LSAC RESEARCH REPORT SERIES

Program of Research on Legal Writing: Phase II: Research on a Writing Exercise

Hunter M. Breland
Sydell T. Carlton
Susan Taylor

Law School Admission Council
Research Report 96-01
January 1998

A Publication of the Law School Admission Council

The Law School Admission Council is a nonprofit corporation that provides services to the legal education community. Its members are 194 law schools in the United States and Canada.

Law School Admission Council fees, policies, and procedures relating to, but not limited to, test registration, test administration, test score reporting, misconduct and irregularities, and other matters may change without notice at any time. To remain up to date on Law School Admission Council policies and procedures, you may obtain a current *LSAT/LSDAS Registration and Information Book*, or you may contact our candidate service representatives.

This study is published and distributed by the Law School Admission Council (LSAC). The opinions and conclusions contained in this report are those of the author and do not necessarily reflect the position or policy of the Law School Admission Council.

Table of Contents

Advisory Committee on Legal Writing

Stephen Armstrong
Shearman & Sterling

Susan Brody
The John Marshall Law School

Mary Lawrence
University of Oregon School of Law

Christopher Rideout
Seattle University School of Law

Helene Shapo
Northwestern University School of Law

Barry Vickrey
University of South Dakota School of Law

Acknowledgments

In addition to the members of the Advisory Committee on Legal Writing, a number of other persons assisted with this project. We are indebted to the following persons:

Dennis Hynes, Professor
K. K. DuVivier, Senior Instructor
University of Colorado School of Law

Cynthia M. Hinman, Assistant Professor and Director/Lawyering Skills
The John Marshall Law School

Frederick M. Hart, Professor
University of New Mexico School of Law

Maria Ciampi, Assistant Dean
Mary Conlon, Registrar's Office
St. John's University School of Law

Jean Henriques, Admissions Officer/Registrar
University of South Dakota School of Law

Louis V. Dibello, Director of Educational Research
Law School Admission Council

Laura Jerry, Lead Research Data Analyst
Educational Testing Service

Executive Summary

Based on the results of a Phase I investigation into the nature of legal writing, a prototype writing assessment, the Diagnostic Writing Skills Test (DWST) for entering law students, was developed. The DWST is composed of two multiple-choice subtests based on prompts and responses to the Law School Admission Test (LSAT) Writing Sample and contains a total of 48 items. The DWST was administered to over 500 entering law students in five different law schools, and data were collected on the performance of these students in first-year legal writing courses and in law school courses generally. The experimental data indicated that the prototype DWST: (1) was completed by 89 percent of students who were allowed 45 minutes of testing time and 97 percent of students who were allowed one hour, (2) was relatively difficult for the students tested, and (3) had relatively low reliability compared to traditional tests. Despite the difficulty and reliability problems, the DWST predicted performance in legal writing courses better than either the LSAT or undergraduate grade-point average. The prototype test was second only to the LSAT in the prediction of law school grade average, and it made a significant contribution to the prediction of law school writing course grade and law school grade average when used along with the LSAT and undergraduate grade-point average in the prediction. Gender and ethnic comparisons indicated that a test like the DWST would reduce gender and ethnic differences slightly if incorporated as part of the LSAT. Next steps are to revise the prototype test to reduce its difficulty; to correct some questions that seemed confusing to students; and to increase its reliability.

Abstract

Based on the results of a Phase I investigation into the nature of legal writing, a prototype writing assessment, the Diagnostic Writing Skills Test (DWST) for entering law students, was developed. The DWST is composed of two multiple-choice testlets based on prompts and responses to the Law School Admission Test (LSAT) Writing Sample and contains a total of 48 items. The DWST was administered to over 500 entering law students in five different law schools, and data were collected on the performance of these students in first-year legal writing courses and in law school courses generally. The pretest data indicated that the prototype DWST contained six items that needed revision because of low percentages answering them correctly, low biserial correlations, or low percentages of a high-scoring group answering them correctly. The reliability of the DWST was estimated in the range .70–.81. This relatively low reliability was attributed to the local dependence of the testlets, possible multidimensionality, the use of a four-choice item structure, lack of motivation among examinees, and the six questionable items, which have since been revised. Despite these problems, the DWST predicted performance in legal writing courses better than either the LSAT or undergraduate grade-point average. The prototype test was second only to the LSAT in the prediction of law school grade average, and it made a significant contribution to the prediction of law school writing course grade and law school grade average when used along with the LSAT and undergraduate grade-point average in the prediction.

Introduction

In Phase I, a project to define legal writing (Breland & Hart, 1994), samples of student legal memoranda were collected from 12 law schools dispersed geographically and representing a range of student abilities. These samples, 237 in all, were subjected to a series of analyses following four sets of ratings of their overall quality (by legal writing instructors and by independent legal consultants) and two sets of ratings (by legal writing instructors) of specific aspects of the memoranda. These analyses indicated that the following specific aspects of legal memoranda were most closely associated with overall quality:

- Legal analysis
- Style (including clarity, conciseness, organization, logical continuity, paragraphing, and focus)
- Mechanics (including grammar and usage, punctuation, editing and proofreading, sentence control, diction, and spelling)

Legal analysis and style had roughly the same relationship with overall quality, while mechanics was somewhat less important.

