DOCUMENT RESUME

ED 482 926 TM 035 401

AUTHOR	Hambleton, Ronald K.; Sireci, Stephen G.; Swaminathan, H.; Xing, Dehui; Rizavi, Saba
TITLE	Anchor-Based Methods for Judgmentally Estimating Item Difficulty Parameters. LSAC Research Report Series.
INSTITUTION	Law School Admission Council, Newtown, PA.
REPORT NO	LSAC-CTR-98-05
PUB DATE	2003-10-00
NOTE	48p.
PUB TYPE	Reports - Research (143)
EDRS PRICE	EDRS Price MF01/PC02 Plus Postage.
DESCRIPTORS	*Adaptive Testing; College Entrance Examinations; *Computer Assisted Testing; *Difficulty Level; Estimation (Mathematics); Field Tests
IDENTIFIERS	*Anchor Tests; *Item Parameters; Law School Admission Test

ABSTRACT

The purposes of this research study were to develop and field test anchor-based judgmental methods for enabling test specialists to estimate item difficulty statistics. The study consisted of three related field tests. In each, researchers worked with six Law School Admission Test (LSAT) test specialists and one or more of the LSAT subtests. The three field tests produced a number of conclusions. A considerable amount was learned about the process of extracting test specialists' estimates of item difficulty. The ratings took considerably longer to obtain than had been expected. Training, initial ratings, and discussion took a considerable amount of time. Test specialists felt they could be trained to estimate item difficulty accurately and, to some extent, they demonstrated this. Average error in the estimates of item difficulty varied from about 11% to 13 %. Also the discussions were popular with the panelists, and almost always resulted in improved item difficulty estimates. By the end of the study, the two expected frameworks that developers thought they might provide test specialists, had merged to one. Test specialists seemed to benefit from the descriptions of items located at three levels of difficulty and from information about the item statistics of many items. Four appendixes describe tasks and contain the field test materials. (Contains 8 tables and 18 references.) (SLD)

LSAC RESEARCH REPORT SERIES

■ Anchor-Based Methods for Judgmentally Estimating Item Difficulty Parameters

Ronald K. Hambleton
Stephen G. Sireci
H. Swaminathan
Dehui Xing
Saba Rizavi

University of Massachusetts at Amherst

■ Law School Admission Council
Computerized Testing Report 98-05
October 2003

A Publication of the Law School Admission Council

LSAC RESEARCH REPORT SERIES

Anchor-Based Methods for Judgmentally Estimating Item Difficulty Parameters

Ronald K. Hambleton
Stephen G. Sireci
H. Swaminathan
Dehui Xing
Saba Rizavi

University of Massachusetts at Amherst

Law School Admission Council
Computerized Testing Report 98-05
October 2003

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 201 law schools in the United States and Canada.

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

Table of Contents

Executive Summary

The Law School Admissions Council (LSAC) is currently investigating the feasibility of computerized adaptive testing. One of the important lessons learned from the recent controversy between the Graduate Record Examination Board and Stanley H. Kaplan Educational Centers Ltd is that large item banks will be needed to support computerized adaptive testing.

One dilemma for the LSAC if computerized adaptive testing is adopted is as follows: large test taker samples of candidates are highly desirable in field testing new items (large samples lead to precise estimates of the item statistics that are very important in implementing a computerized adaptive test), while at the same time, large test taker samples also result in more item exposure that can lead to a loss of item security.

One promising idea for lowering the desired size of test taker samples in item statistics estimation would be to combine the judgments of test specialists about the item statistics along with actual field-test data into a Bayesian item parameter estimation procedure where the information provided by the test specialists serves as a prior distribution. Test taker sample sizes can be reduced because the information loss due to smaller sample sizes is replaced by the information about the item statistics provided by the judgments of test specialists or other persons with knowledge about the test items. A Bayesian item parameter estimation procedure is simply a formal statistical way to combine both the information provided by the test takers and the information provided by the test specialists to arrive at the item statistics.

The purposes of this research study were to develop and field test anchor-based judgmental methods for enabling test specialists to estimate item difficulty statistics. The basic idea of anchor-based methods is that test specialists are provided with a frame of reference for making their judgments—descriptions of items at three levels of item difficulty, and/or many previously calibrated test items along with their item statistics. The task of judging the difficulties of new items amounts to matching the new items to those already calibrated and determining where the new items seem to fit in terms of their difficulty.

