The Listening and Writing Skills Test (LAWS):

Rationale, Field Testing, Reliability and Validity Data for a High-Structure Assessment of Writing Skills

Presenter: Bruce Davey

What Happens If Police Officers Can't Write Well?

- Bad police reports
- Bad communication to public
- Incidents described inaccurately
- Investigative leads lost
- Cases lost, charges dropped
- Embarrassment in Court
- **Legal liability**

Authentic Assessment of Writing

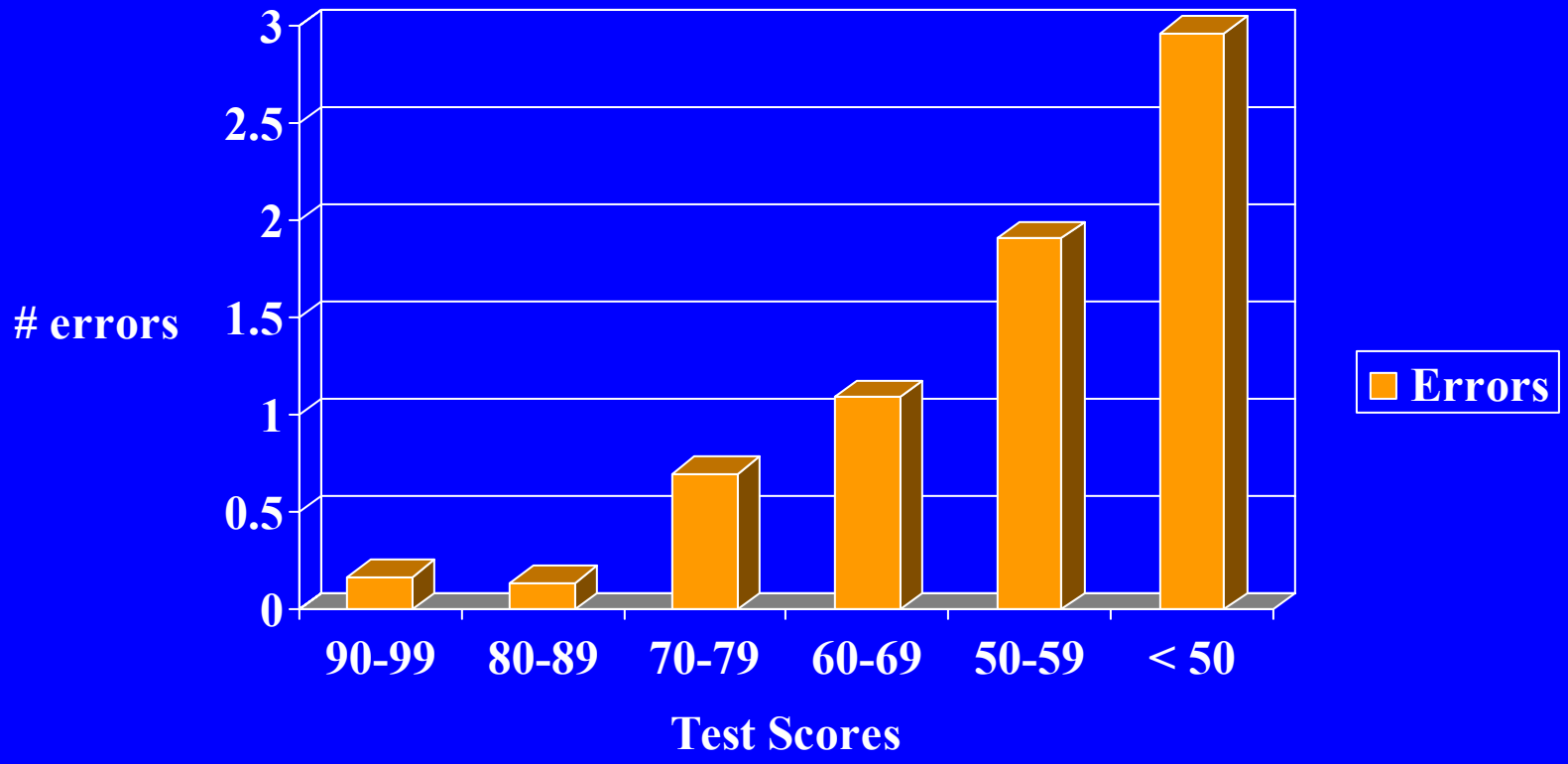
- High content validity
- Relevance is easily established

-- but --

- Problems of reliability/subjectivity
- Requires multiple raters
- Requires about 40 minutes/paper to score

“I have not discussed and will not discuss the content of this test with anyone from another test session.”

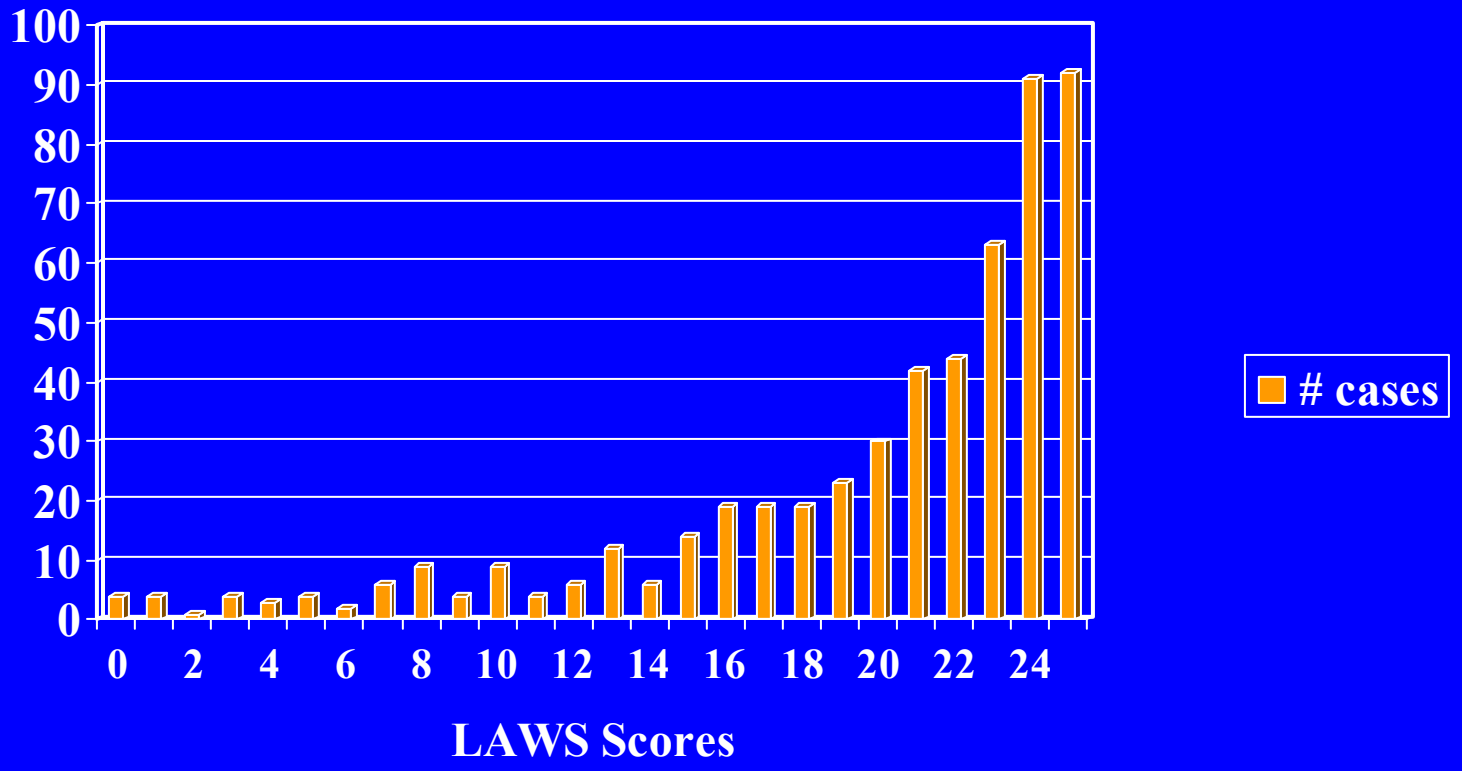
Errors in Declaration Sentence Vs. Written Test Scores