The objective in Phase II was to determine what kind of writing assessment administered before law school would best represent the results from Phase I. Since applicants to law school could not be expected to know anything about legal analysis, it was clear that legal analysis could not be a part of the assessment. Further understanding of the nature of legal writing and the type of writing assessment that would be most appropriate for applicants to law school was obtained from an examination of past research on legal writing.

Past Research

An early study by Coffman and Papacristou (1955) obtained results similar to those of Phase I. In this study, law school professors rated samples of law school student writing. Subsamples of papers were classified as being either well-written or poorly written. Coffman and Papacristou concluded that:

> A study of the papers revealed few of the gross errors in mechanics and grammar which are often found in the papers of high school seniors. The outstanding characteristic of the poor papers was their lack of clarity resulting from poor organization, both in the overall outline and in sentences and paragraphs. In contrast to the papers judged well-written, the poorly written papers tended to exhibit no underlying structure, no smooth flow of ideas to lead the reader from point to point.

A year later, Olsen (1956) reported a study of several experimental tests, including a 30-minute editing test. The 30-minute editing test correlated better with first-year average law school grades than did a 60-minute test of reading comprehension (.36 to .34) and better than a 35-minute test of data interpretation (.36 to .24). Only a 40-minute Law School Admission Test (LSAT) subtest of Principles and Cases predicted better than did the 30-minute editing task (.39 to .36).

Pitcher (1962) studied two experimental writing tests: a 30-minute interlinear exercise (an editing task) and a 50-minute combining sentences test, both of which were scored by human judges. In five different law schools, both types of tests yielded average correlations of .19 with the first-year law grade average; the 175-minute LSAT correlated .31 on average, as did the prelaw record (undergraduate grades). When correlated with first-year writing course grade, the 30-minute editing exercise predicted better than the 175-minute LSAT (.26 to .14). The combining sentences test correlations averaged .23 and the prelaw record .31. Subsequent analyses showed also that the two experimental tests added significantly to the prediction of first-year average grades and writing course grades beyond what was possible using the LSAT alone.

In 1962, a machine-scoreable writing ability test was introduced as part of the LSAT, and the value of writing items as predictors of law school grades was investigated (Pitcher, 1965). In this relatively large study involving ten law schools and over 1,600 law school students, the principal outcome appears to have been that a multiple-choice test of writing skill did not add much predictive validity beyond that possible with the LSAT alone. And yet it was not noted that the LSAT was at that time a test of approximately three hours in length, while the writing ability test was only 40 minutes in length. The same result most likely would have occurred if any of the brief subtests that constituted the LSAT had been analyzed with respect to their ability to add to the rest of the test. The same Pitcher (1965) study showed that the 40-minute writing ability test correlated .21 with first-year average law grades as compared to a correlation of .24 for the prelaw record and a correlation of only .29 for the entire three-hour LSAT.

A subsequent summary of validity studies covering a number of years (Schrader, 1976) indicated that the LSAT was the best predictor, undergraduate grade-point average (UGPA) second, and the writing ability score third. Other studies (Pitcher, 1975, 1976), comparing combinations of LSAT, UGPA, and writing ability, suggested that the writing ability score did not add significantly to the predictions of average law school grades possible using LSAT and UGPA alone. Powers (1980) examined the possibility that the writing ability score would be of special value in the assessment of minority students, but the results were not promising. In 1980, the writing ability test was removed from the LSAT, and an unscored writing sample was introduced in 1982.

The results of these studies were disappointing in part because the only objective at that time was to increase the prediction of first-year law school grades beyond what was possible using the LSAT alone. Given that the LSAT was a 175-minute test, it was difficult to demonstrate that any brief test could add significantly to the prediction of law school grades. Moreover, law school grades may not be the best criterion for evaluating a writing test; a better criterion might be law school writing course grades.

The writing ability test that was being evaluated in 1976 consisted of a 20-minute Error Correction test and a 20-minute Sentence Correction test. The disappointing results obtained in studies of these sentence-level tests may have been because they did not emphasize aspects of style shown to be important in the Phase I study. The clarity of a sentence, phrase, or word cannot be evaluated without reference to the entire piece of writing that it is a part of. Conciseness cannot be evaluated without knowing what has come before, for example, whether a statement has been repeated. Logical continuity cannot be evaluated without reference to the sentences or paragraphs that need to be connected. The assessment of the kinds of writing skills indicated by the Phase I results would require the use of an entire piece of writing.

The Interlinear Exercise

An old form of writing assessment used by the College Board in the 1950s and used in some law school writing assessments at about the same time used an entire piece of writing (see Olsen, 1956, and Pitcher, 1962). This form of writing assessment was known in College Board programs as an interlinear exercise. The interlinear exercise consisted of poorly written material that required the student to find and correct deficiencies. It can be much more complex than simple editing and, unlike the usual multiple-choice tests of writing skill, the assessment is not limited to sentence problems. The ability to reorganize text and to introduce smooth transitions, logical continuity, coherence, precision, proper tone, attention to audience, and clarity can also be tested. Despite its promise and demonstrated validity, this test was discontinued in the 1960s because of the time and cost required to score it.

The interlinear exercise is responsive to current issues in educational assessment in that it would involve the performance of tasks that are valued in their own right (Linn, Baker, & Dunbar, 1991, p. 15) and it is compatible with current efforts to develop assessments which have the purpose of capturing in tests those tasks with "the same demands for critical thinking and knowledge integration as required in desired criterion performances" (Shepard, 1991, p. 9).