The study consisted of three related field tests. In each field test we worked with six Law School Admission Test (LSAT) test specialists and one or more of the LSAT subtests. In the first, we worked with both reading comprehension and logical reasoning; in the second, analytical reasoning; and in the third, analytical reasoning and reading comprehension. In the typical field test, panelists were trained in one of the variations of an anchor-based method for judging item difficulty statistics, then they were given test items to judge. In addition, panelists were informed during training about the factors that typically influence the difficulty of test items: the amount of negation, sentence and paragraph length; location of relevant text; the presence of effective distractors in the item; the novelty of the problem, and so on. Because of a strong belief that discussion among test specialists was valuable in the judgmental process, after panelists provided their independent ratings, discussions always took place about the test specialists' estimates, then they were given the opportunity to revise their ratings if they felt they could improve the estimates.

The three field tests were simply that—an opportunity to try out methods, receive feedback from test specialists on their likes and dislikes about the process, and collect ratings data that could be thoroughly analyzed. Because the items that were being judged by test specialists had been used in previous administrations of the LSAT, their item statistics were known to the researchers and could be used to evaluate both individual test specialist ratings and the first and second group estimates of the item statistics.

The three field tests produced a number of conclusions. A considerable amount was learned about the process of extracting test specialists' estimates of item difficulty. The ratings took considerably longer to obtain than had been expected. Training, initial ratings, and discussion took a considerable amount of time. For example, six hours might be needed to train test specialists and to obtain ratings on 20 test items.

Test specialists felt they could be trained to estimate item difficulty accurately, and, to some extent, they demonstrated this. Average error in the estimates of item difficulty varied from about 11% to 13%. Also, the discussions were popular with the panelists and almost always resulted in improved item difficulty estimates.

We began the study thinking there were at least two frameworks with which we might provide test specialists to improve their item difficulty estimates. By the end of the study, both methods had merged into one. Test specialists seemed to benefit from both the descriptions of items located at three levels of difficulty and from information about the item statistics of many items. Any future work in this area should probably combine both methods.

We completed the project with the feeling that the results were encouraging but would be better with improved training. What we learned was that the test specialists had developed many skills for themselves about what makes items hard and easy, and, therefore, the test specialists could be used more effectively than they were in the study to develop the descriptions of items at three levels of item difficulty and to develop rules for judging test items. With more research and development, we could see a training program for test specialists emerge that would prepare them for judging item statistics as a regular part of their work. This program would build not only on some of the research carried out in this study, and the relevant

literature, but also the insights and experiences of the test specialists themselves. We also share the test specialists' view that there would be some general principles for judging item difficulty but also specific principles linked to the three major areas covered by the test. At least some of the principles of judging item difficulty are specific to the reading comprehension, analytical reasoning, and logical reasoning subtests.

Introduction

One of the important lessons learned from the recent controversy between the Graduate Record Examination (GRE) and Stanley H. Kaplan Educational Centers Ltd is that large item banks will be needed to support computerized testing programs. The cost of item development can be very high for agencies such as the GRE (or Educational Testing Service [ETS]) or the Law School Admission Council (LSAC), especially when cognitively challenging items are needed. Another time-consuming and costly activity is the field testing of new items to obtain stable item statistics, which are needed in the test item selection process to ensure that parallel tests can be constructed or nonparallel tests can be properly equated (as in the case of computer adaptive tests).

But there is a dilemma for testing agencies such as the LSAC: On the one hand, large samples of test takers for field testing new or revised test items are highly desirable, but on the other hand, large test taker samples are associated with higher financial costs and a potential loss of item security (the basic principle is that the more candidates who see an item during a field test, the less likely it is to remain secure). One potential solution to the dilemma might be to use test specialists to estimate item statistics. There is some evidence in the measurement literature that test specialists are capable of estimating item difficulties with reasonable accuracy (e.g., Chalifour & Powers, 1989). Such a strategy could lower the costs associated with obtaining item statistics and ensure item security, at least until items become a part of a regularly administered computer-based test.

One promising idea for obtaining item statistics might be to combine the judgments of test specialists into a Bayesian item parameter estimation procedure where the information provided by the test specialists serves as a prior distribution. The combination of the prior beliefs about the item statistics with item response data from a reduced sample of test takers may function about as well as item statistics based upon considerably larger test taker samples. This point is demonstrated in the paper by Swaminathan, Hambleton, Sireci, Xing, and Rizavi (1999), (see Hambleton & Swaminathan, 1985; Lewis & Sheehan, 1990; Mislevy, 1986, 1988; Mislevy & Bock, 1990; Mislevy, Sheehan, & Wingersky, 1993; Swaminathan & Gifford, 1982, 1985, 1986).