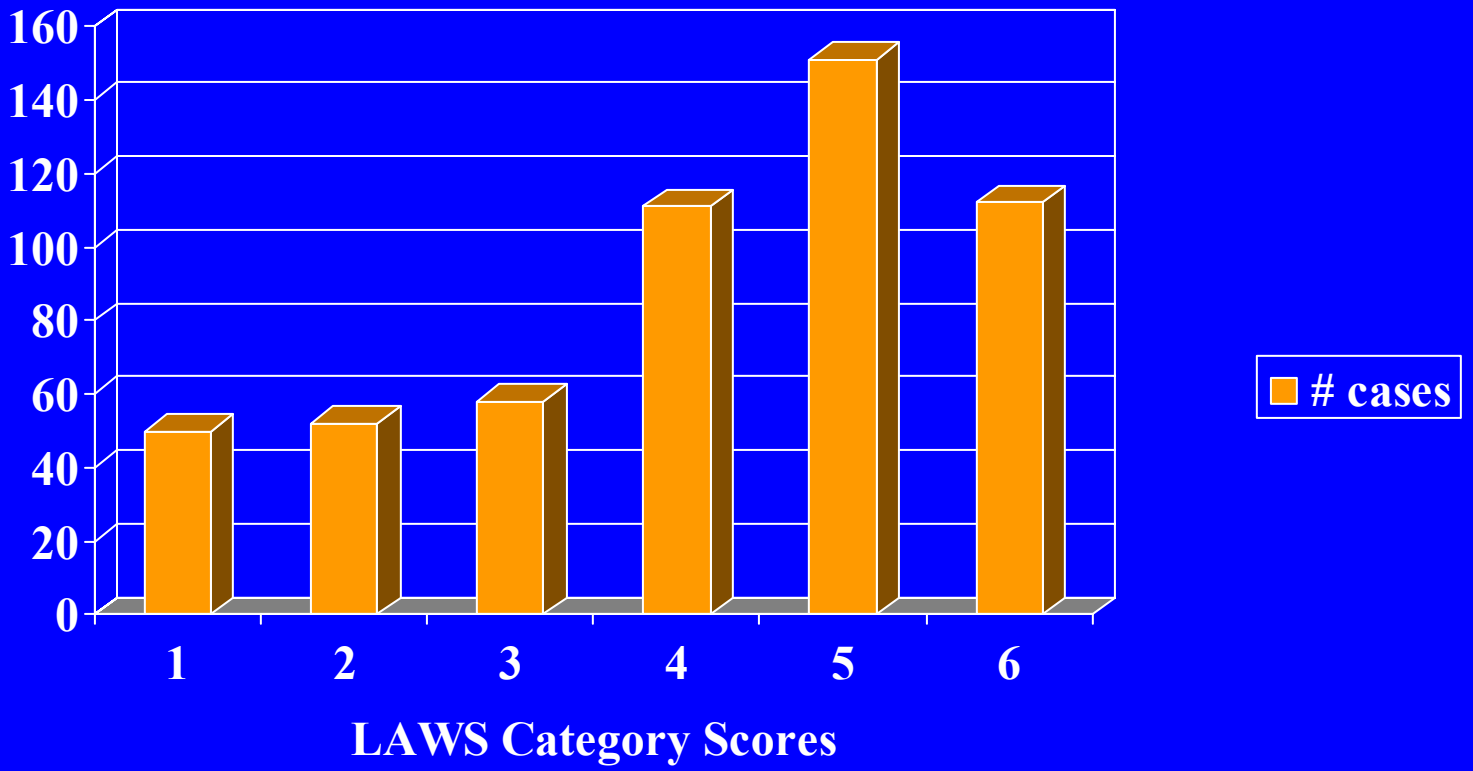
LAWS Language Level

	Form 1	Form 2
• # Sentences	5	5
• # Words	102	94
• Flesch		
Reading Ease	74	68
• Grade Level	7.9	8.4

Frequency Distribution: LAWS Raw Scores



Frequency Distribution: LAWS Category Scores



LAWS Reliability Data

Type	Reliability:	N
Inter-rater	.96-.99	481
Internal consistency	.91	534
Test-Retest	.79	149
Alternate Form	.77	131

Scoring Time Per Paper

	Authentic	LAWS
• # Raters needed	2	1 or 2
• Scoring time per paper	15 min	3.5 min
• Discuss time per paper	5 min.	1 min
• Total time per paper	40 min	9 min (4 min if one rater)

LAWS Construct Validity Data

Correlations With External Measures:

	r	N
Cognitive Ability MC Test	.56**	534
Educational Level	.29**	510
Self Assessment of Writing	.37**	512
Oral Examination Score	.31**	158
CIS	.41**	158

Writing Ability Self-Rating

104 **WRITING ABILITY**
Ability to write clear, accurate and understandable reports and other communications.