Messick's (1989) argument concerning the consequential basis of validity is relevant here as well. If an assessment leads teachers and coaching schools to focus on the assessment format rather than the desired skill, then these consequences should be considered in judging the validity of the assessment. The interlinear exercise, on the other hand, could encourage coaching schools and others to emphasize skills that effect permanent improvement in an applicant's legal writing ability.

One study conducted before the College Board discontinued the interlinear exercise (Godshalk, Swineford, & Coffman, 1966), attests to the predictive effectiveness of this type of test. For that study a comprehensive writing ability criterion was developed that consisted of five free-writing tasks, each scored by five different English teachers. A number of more efficient tests were then used as predictors of the writing ability criterion including the Scholastic Aptitude Test (SAT), multiple-choice tests of sentence correction, and the interlinear exercise. In the full sample of 646 cases studied, an interlinear exercise correlated .67 with the writing ability criterion. In a subsample of 158 students for whom SAT scores were available, an interlinear exercise correlated better with the writing ability criterion than did the SAT Verbal score (.68 to .63). That is, a 30-minute interlinear exercise predicted writing ability better than did a one-hour verbal aptitude test. The interlinear exercise was also included in a number of multiple-regression analyses along with the SAT and other test item types. Even when combined with the SAT and a multiple-choice test of sentence correction skills, the interlinear exercise made a significant contribution to the prediction of writing ability. Godshalk et al. (1966) concluded: "The interlinear exercise is one of the more valid types of question and often contributes uniquely to the multiple correlation" (p. 24); and "The interlinear exercise makes a singular contribution to the prediction of writing ability" (p. 29).

Some research evidence indicates that a single 30-minute interlinear exercise would have a validity for predicting later writing performance of about the same magnitude as two 45-minute free-writing tasks. Breland, Camp, Jones, Morris, and Rock (1987) conducted a study similar to the Godshalk et al. study and found that two 45-minute writing tasks correlated .66 and that two different 45-minute writing tasks correlated .72 with a comprehensive writing ability criterion for first-year college students. The average of these two predictive correlations (.69) is not significantly different from the .68 correlation that Godshalk et al. obtained for a 30-minute interlinear exercise.

Validity is not the only reason that it would be useful to consider the administration of a writing exercise with the LSAT. Other reasons include the following:

1. If writing is important in the law, and almost everyone agrees that it is, then it seems logical that it should be a part of the LSAT test battery.

2. Currently, writing skills are not emphasized heavily in law school admission, since the present writing sample is not scored. In a sense, this represents a lost opportunity for women because research shows that women perform better on writing tasks than do men, on average. If a scored writing task were administered, and incorporated as a part of the total LSAT score, it would quite likely boost the average LSAT scores of women and make admission to law school more likely for them.

3. An interlinear exercise would hold considerable promise for computerized administration and constructed-response scoring. If there is interest in computerizing the LSAT at some future time, the interlinear exercise would fit well with such an interest.

Development of a Prototype

In developing a prototype writing assessment for use in law school admission, some of the first issues faced were those of how the test would be scored, how much testing time could be allowable, and where to obtain text material for such a test. Interlinear exercises of the type described previously in this report had traditionally been scored by human judges. Since it was the cost and inefficiency of human scoring that led to their demise, the first decision made was that the test would not be scored by human readers. This meant that a test type originally intended for human scoring would have to be adapted such that it could be machine scored.

How long could the test be? The current sample collected for the LSAT requires only 30 minutes, so initial timing considerations assumed that a new prototype test should require about the same amount of time. Nevertheless, it was known that earlier machine-scored writing tests used by the Law School Admission Council (LSAC) had required 40 minutes of testing time. It was also known that the reliability of the prototype test would be directly related to its length. Consultation with staff at LSAC on the timing issue indicated that a prototype test requiring 45 minutes of testing time would be acceptable, so a decision was made to design the prototype test for administration in 45 minutes.

The text material for the test needed to be appropriate for law school applicants and free of copyright restrictions. The copyright problem could be avoided by creating original material, but that would require more test development time and one could not be certain that the material would be appropriate for law school applicants. A source of material that would avoid both of the appropriateness and copyright problems was the LSAT itself, namely the LSAT Writing Sample. These samples had been generated by law school applicants in response to prompts developed by LSAC and thus should be appropriate. While it would be necessary to modify text as written to create a useful prototype, the existence of a variety of prompts and responses to them served as a useful starting point. After examination of a number of LSAT Writing Sample prompts and responses, two topics were chosen, and an assessment task was developed for each.

Rationale

The tasks developed ask the examinee to read a hypothetical response made to the topic and to respond to a number of questions about the response. The questions focus primarily on organization, audience, transition, conciseness, and clarity. These categories represent areas identified in the Phase I study of legal writing, and they are also areas in writing competency that trouble educated, upper-level writers in college and the professions in general. This conclusion is based on observation of several populations—first-year law school students, students writing essays for the LSAT, juniors and seniors in science and humanities at the university level, and researchers and management personnel in research companies.

Probably the greatest deficiency in upper-level writers is in the area of sensitivity to audience. This facility is necessary if the writer is to include appropriate detail, presented at a syntactic level and in a diction accessible to targeted readers. Organization includes the ability to establish a thesis and support it in a logical sequence of paragraphs that are themselves organized logically. We included items on the introductory paragraph, since a writer ought to understand how to exploit that paragraph to help the reader form a mental map of the rest of the document. Introductions also fall under the audience category, as they signal what kind of reader the writer is expecting.

