces.ed.gov/pubs2004/pirlspub/14.asp

(For information comparing *NAEP* and PIRLS)
http://nces.ed.gov/pubs2004/pirlspub/pdf/2003073_2.pdf

UNESCO Second Regional Comparative and Explanatory Study (SERCE): Student Achievement in Latin America and the Caribbean
http://unesdoc.unesco.org/images/0016/001610/161045e.pdf pp. 15-17

Representative US State Tests
California Standards Test—English Language Art (Grade 2-11)
http://www.cde.ca.gov/ta/tg/sr/css05rtq.asp

Massachusetts Comprehensive Assessment System—Reading Comprehension (Grade 3-8 & 10)
http://www.doe.mass.edu/mcas/2011/release/g4ela.pdf

About the Author
Author Dick Schutz
Affiliation information 3RsPlus, Inc.

Email: 3RsPlus@usinter.net
Biographical information in this paragraph.
Dick Schutz is CEO of 3RsPlus, Inc. a firm conducting R&D and constructing educational products. He was formerly Professor of Educational Psychology at Arizona State University and Executive Director of the Southwest Regional Laboratory for Educational Research and Development. He has served as the founding editor of the Journal of Educational Measurement, the founding journal editor of the Educational Researcher, and editor of the American Educational Research Journal. His recent technical papers can be accessed from the Social Science Research Network http://ssrn.com/author=1199505

education policy analysis archives

Volume 21 Number 90 December 20th, 2013 ISSN 1068-2341