(A) below average
(B) about average
(C) above average
(D) very high (top 10%)
(E) outstanding (top 1-2%)

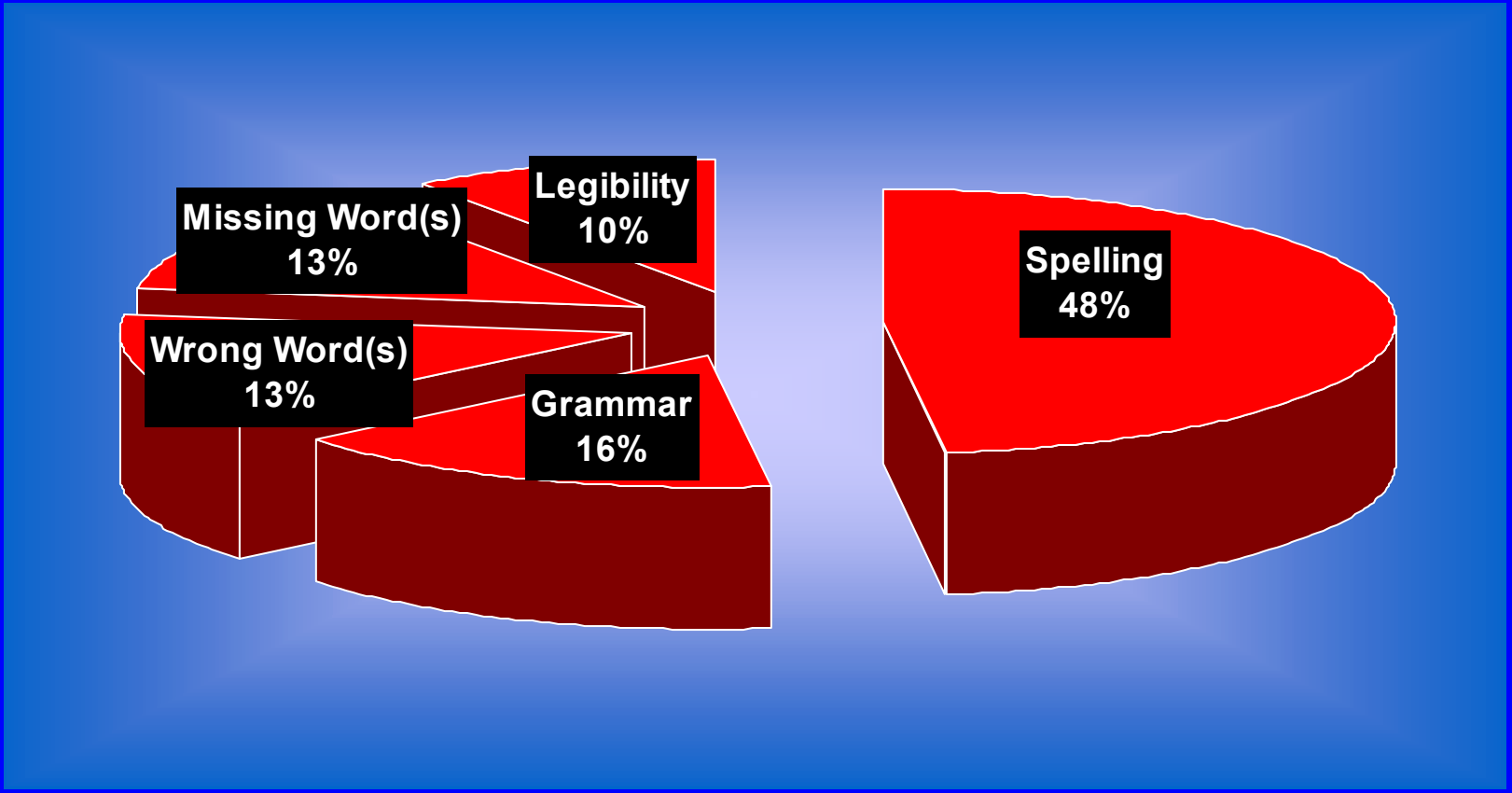
Note: This rating is one among 28 made about 30 minutes before the test battery is administered.

r with LAWS = .37 corrected r = .51 (N=512)

Correlation Between Laws and Authentic Writing Assessment

Writing Exercise Score for:	r	N
Theme Organization	.38*	38
Clarity of Communication	.28	38
Writing Mechanics	.61**	38
Overall Writing Score	.51**	38

Types of Errors Made on LAWS



Effects of Retesting

	N	r between testings:	Improvement:
• First Testing to second	103	.79	+ .12 SDs
• Second Testing to Third	36	.75	+ .42 SDs
• Third Testing to Fourth	10	.85	- .48 SDs
• Average Practice Effect	149	.79	+ .14 SDs