To test skill in transition and development, we created items that work on two levels. Some items simply test the examinee's ability to understand the logical relationships signaled by particular transitional terms or phrases. Others ask the examinee to judge, from a sentence's context, what kinds of transitions are needed, and when.

The items chosen to test conciseness and clarity were not intended to test basic grammatical knowledge; rather, the items chosen contain the errors that appear when more sophisticated writers attempt to give shape to complex ideas or solve a problem in transition. In particular, the items are intended to test a writer's ability to strategize, to choose a sentence pattern that will best emphasize a particular point. Items also ask the examinee to judge whether certain expressions are precise, inflated, or set at an inappropriate level of diction for the targeted reader.

Although upper-level writers in college and the professions tend to have fewer problems with grammar and punctuation, they are not immune to such problems. Excluding those who speak English as a second language, most writers at this level do not show significant punctuation or grammatical weakness, with one exception. Comma splice errors (joining two sentences with only a comma) are, surprisingly, very common. This most likely has to do with the mature writer's "ear" for the sound of a close logical relationship between two sentences, along with some persisting uncertainty about the use of semicolons and the grammatical function of the transitional word, however. We have included a small number of items addressing comma splice errors because they are prevalent in a skilled writing population and because they involve syntactical and organizational decisions.

Test Construction

As previously noted, two tasks were developed based on the two LSAT Writing Sample topics selected. The first task was called "Valerie" and the second "Winfield," since these words were descriptive of the topics. Valerie consisted of 23 items and Winfield consisted of 25 items. Table 1 lists the classifications of the item types in Valerie and Winfield. Two forms of the prototype test were constructed by ordering Valerie first and Winfield second in one form and Winfield first and Valerie second in the second form. The form having Valerie first was labelled Form V, and the form having Winfield first was labelled Form W. Table 1 also lists, in the last column, diagnostic categories into which items were classified.

TABLE 1
Classification of item types

Form	Item Number	Classification	Diagnostic Category
V	1 26	Organization/Thesis	Organization
	2 27	Organization/Introductory paragraph	Organization
	3 28	Precision/Clarity	Clarity
	4 29	Transition	Transition
	5 30	Usage/Transition	Transition
	6 31	Clarity	Clarity
	7 32	Audience/Strategy	Strategy
	8 33	Transition	Transition
	9 34	Logic/Clarity	Clarity
	10 35	Subordination/Clarity	Clarity
	11 36	Strategy	Strategy
	12 37	Conciseness	Conciseness
	13 38	Organization	Organization
	14 39	Audience/Logic/Strategy	Strategy
	15 40	Conciseness	Conciseness
	16 41	Subordination/Transition	Transition
	17 42	Usage/Conciseness	Conciseness
	18 43	Transition	Transition
	19 44	Organization	Organization
	20 45	Organization/Development	Organization
	21 46	Organization/Conclusion	Organization
	22 47	Organization/Development	Organization
	23 48	Organization	Organization
W	1 24	Organization/Thesis	Organization
	2 25	Organization/Introductory paragraph	Organization
	3 26	Precision/Conciseness	Conciseness
	4 27	Precision/Clarity	Clarity
	5 28	Transition/Precision/Pronoun reference	Transition
	6 29	Organization	Organization
	7 30	Conciseness	Conciseness
	8 31	Transition	Transition
	9 32	Subordination/Transition	Transition
	10 33	Usage/Clarity	Clarity
	11 34	Conciseness	Conciseness
	12 35	Organization/Development	Organization
	13 36	Parallelism/Conciseness	Conciseness
	14 37	Transition	Transition
	15 38	Transition	Transition
	16 39	Organization/Development	Organization
	17 40	Usage/Clarity	Clarity
	18 41	Precision/Clarity	Clarity
	19 42	Transition	Transition
	20 43	Organization	Organization
	21 44	Subordination/Strategy	Strategy
	22 45	Audience/Strategy	Strategy
	23 46	Development/Strategy	Strategy
	24 47	Development/Strategy	Strategy
	25 48	Organization/Conclusion	Organization

A more accurate description of the test specifications is given by using all of the item descriptions in Table 2. Here, each test objective is listed separately even though individual items may test more than a single objective. Copies of Form V and Form W of the prototype test are included in this report in Appendix A, and an illustration of a diagnostic score report is included in Appendix B.

TABLE 2
Test specifications

Classification	Number of Occurrences	Percentage of Occurrences
Organization	15	19
Transition	11	14
Clarity	8	10
Conciseness	7	9
Development	6	8
Precision	4	5
Subordination	4	5
Usage	4	5
Audience	3	4
Thesis	2	3
Introductory paragraph	2	3
Logic	2	3
Parallelism	1	2
Pronoun reference	1	2
Total	77	

Note. An item may test more than a single classification; thus, the concepts tested total more than the 48 items of prototype test.

Test Timing

On the basis of informal administrations of the test to staff and college students, initial estimates were that the test could be administered in approximately 50 minutes, with 45 minutes of actual testing time.

Comments on the Prototype Test

There are a number of characteristics of the test that are worthy of comment. The arrangement of the test into two testlets raises the issue of local dependence, which can be expected to produce a lower test reliability than a test made up of all independent items. Nevertheless, it would not be possible to cover the test objectives using all independent items. The hope, in a test of this type, is that some reliability might be traded for validity (Wainer & Thissen, 1996). Reliability can also be reduced if a test is multidimensional as seems likely in the case of the prototype constructed.

Two further characteristics of the prototype test as constructed will reduce reliability. Most multiple-choice tests have five-choice options (that is, the examinee has five choices among which to select for each question). In an effort to speed up the prototype test to fit within a 50-minute administration time limit, the decision was made to use four-choice options for each question. The use of four choices reduces the item variance and thus the test reliability. Finally, reliability will be diminished in the prototype test because some of the questions are inefficient. It is time-consuming to read a passage over and over to determine where best to place any particular sentence, for example. And yet, this kind of continuity testing was deemed important by the Advisory Committee and Phase I results supported such a judgment.

In summary, the reliability of the prototype test would be expected to be relatively low. The hope would be that reduced reliability would be compensated for by increased validity. Additionally, the reliability of diagnostic subscores would be expected to be even lower than the total score and thus directions for test use would have to be carefully stated.

Pretest of the Prototype

Five law schools agreed to conduct a pretest of the prototype writing skill assessment during the summer and fall of 1995. Two of these schools wanted to try the test with their entire entering classes while the others had agreed to ask for volunteers or to administer the test in selected classes. These institutions were sent instructions for administration of the test (Appendix C), a student consent statement to be read to students before they were administered the test (Appendix D), and student rosters for recording data (Appendix E). Two student rosters were to be completed, one including the students' names (which would be retained at the institution) and another with the students' names replaced by codes to be obtained from the LSAC. The rosters without the student names were to be returned to Educational Testing Service (ETS) to ensure student anonymity.

Test Timing

Since the time available for testing varied for each institution, and since precisely how much time to allow for the pretest was unknown, it was decided to allow a range from 45 to 60 minutes. The informal tryouts had shown that the subjects used could complete the test in 45 minutes, but it was not certain how much time students in the participating institutions would require. Additionally, some time would be required for orienting the students to the test and for reading the instructions. Institutions were requested to record the total time available for instructions and actual testing, the total time used for instruction, and the total time used for actual testing on the Test Administration Questionnaire (Appendix F).

Student Motivation

In a pretest situation, the motivation of the examinees is always a concern. If examinees do not do their best, it is not clear what the results mean. At the same time, students must be told that the test is experimental and that it will have no effect on their grades or other school outcomes. The consent statement read to students also told them that, if for any reason, they did not wish to take the test, they did not have to do so. Accordingly, in this kind of testing environment, one can never be certain that students were motivated to do well on the test.

Data Obtained

The five participating institutions were coded A, B, C, D, and E. A total of 19 persons took the test in Institution A, 227 in Institution B, 58 in Institution C, 189 in Institution D, and 11 in Institution E. Three of the persons taking the test in Institution A and one of those persons in Institution C were instructors. The participating institutions also provided writing course grades and law school grade average for most students who took the prototype test. Course grades and law school average were for the first semester of law school except for the writing course grades in Institution D, which were for a course covering the entire first year of law school.

Test Analyses

A total of 507 persons took the prototype DWST in the five participating institutions, with 254 taking Form V and 253 taking Form W. The overall mean score was 29.27 (out of a maximum of 48 items). Those administered Form V obtained a mean score of 28.68 and those administered Form W obtained a mean score of 29.86. The standard deviation was 5.73 overall, with a standard deviation of 5.40 for Form V and a standard deviation of 5.99 for Form W. The scores ranged from zero to a high of 42 (out of 48 items). The range for Form V was zero to 42 and that for Form W was 3 to 41. Three of the answer sheets for Form V were blank, however, and when these were excluded, the range of scores for Form V was from 3 to 42.

The frequency distribution of scores is shown in Figure 1. This distribution is essentially normal, except for a slight bimodality at the peak and a curious cluster of scores below 20. The cluster of low scores, together with the three blank answer sheets that were excluded from Figure 1, suggests that some examinees may not have been highly motivated. Random guessing alone would have produced an average score of 12, so that any score near 12 could be the result of random guessing. In fact, if one takes scores in the range from 8 to 15 and computes a weighted average, that average is 12.5. One can imagine that some of the scores above 15 resulted from a combination of guessing and some (but not much) actual reading of the questions. We believe that this clustering of scores in the lower tail of the distribution is important and more will be said about it below.