LAWS Results by Race & Gender

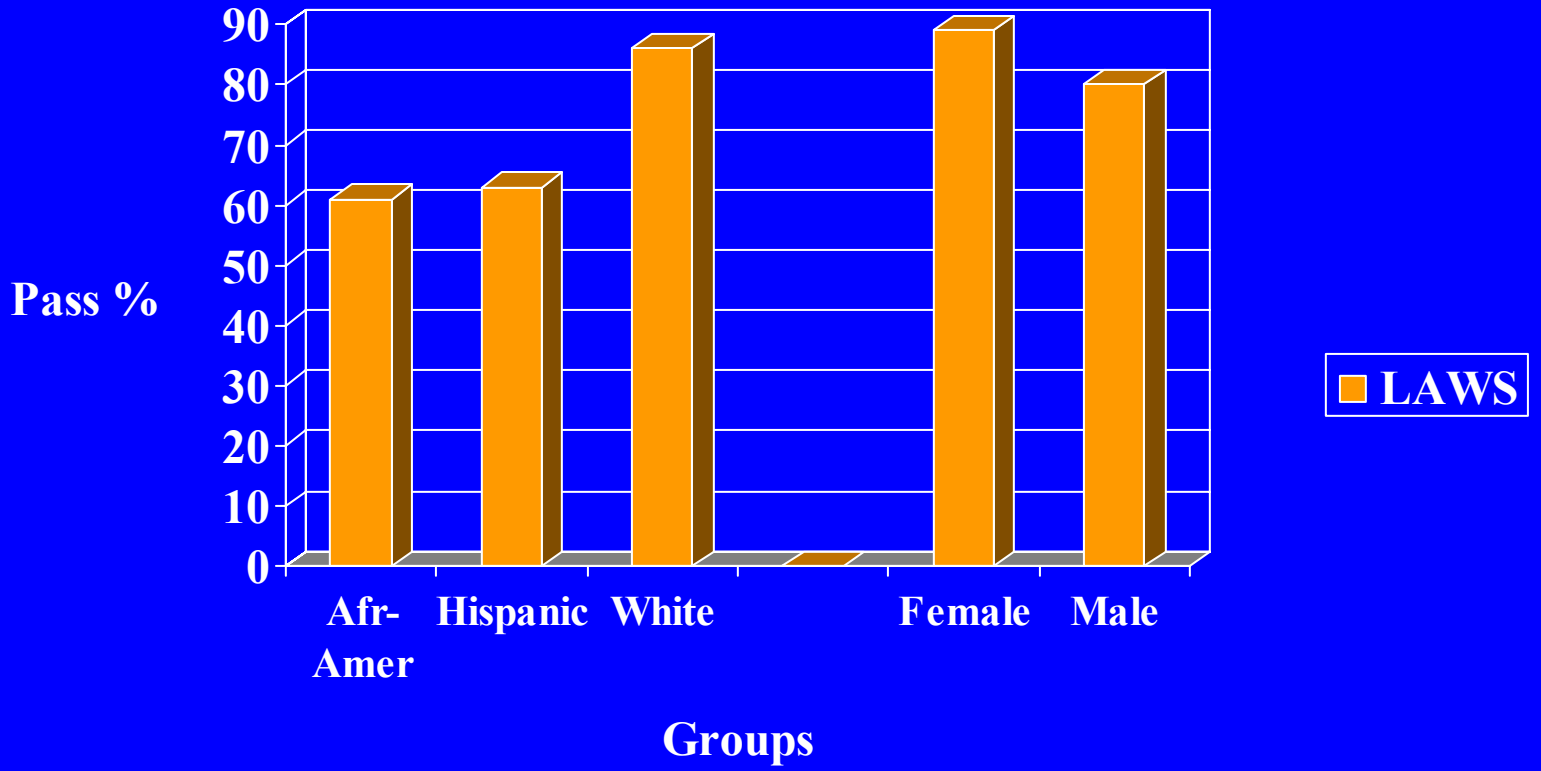
- Cognitive Ability Test

Group	Diff	N
White		412
Afr-Amer	-.80 sd	41
Hispanic	-.91 sd	65
Male		460
Female	-.04 sd	70

- Laws Category Scores

Group	Diff	N
White		412
Afr-Amer	-.56 sd	41
Hispanic	-.68 sd	65
Male		460
Female	+.27 sd	70

LAWS "Acceptable Rates" By Group



LAWS CATEGORY DEFINITIONS
Acceptable Categories

Excellent

No errors made in writing exercise. Able to listen well and convert what he or she hears to accurate, correct written narrative. Excellent spelling, grammar, clarity and sentence formation.

Very Good

Did very well on the writing exercise. Able to listen well and convert what he or she hears to correct written narrative with few errors. No evident problems in spelling, grammar, clarity and sentence formation.

Good

A few minor errors on the writing exercise. Generally able to listen well and convert what he or she hears to correct written narrative, making occasional errors in spelling, grammar, clarity or sentence formation. Based on this test, appears to be in the average range of writing ability.

Acceptable

Made several errors on the writing exercise, but overall performance on spelling, grammar, clarity and sentence formation was acceptable. Makes some errors in producing written narrative based on what he or she hears, and shows room for improvement in writing skills, but is in the acceptable range.

LAWS CATEGORY DEFINITIONS
Less Than Acceptable Categories

Marginal

Made a number of errors on the writing exercise, so that overall performance was questionable. May have difficulty producing clear, correct written narrative based on what he or she hears. Further assessment is advised to determine whether writing skills meet the department's standards.

Unacceptable

Was not able to write clear and acceptable, problem-free sentences based on what he or she heard. Tends to make a variety of errors in producing written narrative. Based on these results, appears to need considerable remediation before writing skills can be considered acceptable.

THE LISTENING AND WRITING SKILLS (LAWS) TEST: Rationale, Field Testing, Reliability and Validity Data for a High-Structure Assessment of Writing Skills

Presentation at the 2003 IPMAAC Annual Conference

Presenter:

Bruce Davey, Director
Bruce Davey Associates
PO Box 1333
Glastonbury, CT 06033

June 2003

THE LISTENING AND WRITING SKILLS (LAWS) TEST: Rationale, Field Testing, Reliability and Validity Data for a High-Structure Assessment of Writing Skills

This report describes a highly structured approach to the assessment of basic writing skills. The Listening and Writing Skills (LAWS) test was developed in 2002 and is currently undergoing extensive field testing and analysis. This report presents its rationale, describes its format, and summarizes the results of the first 9 LAWS administrations, including normative data, reliability and construct validity data.

LAWS Rationale and Format

The Listening and Writing Skills (LAWS) test is an assessment of basic writing skills. It was developed by Bruce Davey Associates (BDA) in response to problems seen with traditional methods of assessing basic writing skills. Two traditional methods are:

(a) Scoring of a writing exercise in response to a prompt such as "Tell us why you want this job" or "Write about a recent event in which you had to help someone with a problem"; and

(b) Multiple-choice assessment of writing ability, using items designed to assess whether test takers can detect writing errors in sentences presented, or select the correct or most appropriate word usage or language usage from among choices presented, or demonstrate their vocabulary.

Approach (a), sometimes called "authentic writing assessment," has the virtue of high content-relevance -- test takers are required to produce actual writing samples. However, scoring written narrative is difficult and time-consuming. Because this approach often poses problems of reliability, it generally requires scoring by two or more persons followed by consensus discussions in order to achieve reliable results. Moreover, even with multiple raters there are often reliability problems encountered with these exercises. Because each paper to be scored is quite different in terms of its content, style, length, word usage, etc., scoring is difficult and depends heavily on rater judgment.

Typically it requires approximately 30 minutes of rater time to score a 200-word paper (10-15 minutes per rater plus a 5-minute consensus discussion period for each paper). This extent of time commitment raises questions about the cost-effectiveness of this approach. Finally, the themes and ideas presented in such writing samples often introduce error variance due to content interference. Even if raters are instructed to score only writing mechanics, theme content can interfere with scoring nonetheless. For example, a test taker with good ideas may be scored higher on their writing ability than warranted because of "halo effect" while a test taker with well-written but objectionable or uninspired ideas may be unduly penalized.

Approach (b) is far more efficient, since multiple choice writing exercises can be machine-scored, and a large number of items can be very efficiently administered and scored. It is also possible to produce reliable subscores which may be useful for diagnostic purposes (vocabulary, grammar, word usage, etc.). However, a complaint often registered about this approach is its low fidelity. Clearly the multiple choice approach does not have the content-relevance of an authentic writing sample. For example, far more inference would be needed to state that a score of 63% on a multiple choice test of written communication skills is an accurate indicator of a writing skills deficit.

The LAWS test attempts to combine some of the strengths of authentic writing assessment with the greater standardization that accompanies a multiple-choice assessment. Instead of presenting a writing theme or prompt requiring one to compose a narrative which will differ markedly from one test taker to the next, the LAWS test presents a series of standardized sentences for test takers to listen to and then reproduce. Scoring a LAWS exercise is analogous to applying an objective scoring template

to the writing produced. Since the test taker's task is to correctly produce a series of standardized sentences, word for word, there are right and wrong answers, and their accuracy can be objectively and reliably assessed.

In the current research versions of LAWS (Forms 1, 2 and 3), there are five sentences presented. Each sentence is approximately 20 words long, with rather basic vocabulary and complexity (eighth grade level -- see Figure 1 below). The sentences are presented via high-quality digital recording. Each sentence is recited slowly, and repeated twice after the first recitation. Test takers are instructed to try to write as much of the sentence as they can the first time they hear it, and to fill in missing parts on the second and third recitations. In scoring, deductions are made for spelling errors, grammar errors, wrong words or phrases, missing words or phrases, and illegible words or phrases.