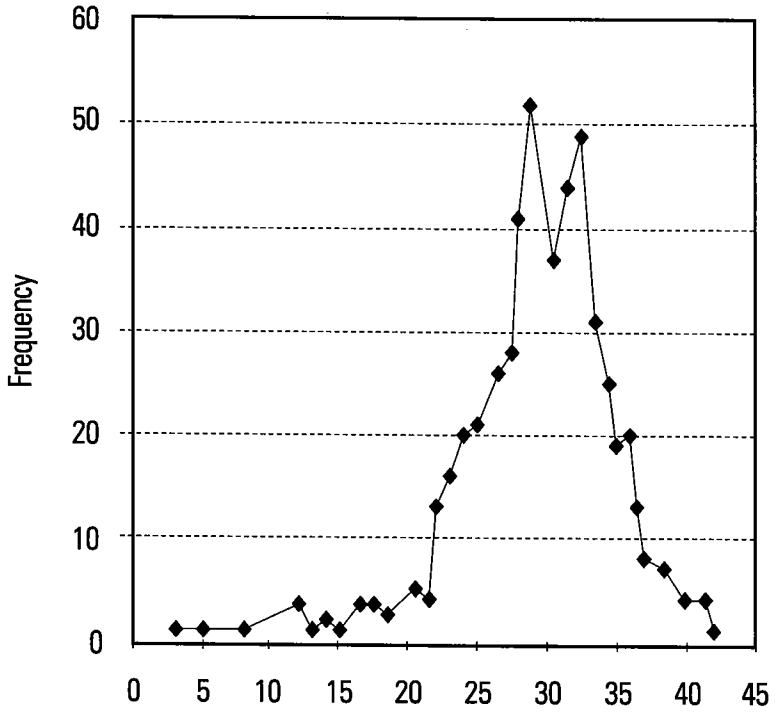

FIGURE 1. *Distribution of DWST scores*

Table 3 gives the total and subscore means, standard deviations, and ranges for each of the five participating institutions. Institution A has the highest total mean score (34.37), but this includes the scores of three instructors. When the instructors' scores are removed, this mean is reduced to 33.20. With the instructors' scores removed, Institution A still has the highest observed mean score on the DWST. Institution E had the lowest mean score (25.73), but only 11 students took the test. The total score ranges are especially interesting, since one student scored as low as 3 (in Institution D), which is below the score of 12 that could be expected from guessing alone.

TABLE 3

DWST diagnostic and total score means, standard deviations, and ranges by institution

Score	Items	Institution A* (N = 19)	B (N = 227)	C** (N = 58)	D (N = 189)	E (N = 11)
Organization						
Mean	15	9.58	7.78	9.02	8.14	6.54
S.D.		2.35	2.02	1.86	2.16	2.19
Range		5–14	1–13	5–13	0–13	3–10
Transition						
Mean	11	8.79	7.12	7.84	7.57	6.36
S.D.		0.77	1.69	1.76	1.65	2.60
Range		7–10	1–11	3–11	1–10	2–11
Strategy						
Mean	7	5.10	4.39	4.86	4.70	4.27
S.D.		1.12	1.27	1.31	1.28	1.14
Range		3–7	0–7	1–7	0–7	2–6
Conciseness						
Mean	7	5.05	3.81	4.69	4.19	2.91
S.D.		0.76	1.40	1.16	1.32	0.67
Range		3–6	0–7	2–7	0–7	2–4
Clarity						
Mean	8	5.84	4.91	5.88	5.43	5.64
S.D.		1.31	1.38	1.13	1.47	1.55
Range		2–8	2–8	3–8	1–8	3–8
Total						
Mean	48	34.37	27.98	32.39	30.04	25.73
S.D.		4.24	4.75	4.58	5.34	5.61
Range		25–41	12–37	19–42	3–41	13–34

**Includes three instructors

**Includes one instructor

Student Motivation

The relatively high percentage of students below, at, or near a guessing level of performance suggests that some students may not have been highly motivated when they took the test. This might readily be expected, since students were told that the test would not affect their law school grades and, in fact, that they did not have to take the test at all if they did not want to. Nevertheless, since most students appeared to have taken the test seriously, useful analyses can be conducted.

DWST Speededness

Another concern about the experimental administration was whether students would have adequate time to complete the test. Of the 504 persons who submitted answer sheets, 455 completed the test, or 90 percent. Four of those who completed the test were instructors, however. Of the 500 students, four made no attempt to complete the test, responding to less than 10 items (two students) or working on only one of the two parts of the test (two students). If these four students are excluded from the total, 451 of 496 students (91 percent) completed the test. By Swineford's (1956) guidelines on test speededness, a test is considered unspeeded if: (1) 80 percent of examinees complete the entire test, and (2) 100 percent of examinees complete 75 percent of the test. The DWST clearly qualifies as unspeeded on the basis of the first criterion. But it would not qualify as unspeeded by the second criterion because 11 students completed less than 75 percent of the test, or 36 items. If it is reasoned, however, that these 11 students were not motivated and could have completed the test had they wanted to, the DWST could still be considered unspeeded.

Table 4 gives an analysis of completion rates within institutions and within course sections. Low completion rates tended to occur primarily in some sections of some institutions, suggesting that perhaps something occurred in those sections to reduce completion rates. Since different amounts of time were available in different institutions and sections, it is of interest to examine completion rates in relation to time available. Of those students who were allowed 50 to 60 minutes to complete the test, 244 of 252 (or 97 percent) did complete it. Of those allowed 45 to 47 minutes, 206 of 234 (or 88 percent) completed the test.