The purpose of repeating each sentence three times is to reduce the extent to which the LAWS test is dependent on "listening skills" or pure auditory memory, and to increase the focus on writing skills. However, listening is believed to interact with verbal comprehension in LAWS assessments in a way which is similar to what occurs in CLOZE assessments. Because one must listen to and write down a sentence which is approximately 20 words long, some reconstruction of what was heard is necessary to accomplish the task. The test taker is likely to retain most of the sentence, but not all of it. Just as in CLOZE assessments, the context of the sentence provides cues to any "missing" parts, or parts the test taker cannot quite recall. A person who has strong verbal skills will be able to "fill in the blanks" with greater accuracy than a person whose reconstruction of the sentence is not aided by strong language development. Even though the sentence is repeated three times, it appears to be difficult for those with poor language skills to accurately reconstruct all that was heard. The most common errors involve spelling, but surprisingly, more than half the errors involve grammar, wrong words, missed words and illegible words.

Figure 1 Grade Level of Sentences used in LAWS Form 1 and Form 2

	LAWS Form 1	LAWS Form 2
Number of Sentences	5	5
Number of Words	102	94
Number of characters	444	425
Flesch Reading Ease	74.1	68.0
Flesch-Kinkaid Grade Level	7.9	8.4

LAWS Scoring Time

Scoring of the five LAWS sentences during the tryout phase has averaged approximately 3-4 minutes per paper. The scoring would be considerably shorter (perhaps two minutes per paper) if the raters were merely counting errors; however, during the research phase they are required to not only tally the errors but place them in one of five categories (spelling errors, grammatical errors, wrong words or phrases, missing words or phrases, illegible words or phrases). A more basic error-counting approach would be likely to require less than half of this time, but diagnostic information would be lost. With this categorization approach, scoring is still somewhat time-intensive (e.g., scoring 100 papers would require about six hours for a single rater to score). However, this is far less than the time that would be required to score a typical writing theme. BDA's experience has been that scoring a 150-250 word writing sample requires about 15 minutes per paper, and clearly requires at least two raters, and either a discussion or a third rater to resolve differences. Scoring the LAWS test requires about 3.5 minutes per paper for a single rater. This is at least a 400% improvement in scoring time, accompanied by greatly increased scoring objectivity. Because of the greatly increased reliability (see Reliability section of this report), scoring could be done by a single rater, which reduces scoring time to approximately one-eighth that of traditional writing assessments. Therefore, while LAWS scoring is

still a time-intensive activity, it allows the reliable scoring of basic writing skills within a time frame more proportionate to the value of the exercise.

LAWS Tryout and Results

The LAWS test was administered nine times between May 2002 and May 2003. Each of these LAWS administrations took place in Connecticut and preceded the administration of BDA's Municipal Police Selection Test. The 9 town governments for whom the LAWS was administered appear in Figure 2.

Figure 2 LAWS Administrations

Town	Date	Form	# Tested
Manchester, CT	05/18/02	1	44
Vernon, CT	06/08/02	1	38
Bristol, CT	06/15/02	1	117
South Windsor, CT	07/13/02	2	56
Bloomfield, CT	08/24/02	2	94
Meriden, CT	10/12/02	2	170
CT Police Corps	3/22/03	1	19
Middletown, CT	5/03/03	1	94
Brookfield, CT	6/07/03	3	55
Form 1 (5 administrations)			312
Form 2 (3 administrations)			320
Form 3 (1 administration)			55
All administrations			687

While Figure 2 shows that there have been 687 LAWS tests completed, a number of people have taken the test more than once. The 687 administrations above involve 534 different persons. The distribution of single and multiple testings is as follows:

Group A.	Persons taking the LAWS test for the first time:	534
Group B.	Persons from above who took LAWS a second time:	103
Group C.	Persons from above who took LAWS a third time:	36
Group D.	Persons from above who took LAWS a fourth time:	10
Group E.	Persons from above who took LAWS a fifth time:	4

Those who were tested more than once were persons who applied for police officer positions in more than one of the cities or towns shown in Figure 2.

Most of the analyses described in this report focus on Group A above, e.g., those taking the LAWS for the first time. This was done for all analyses except those designed to compare performance across testings, such as retest reliability and score improvement analyses. The "first occasion only" rule was invoked to avoid artificially inflating the N by counting some individuals more than once in the analysis. However, analyses focusing on retest reliability and score improvement of necessity involved analysis of the scores of Groups B-D above.

Forms Equivalence

Means and standard deviations were computed for Forms 1, 2 and 3 of LAWS, to determine whether data could be meaningfully combined across the three forms. Basic statistics for the three forms (Figure 3) showed no significant differences in means or variances for either raw scores or category scores (NOTE: category scores are explained in the next section). It should be noted that Form 3 is newly developed and data for it is currently based on a rather small N. However, in its limited sample, its score mean and standard deviation was quite similar to Forms 1 and 2.

Figure 3 Means and Standard Deviations of LAWS Form 1 and 2

	Raw LAWS Scores		Category Scores		
	Mean	SD	Mean	SD	N
Form 1	20.23	5.26	4.14	1.50	262
Form 2	19.74	6.29	4.05	1.66	220
Form 3	21.04	4.55	4.31	1.46	52
All Forms	20.11	5.65	4.12	1.57	534
Form 1-2 Diff	0.49	(.09 SD)	0.09	(.06 SD)	
T test	0.93	(ns)	0.63	(ns)	
F test	1.43	(ns)	1.22	(ns)	

Since there no significant differences in means or variances for the three test forms, data were combined across forms for the analyses which follow, except where specifically noted.

Score Distribution

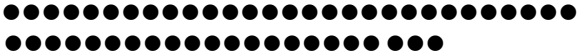
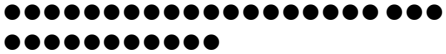
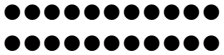
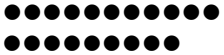
The LAWS score distribution is markedly skewed. On these five-sentence test forms, 17% of all test takers achieved a perfect score of 25, meaning no errors made in writing the five sentences. The median score was 22, corresponding to 3 errors made; but 8% of those tested made 15 or more errors. The frequency distribution appears in Figure 4, and the degree of truncation is obvious. While the high end of the distribution raises a question about whether the test should be made more difficult, the lower end of the distribution suggests that the test is already beyond the grasp of many in the police candidate pool, 8% of whom made three or more errors per sentence.

Figure 4 Raw Score Distribution of LAWS Forms 1 and 2

		N	%		Cumulative %
25	(O errors)	92	17.2%	●●●●●●●●●●●●●●●●●	17.2%
24	(1 error)	91	17.0%	●●●●●●●●●●●●●●●●●	34.2%
23	(2 errors)	63	11.8%	●●●●●●●●●●●●	46.0%
22	(3 errors)	44	8.2%	●●●●●●●●	54.2%
21	(4 errors)	42	7.9%	●●●●●●●●	62.1%
20	(5 errors)	30	5.6%	●●●●●●	67.8%
19	(6 errors)	23	4.3%	●●●●	72.1%
18	(7 errors)	19	3.6%	●●●●	75.7%
17	(8 errors)	19	3.6%	●●●●	79.3%
16	(9 errors)	19	3.6%	●●●●	82.9%
15	(10 errors)	14	2.6%	●●●	85.5%
14	(11 errors)	6	1.1%	●	86.6%
13	(12 errors)	12	2.4%	●●	89.0%
12	(13 errors)	6	1.1%	●	90.1%
11	(14 errors)	4	0.7%	●	90.8%
10	(15 errors)	9	1.7%	●●	92.5%
9	(16 errors)	4	0.7%	●	93.2%
8	(17 errors)	9	1.7%	●●	94.9%
7	(18 errors)	6	1.1%	●	96.0%
6	(19 errors)	2	0.4%		96.4%
5	(20 errors)	4	0.7%	●	97.1%
4	(21 errors)	3	0.6%	●	97.7%
3	(22 errors)	4	0.7%	●	98.4%
2	(23 errors)	1	0.2%	●	98.6%
1	(24 errors)	4	0.7%	●	99.3%
0	(25 errors)	4	0.7%	●	100.0%

Because of the shape of the raw score distribution, a score transformation was considered necessary to map scores onto a distribution approaching normality. Clearly, if a transformation had not been done, analyses of raw scores and their relationships to other variables would have yielded results which were overly influenced by extreme cases at the tail of the distribution. Since it was important to analyze the predictive effectiveness of scores across the full range of the distribution, score categories were created which produced a more normal distribution, as shown in Figure 5.

Figure 5 Distribution of Category Scores for LAWS Forms 1 and 2

Category:	# errors	N	%		Cumulative %
6	0	92	17.2%		17.2%
5	1,2	154	28.8%		46.0%
4	3-5	116	21.7%		67.7%
3	6-8	61	11.4%		79.1%
2	9-14	61	11.4%		90.5%
1	>14	51	9.6%		100.0%

While still possessing some degree of skew, the six-category forced distribution was sufficiently similar to a bell-shaped normal distribution to allow meaningful correlational analyses.

Reliability

The reliability data are extremely impressive in light of the fact that the LAWS is currently a five-sentence test which requires approximately 8-10 minutes to administer from start to finish. Reliability figures appear below.

Figure 6 LAWS Reliability Measures

Reliability Measure	Obtained reliability	N	
Internal consistency (coefficient alpha)	.91	534	
Interrater reliability (independent ratings)	.96	418	All papers scored by two raters without discussion to resolve differences
Interrater reliability (after consensus discussion)	.99	63	
Retest reliability (raw scores)	.88	149	Correlation between consecutive testings (note: too heavily influenced by lower scores)
Retest reliability (category scores)	.79	149	Correlation between consecutive testings
Alternate form reliability (category scores)	.77	131	Correlation between Form 1 and Form 2

The internal consistency reliability figure, based on Cronbach's alpha applied to the combined score of the five items, was .91. This indicates that the construct or constructs being measured by the LAWS test are quite stable; the laws sentences show high intercorrelations. Figure 8 provides information on the difficulty level and intercorrelations of the LAWS sentences for Form 1 and Form 2 separately.

The interrater reliability reported is based on independent ratings without discussion. For two of the test's administrations, the raters discussed and reconciled their ratings, and for those cases the interrater r was .992. The intercorrelation between independent raters was .96; combining the two raters' scores produces a score which of course has a higher reliability (.98).

Retest reliability data are based on the correlation between first and second testings in all cases where such data were available. Using raw scores, the retest reliability is .887 but this is clearly inflated by the impact of skewed-low scores, some of which were 4 standard deviations from the mean. Retest reliability based on category scores (.79) is the more appropriate figure to use.

Inter-rater reliability by LAWS sub-category is also high, especially in light of the low variance of the category scores. Figure 7 shows that most of the LAWS subscores were able to be scored reliably. The only sub-category that appeared difficult to rate reliably was Legibility, a category which required raters to make a relatively subjective judgment. The correlation between raters was only .65 for this subscore, which projects to a reliability of .79 when there are two raters.

Based on earlier data, the scoring of the "legibility" category was revised to increase the likelihood that it could be scored reliably in the future. The category's scoring decision rules were made far more objective, and interrater agreement is likely to increase on future scorings.

It should be noted that some of the differences between raters on sub-scores are not the result of errors which were not counted but errors which were categorized differently by raters. For example, if a candidate was supposed to have written "the girl was unharmed" and instead wrote "the goal was unharm" the raters would clearly agree that there were two errors; but one rater might consider both to be spelling errors while another rater considered the first error to be a wrong word and the second error to be a grammar error. These differences in categorization would not reduce the reliability of the overall score but would reduce the reliability of the sub-scores involved.

Figure 7 Inter-rater Reliability for LAWS Category Scores

Sub-scores	r between raters	r_{xx}
Spelling	.88	.94
Grammar	.92	.96
Wrong words/phrases	.90	.95
Missing words/phrases	.94	.97
Inappropriate commas	.93	.96
Legibility	.65	.79

Figure 8 LAWS sentences: Difficulty and Intercorrelations

FORM 1 (N=260)	Mean*	SD	Range of r's with other items	Mdn r	r with TOTAL
Practice Item	0.44	0.79	.50 - .58	.52	--
Sentence 1	1.61	1.64	.51 - .73	.65	.87
Sentence 2	1.09	0.88	.53 - .73	.65	.88
Sentence 3	0.88	1.36	.52 - .66	.64	.83
Sentence 4	0.69	1.26	.53 - .68	.61	.82
Sentence 5	0.68	1.19	.58 - .68	.66	.84

FORM 2 (N=222)	Mean*	SD	Range of r's with other items	Mdn r	r with TOTAL
Practice Item	0.51	0.84	.26 - .55	.48	--
Sentence 1	1.64	2.29	.55 - .79	.68	.87
Sentence 2	0.82	1.27	.52 - .82	.69	.87
Sentence 3	1.05	1.58	.53 - .82	.72	.84
Sentence 4	0.80	1.29	.26 - .74	.61	.81
Sentence 5	0.88	1.25	.45 - .72	.64	.85

NOTES
*Mean values are equal to average number of errors made on the sentence. Thus a high figure equals more errors.
Correlations with total are not corrected for part-whole overlap.
Practice item is not correlated with total because it is not part of the total score.

Construct Validity

Two studies comparing test scores to police academy training grades are in progress. Several completed construct validity studies are reported below.