TABLE 4

DWST timing and completion rates

Institution Code/Section	*N** Total	*N** Completing Test	Testing Time	Completion Rate
A	16	16	50	100
B	219	194	45–60	89
B-1	26	24	55	92
B-2	24	23	60	96
B-3	26	15	45	58
B-4	19	19	55	100
B-5	19	16	45	84
B-6	16	15	45	94
B-7	23	19	47	83
B-8	25	25	45	100
B-9	14	14	45	100
B-10	14	12	45	86
B-11	13	12	45	92
C	55	50	45	91
C-1	10	9	45	90
C-2	15	15	45	100
C-3	4	4	45	100
C-4	15	11	45	73
C-5	5	5	45	100
C-6	6	0	45	100
D	86	181	60	97
D-1	70	68	60	97
D-2	57	57	60	100
D-3	59	56	60	95
E	10	9	45	90
Total	486	450	45–60	92
60 Minutes	210	204	60	97
55 Minutes	26	24	55	92
50 Minutes	16	16	50	100
47 Minutes	23	19	47	83
45 Minutes	211	187	45	89
50 + Minutes	252	244	50+	97
50 - Minutes	234	206	50-	88

*Excludes instructors, blank answer sheets, and students who withdrew.

DWST Item Difficulties

Table 5 shows that the largest proportion of DWST item difficulties were in a midrange of difficulty (40 to 70 percent of students answering correctly). For the Form V sample of examinees, 48 percent of items were in this range and, for the Form W sample of examinees, 40 percent. Only 16 percent of Form V items and 16 percent of Form W items were more difficult, while 35 percent of Form V items and 44 percent of form W items were less difficult. Since the DWST used a four-choice item structure, items with difficulties near or below $p = .25$ were probably answered predominantly by random guessing. Thus, in Table 5, the six items in the 10 to 30 percentage correct range may have been the result of guessing. Overall, the average p value for Form V was .61 and that for Form W was .64. The optimum average p value for a test with a four-choice item structure is approximately .75, since 25 percent of correct responses could result from guessing alone, and since .50 would be optimum for a test which allowed no guessing. This would suggest that the DWST is slightly more difficult than would be optimum, but difficulty level is readily adjusted.

TABLE 5
DWST item difficulties

Percentage Correct	Examinee Sample			
	Form V		Form W	
	N	Percent	*N*	Percent
Range				
90 +	5	10	5	10
80 - 90	7	15	7	15
70 - 80	5	10	9	19
60 - 70	11	23	7	15
50 - 60	4	8	11	23
40 - 50	8	17	1	2
30 - 40	2	4	2	4
20 - 30	5	10	5	10
10 - 20	1	2	1	2
< 10	0	0	0	0
Average				
61 (FormV)				
64 (FormW)				

DWST Reliability

Table 6 summarizes reliability analyses conducted of the DWST, its subscores, and its two parts. Coefficient Alpha based on the Form W sample was .76 and Coefficient Alpha based on the Form V sample was .70. Alpha estimates for the Part 1 and Part 2 scores are of special interest because of what appears to be a positioning effect. Part 1 of Form V, the 23 items about Valerie, produced an alpha estimate of .45 even though the same 23 items, as Part 2 of Form W, produced an alpha estimate of .68. A similar but less dramatic effect occurred for Part 1 of Form W, the 25 items based on Winfield. When these same 25 items were ordered second for Form V, the alpha estimate increased from the .60 obtained for Part 1 to .66. Since the higher alpha estimates both occurred for Part 2, it could be that examinees feel more comfortable with the test format after they get to the second half of the test.

Because alpha estimates of reliability are lower-bound estimates (Lord & Novick, 1968, p. 87), it is of interest to consider what might be upper-bound estimates of reliabilities for Forms V and W. If one takes the Part 2 alpha estimates and applies the Spearman-Brown prophecy formula (assuming the two parts are of equal length), an upper-bound estimate for Form V would be .80 and an upper-bound estimate for Form W would be .81.

TABLE 6

DWST subscore, part, and total reliabilities

Score	Items	Coefficient Alpha	
		Form V	Form W
Organization	15	0.33	0.44
Transition	11	0.41	0.46
Strategy	7	0.43	0.38
Conciseness	7	0.33	0.34
Clarity	9	0.35	0.34
Part 1	*	0.45	0.60
Part 2	*	0.66	0.68
Total	48	0.70	0.76

*Form V has 23 items in Part 1 and 25 in Part 2; Form W has 25 items in Part 1 and 23 items in Part 2.

The alpha estimates for the subscores are, of course, somewhat less than those for the total test. It is interesting, however, that some of the subscore estimates for Form V are almost as high as the .45 estimate for Part 1 of Form V. Why this might be is not clear, but one suspects either that Form V is not unidimensional or that something odd may have occurred within the sample that took Form V. Since Form W consists of the same items as Form V, but in a different order, the hypothesis of a dimensionality seems less likely than one related to the nature of the sample.

As noted previously, there are a number of reasons why one would expect relatively low reliabilities for a test like the DWST:

1. The construction of the test as two testlets results in what is known as local dependence (see Wainer & Thissen, 1996).

2. The use of a four-choice item structure, rather than the usual five-choice structure, increases the probability of guessing correct answers.

3. The inherent inefficiency of some of the items, as well as the inefficiencies resulting from some imperfect prototype items, resulted in fewer effective items in the test.

4. No effort was made to limit the items in the test on the basis of homogeneity, so it is quite likely that the items are relatively heterogeneous.

Questionable Items

A total of six of the 48 items of the DWST were identified as questionable because of p-values near or below .25, very low biserial correlations, or low p-values in an extremely high-scoring group. A summary of the analyses made of these items is given in Table 7. These items were double-keyed in an attempt to make them more effective, but Table 7 shows that the double-keying usually only made the items less difficult.