Correlation with self-ratings of Writing Ability

At each of the five administrations, before the written or LAWS tests were administered, test takers were given a Self-Rating Form (SRF) to complete voluntarily. The SRF is a 23 item questionnaire requiring persons to rate themselves on a series of basic abilities and competencies. The SRF has been used with BDA's testing programs since 1989, and a great deal of reliability and validity data is available on this instrument and its items. One particularly relevant item of the SRF is shown below. Test takers rated themselves on the following trait approximately 30 minutes before completing a series of exercises and tests which included the LAWS. The wording of the self-rating trait and the rating scale used appears in Figure 9.

The raw correlation between the LAWS category score and the above self rating was +.37 (n=512, p<.001). This is an encouraging correlation when one considers that it is based on a single self-rating on a five point scale which, naturally, has a fairly low retest reliability. A large-scale study (N=5,557) has shown the writing ability self-rating to have a retest reliability of .52. When the r of .37 is adjusted for the unreliability of the criterion measure, the corrected r is +.51. Thus the LAWS test shows a robust correlation with applicants' own assessments of their writing skills.

Figure 9 Correlation Between LAWS and a Self-Rating of Writing Ability

104	**WRITING ABILITY** Ability to write clear, accurate and understandable reports and other communications.	(A) below average (B) about average (C) above average (D) very high (top 10%) (E) outstanding (top 1-2%)

Correlation between LAWS and self-rating of WRITING ABILITY = +.37 (N=512, p<.001)

Re-test reliability of self-rating of WRITING ABILITY = .52 (N=5,557)

Corrected r between LAWS and self-rating of WRITING ABILITY = +.51 (N=512, p<.001)

Correlation with scored 150-word essays

For 38 of those who have taken the LAWS, more traditional narrative writing exercises were also administered and scored. These exercises were scored for WRITING MECHANICS (grammar, punctuation, spelling, proper sentence structure) as well as ORGANIZATION of the theme and the CLARITY with which the message was communicated. An overall score was obtained by summing the three scores. Each paper was scored by two persons who then discussed their ratings afterward and adjusted as warranted. The correlations between the LAWS test and these writing exercises appears in Figure 10. While the data are encouraging and indicate a strong positive relationship, the N is small and more research is clearly needed in this area.

Figure 10 Correlation Between LAWS and a Traditional Writing Exercise

Writing Exercise Subscore:	r with LAWS:	signif.	N
Theme Organization	+.38	(p<.05)	38
Clarity of communication	+.28	ns	38
Writing Mechanics	+.61	(p<.001)	38
Overall Writing Score	+.51	(p<.001)	38

Correlation with oral examination/interview scores

For 156 of the test takers in the pool, BDA also provided oral examination services in which the test takers were interviewed by a panel of three police officials and scored on a number of traits including their career preparation, judgment, work record, and communication/interaction skills. The interviewers had no knowledge of test takers' LAWS scores nor their written test scores, to assure independent ratings. Analysis of the oral examination data showed a correlation of +.308 (p<.001) between LAWS category scores and summary oral scores, and a correlation of +.411 (p<.001) between LAWS category scores and ratings of Communication/Interaction Skills (CIS). CIS is essentially defined as how effectively the candidate communicated and interacted with the panel during the oral examination. It is therefore a combination of oral communication skills and the ability to interact and relate well with examiners. The +.41 correlation between LAWS scores and CIS is particularly interesting because it indicates a "crossover" from written to oral communication skills, and

its relative strength suggests that the LAWS test may be a broader measure of basic language skills than it was originally thought to be.

Figure 11 Correlation Between LAWS and Oral Examination Scores

Score:	r with LAWS:	signif.	N	correction for range restriction
Oral Exam Total Score	+.31	(p<.001)	158	+.34
Communication/Interaction Skills	+.41	(p<.001)	158	+.45

Correlation with Educational Level

Test takers were asked to describe their educational level on a simple five-point scale ranging from (1 = less than high school) to (5 = 4-year college degree), as illustrated in Figure 12. This scale produced a mean of 3.50 and a SD of 1.10 in this group. The correlation between LAWS and this basic measure of educational level is +.29 (n=510, p<.001). While this is a robust correlation based on a somewhat gross scale, the researcher is surprised that the relationship between LAWS scores and educational level was not higher. However, Figure 12 does demonstrate a strong linear relationship between educational level and number of errors made on the LAWS test.

Figure 12 Relationship Between LAWS Scores, Writing Errors and Educational Level

Educational Level:	N	# LAWS errors
1 Not a high school graduate	2	8.00
2 High School, no college	94	8.26
3 Some College, no degree	204	5.30
4 Two-year Degree	65	4.40
5 Four-year college degree	145	2.65

Correlation with cognitive ability

As mentioned earlier, test takers took an entry-level police officer examination after completing the LAWS test. This test, known as the Municipal Police Selection Test (MPST) is a 100-question multiple choice test. There are six different forms of this test, all of relatively equal difficulty. There are two distinct components of the MPST: cognitive ability and a non-cognitive measure (a personality/interest questionnaire). Correlations between LAWS and the MPST components are shown in Figure 13.

Figure 13 Correlation Between LAWS and Written Examination Scores

Score:		r with LAWS:	signif.	
Total score	(100 items)	+.49	(p < .001)	
Cognitive ability	(70 items)	+.56	(p < .001)	N = 534 for
Non-cognitive (VIQ)	(30 items)	+.11	(p<.01)	all cases

The items of the MPST primarily involve study skills, reasoning and reading comprehension rather than more pure verbal ability items such as vocabulary or writing mechanics. Based on the above data, it is likely that the LAWS would correlate higher, perhaps .60-.70, with a multiple-choice measure of pure verbal ability.

LAWS Sub-Scores Correlated With External Criteria

An important question regarding the construct validity of the LAWS test relates to how its various score components contribute to the total LAWS score, and how those sub-scores correlate with other measures. Figure 14 presents such results.

Figure 14 Correlation Between LAWS Score Components and External Criteria

	r with LAWS	cognitive ability	educ	writing abil self-rating	writing sample	oral communic
LAWS Category Scores	1.00	.56***	.29***	.37***	.51***	.41***
Spelling errors	-.83***	-.46***	-.25**	-.29***	-.47**	-.31***
Grammatical errors	-.65***	-.48***	-.23*	-.19***	-.28	-.27**
Wrong words or phrases	-.51***	-.42***	-.18**	-.10*	-.48**	-.19*
Missing words or phrases	-.52***	-.40***	-.20***	-.11*	-.55**	-.26***
Inappropriate commas	-.15*	-.05	-.10*	-.03	-.08	-.08
Illegibility	-.38***	-.20***	-.08	-.18*	-.00	-.09
	N=534	N=534	N=510	N=516	N=38	N=158

NOTE: Except for the LAWS category score, signs of correlation coefficients are all negative because the variables represent number of errors made.