The p_{high} figures in Table 7 were computed from a subsample of nine examinees who responded correctly to at least 40 of the 48 items on the test. Three of these examinees were instructors in one of the participating institutions. When these items were single-keyed, five of the six items were answered correctly by only one-third of the top-scoring examinees. One of the six items was answered correctly by five of the nine top-scoring examinees, but its overall p-value was only .25. The r_{bis} correlations (biserials) were computed as the correlation between the item score and the total score, with that item removed, and then averaged for the two test forms. The p-values shown were also the average for the two forms.

The double-keying was conducted by identifying the option selected most often as correct and counting that as correct as well as the keyed option. This double-keying, of course, increased the p-values for all six items.

The biserial correlations for only three of the items were increased by double-keying, however, and for two items the biserial correlations decreased.

Alpha reliability estimates were made both after these items were removed from the test and after they were double-keyed. Removing these six items reduced the alpha for Form V from .70 to .66 and for Form V from .76 to .62. Double-keying the items had about the same effect: The Form V alpha estimate was reduced from .70 to .66, and the Form W alpha estimate was reduced from .76 to .61.

TABLE 7
*Questionable items**

Item	Single-keyed			Double-keyed		
	p	p_{high}	r_{bis}	p	p_{high}	r_{bis}
W-3/V-26	0.43	0.33	0.12	0.80	1.00	0.09
W-20/V-43	0.20	0.33	0.13	0.64	0.67	0.13
W-21/V-44	0.25	0.56	0.20	0.85	1.00	0.36
W-41/V-16	0.45	0.33	-0.06	0.82	0.78	0.10
W-46/V-21	0.14	0.33	0.14	0.63	0.78	0.02
W-47/V-22	0.25	0.33	0.08	0.65	0.89	0.14

*These items were revised in January 1997, but no new data have been collected for them. The revised items are included in the Appendix A revisions of Forms W and V.

Predictive Validity of the DWST

This section compares the validity of the DWST with that of the UGPA and the LSAT for predicting writing course grades and average law school grades. To make these comparisons possible, the data collected for the current study were matched with data of LSAC. In some cases, no match was made because the LSAC identification numbers provided by the participating law schools did not match those on record at LSAC. Additionally, some cases were excluded because LSAT scores were on an old score scale that could not be compared with the current score scale. Other cases were excluded because students withdrew from courses and thus had no course grades. Table 8 summarizes the data available for predictive validity analyses.

Correlational Analyses

Table 9 compares correlations between UGPA, LSAT scores, and DWST scores and law school writing course grades. For the DWST, correlations were computed for both the single-keyed version and the version with double-keying of six items. The correlations vary considerably for each institution, but the average of the correlations weighted by the number of cases analyzed for each institution indicates that the DWST was the best predictor of writing course grade (.28–.29), the UGPA second best (.23), and the LSAT worst (.21). Double-keying of the six questionable items on the DWST made little difference in predictive validity. These average DWST correlations with writing course grades are not quite as high as the .36 figure obtained by Olsen (1956) using a 30-minute editing test (human scored), but they are slightly better than the .26 correlation obtained by Pitcher (1962) using a 30-minute editing test and a .23 correlation using a combining sentences test (both human scored).

TABLE 8

Number of cases, means, and standard deviations for UGPA, LSAT, and DWST in five law schools

	Institution				
Variable	A	B	C	D	E
UGPA					
N	16	192	49	178	9
Mean	3.36	2.93	3.25	3.06	3.30
S.D.	0.48	0.46	0.43	0.41	0.25
Minimum	2.30	1.79	2.22	2.07	2.84
Maximum	3.95	3.96	3.83	3.95	3.76
LSAT					
N	16	193	49	177	9
Mean	162	150	157	155	153
S.D.	6.62	5.76	6.3	5.22	2.98
Minimum	147	130	142	141	150
Maximum	175	165	176	175	159
DWST(Single)					
N	16	193	51	181	9
Mean	33.2	28.5	33.1	30.4	27.8
S.D.	3.96	3.34	4.34	4.58	3.77
Minimum	25	12	23	8	24
Maximum	41	37	42	41	34
DWST(Double)					
N	16	193	51	181	9
Mean	36.4	31.5	35.7	33.4	31
S.D.	4.44	4.45	4.15	4.57	4.24
Minimum	26	15	26	11	25
Maximum	42	41	45	44	37

TABLE 9

A comparison of UGPA, the LSAT, and the DWST as predictors of law school writing course grade

		Correlations			
Institution	*N*	UGPA	LSAT	DWST (Single)	DWST (Double)
A	16	0.29	0.48	0.71	0.67
B	188	0.20	0.00	0.24	0.24
C	48	0.29	0.40	0.14	0.23
D	170	0.21	0.39	0.34	0.34
E	9	0.64	-0.25	-0.05	-0.09
Weighted average		0.23	0.21	0.28	0.29

Table 10 makes the same comparison when the criterion is changed to law school grade-point average. As in the prediction of law school writing course grade in Table 9, the correlations vary considerably across institutions. The weighted average of these correlations, however, indicates that the LSAT was the best predictor of law school GPA (r = .37), the DWST second best (.31–.32), and UGPA third best (r = .19). As was the case of predicting writing course grade, the double-keying of six questionable items of the DWST made little difference.

With the exception of the UGPA correlation, these average correlations with law school GPA are similar to the average for a large number of studies summarized by Schrader (1976). A total of 150 different law