In ascertaining what the LAWS test measures, it is also useful to review the distribution of errors made by test takers on the LAWS sentences. These are as follows:

Spelling errors =	45%	of all deductions
Grammatical errors	15%	
Wrong word/phrase	11%	
Missing word/phrase	11%	
Inappropriate comma	10%	(no longer considered)
Legibility	8%	

In interpreting these data along with Figure 14, it is important to remember that spelling in the LAWS test is not at the traditional level of a spelling test. Because the language of the LAWS sentences is essentially at the 8th-grade or conversational level, spelling errors generally involve day to day word usage rather than spelling errors involving sophisticated terminology or rarely-used words. While many very intelligent people have trouble spelling difficult words, an inability to spell everyday words to which we have a high exposure indicates a more fundamental language problem. This is most likely why the LAWS spelling subscore correlates so well with so many other variables.

During early tryouts of the LAWS scoring system, deductions for inappropriate commas were part of the system. Analysis of the results for the first 250 candidates led to the decision to eliminate deductions for “inappropriate commas” from future LAWS testings. This subscore did not correlate

with any other measure of overall writing ability, nor with cognitive ability nor education. As such, its value to the exercise appeared to be negative, and it has been eliminated as a scoring factor.

The scoring of the Legibility subscore was also refined as a result of these data. Earlier data showed that the "legibility" factor had relatively low inter-rater reliability, and likewise lower levels of construct validity. More scoring guidance has been developed for future administrations, to reduce rater subjectivity and improve reliability. The reliability of this subscore has been around .80 for more recent testings.

Demographic Breakdowns of LAWS Scores

For all nine test administrations, test takers were asked to voluntarily complete a race/gender self-report form. Of the 534 test takers taking the LAWS test for the first time, 518 identified their race and 530 identified their gender. The breakdown of their scores on the LAWS test and, for comparative purposes, a cognitive ability test, appears in Figure 15.

Figure 15 Demographic Breakdowns of LAWS Scores

	Cognitive Ability Scores				LAWS Category Scores			
	Mn	SD	N	SD diff*	Mn	SD	N	SD diff*
All test takers	55.07	8.08	534		4.12	1.57	534	
White	56.55	7.16	412		4.34	1.45	412	
Black	50.10	7.84	41	-0.80 sd**	3.46	1.83	41	-0.56 sd**
Hispanic	49.20	9.36	65	-0.91 sd**	3.28	1.68	65	-0.68 sd**
Minority (B+H)	49.79	7.55	106	-0.84 sd**	3.35	1.73	106	-0.63 sd**
Other or unidentified	53.44	9.57	16		3.38	1.49	16	
Male	55.10	7.96	460		4.07	1.59	460	
Female	54.79	8.91	70	-0.04 sd	4.50	1.35	70	+0.27 sd*

** Significant at .001 level * significant at .05 level

It should be noted that the sample sizes (106 minorities and 70 females) are relatively small, so that definitive conclusions cannot be made. However, the preliminary data are encouraging. The LAWS test showed less adverse impact than a standardized cognitive ability test in this sample of 534 persons. This is an encouraging finding, especially in light of the fact that the LAWS test is a highly reliable, objectively scored, purely cognitive test.

Females within this sample outscored males by 0.27 standard deviations on the LAWS. Again, the sample size is quite small. Interpretations are therefore premature..

Retest / Practice Effects

One concern with any test which may be re-administered is the effect of practice on future scores. Since there were 149 persons who took the LAWS test two or more times in a relatively short period (ranging from 7 days to 148 days and averaging about 30 days). Their scores appear in Figure 16.

Figure 16 LAWS Re-Administrations: Practice Effects

	N	r between testings	Number of LAWS errors Mn	SD	Mn	SD	Improvement:	
First testing to second	103	.789	4.33	1.50	4.51	1.42	0.18	(+0.12 SD)
Second testing to third	36	.747	4.64	1.18	5.14	1.13	0.50	(+0.42 SD)
Third testing to fourth	10	.846	5.10	0.83	4.50	1.12	-0.40	(-0.48 SD)
Average Practice Effect	149	.793	4.46	1.45	4.66	1.38	0.20	(+0.14 SD)

Figure 16 shows that on a second testing, the LAWS category score improved by 0.18, which is 0.12 SD's. Improvements related to third and fourth testings suffer from very small N's (36 and 10 cases respectively and do not produce a consistent pattern. However, it is clear that practice effects are small; the average across all the above data is an improvement of 0.14 SD's. The correlation between testings also shows high consistency between scores on a re-test.

Reporting LAWS Scores

While LAWS scores have been considered experimental until recently, a format and approach for reporting them has been developed. At this point BDA is recommending to its clients that LAWS scores be advisory rather than treated on a pass/fail or ranking basis, and that they be used as one of many pieces of information considered when making final employment decisions.

In designing the rubrics for reporting candidate performance, BDA used the six-category system described throughout this report. In labeling each category, BDA tried to develop low-inference descriptors, i.e., descriptors which closely matched the content of candidates' actual test behavior rather descriptors which inferred behavior beyond what was directly observed. Figure 16 illustrates these content-relevant descriptions and normative information provided to guide decision-making.

Summary

The Listening and Writing Skills (LAWS) test is a new approach to writing assessment which has demonstrated high levels of content and construct validity. Based on the research evidence presented in this report, the LAWS test appears to assess basic written language skills rather directly and with high reliability, while displaying correlations with a wide range of other measures, both written and oral. Further criterion-related research and LAWS refinements are underway.

Figure 17 Laws Scoring and Normative Categories

Excellent*

No errors made in writing exercise. Able to listen well and convert what he or she hears to correct written narrative. Excellent grammar, spelling and sentence formation.

Very Good*

Did very well on the writing exercise. Able to listen well and convert what he or she hears to correct written narrative with few errors. No evident problems in spelling, grammar, clarity and sentence formation.

Good*

A few minor errors on the writing exercise. Generally able to listen well and convert what he or she hears to correct written narrative, making occasional errors in spelling, grammar, clarity or sentence formation. Based on this test, appears to be in the average range of writing ability.

Acceptable*

Made several errors on the writing exercise, but overall performance on spelling, grammar, clarity and sentence formation was acceptable. Makes some errors in producing written narrative based on what he or she hears, and shows room for improvement in writing skills, but is in the acceptable range.

Conditionally Acceptable*

Made a number of errors on the writing exercise, so that overall performance was questionable. May have difficulty producing clear, correct written narrative based on what he or she hears. Further assessment is advised to determine whether writing skills meet the department's standards.

Unacceptable*

Was not able to write clear and acceptable, problem-free sentences based on what he or she heard. Tends to make a variety of errors in producing written narrative. Based on these results, appears to need considerable remediation before writing skills can be considered acceptable.

	*Percent of Candidates who are:		
	Above this Category	In this Category	Below this Category
Excellent	0	17%	83%
Very Good	17%	29%	54%
Good	46%	22%	32%
Acceptable	68%	11%	21%
Conditionally Acceptable	79%	11%	10%
Unacceptable	90%	10%	0%