At the same time, the use of test specialists to estimate item statistics is no guarantee of success, since the balance of the research evidence suggests that test specialists are not especially good at estimating item difficulties or other item statistics (Hambleton, Bastari, & Xing, 1998). But our reading of the research suggested that the methods used in many of these research studies were often inappropriate or poorly implemented. Panelist training and practice were often minimal, confusing, or used unknown scales for estimating item difficulties (e.g., panelists might be asked to provide estimates of item statistics on the ETS delta scale, 1 to 25, or the latent-ability scale, -3.0 to +3.0), and rarely were panelists given a frame of reference for their judgments (such as the item statistics of a representative sampling of items from the item bank of interest, or detailed information about the population of candidates).

The purposes of this research study were to develop and field test anchor-based judgmental methods for enabling test specialists to estimate item difficulties. The basic idea of anchor-based methods is that panelists are provided with a frame of reference for making their judgments. We chose to focus on a single item statistic (item difficulty) in our research because of evidence in the literature that test specialists were rarely able to estimate item-discrimination indices. Our goal was to overcome many of the shortcomings of previous research in the area. Three field tests were conducted. These field tests will be considered separately below, and then conclusions will be offered in the final section of the paper. First, however, some of the relevant research on estimating item difficulties will be considered (see also, Bejar, 1983; Chalifour & Powers, 1989; Freedle & Kostin, 1993; Sheehan & Mislevy, 1990, 1994; Thorndike, 1982).

We would like to thank Peter Pashley and Lynda Reese for their assistance in conducting the study and, most importantly, we want to thank the test specialists at the LSAC for participating in the training and the item-rating process.

Overview of the Research Findings

The measurement literature on estimating item statistics reveals some inconsistent findings. Some research studies indicate that panelists are unable to do the task and others have produced considerably better results (e.g., Chalifour & Powers, 1989). The research findings suggest too that panelists are much better at estimating item difficulties than they are item discrimination indices. What follows are some highlights of our literature review:

1. With training, raters get better. (They develop more complex models of item difficulty.)

2. Global ratings are usually better predictors of item difficultly than component models.

3. In many studies, raters are asked to use nonfamiliar scales (delta, latent ability, 1 to 10, etc.) in judging item difficulty. These scales are often problematic for panelists. They have no frame of reference for a delta of (say) 11.3.

4. Predictors of item difficulty (i.e., the factors which influence item difficulty) vary as a function of the item type (e.g., what works with verbal analogies may not work with reading comprehension or analytical reasoning).

5. Securing panelist agreement about item difficulty is preferable (the estimates are better) to averaging totally independent ratings and using the averages as the estimates.

6. Some types of items are easier to judge than others (e.g., quantitative items are easier to judge for item difficulty than verbal items).

7. Administrative details can be influential in item difficulty (e.g., Is the test slightly or moderately speeded? Are test takers permitted the use of calculators?). Also, panelists may underestimate the difficulty of items when items appear towards the end of a slightly or moderately speeded test.

8. Panelists need to be trained to look at (a) structural characteristics (e.g., item format, number of answer choices, number of operations needed to solve the problem); (b) surface features (e.g., sentence lengths, use of uncommon words); and (c) the psychological component (e.g., which cognitive skills are needed to answer the question?) of test items. The common shortcoming of panelists is to focus their judgments of item statistics on only one of the dimensions.

9. The test taker population is important (e.g., the age, gender, and ethnic group of test takers; levels of test anxiety and motivation to perform well).

10. Item placement in a test is important (especially if a test is speeded). Item statistics may vary as a function of item placement.

11. There is considerable evidence to suggest that item difficulty levels can be predicted, but predicting item discrimination has not been successful at all.

12. Average ratings of item difficulty statistics across panelists (i.e., raters) are much more highly correlated with actual item difficulties than individual panelist ratings. (The conclusion that follows is: Use multiple panelists in judging item statistics.)

All of the features that were relevant for this study (6, 9, and 10 were not judged as relevant) were incorporated into the design of the three field tests.

There are also some specific suggestions in the research literature for the factors affecting item difficulty.

1. Negations: the greater the number, the more difficult the test item.

2. Referential expressions: the greater the number, the more difficult the test item.

3. Vocabulary: the more multisyllabic words and hard words used, the more difficult the test item.

4. Sentence and paragraph lengths affect item difficulty.

5. Abstraction of text: the more abstract the text, the harder the test items will be.

6. Location of relevant text: apparently, when the relevant material is in the middle of a passage, the item is harder for the test taker.

7. The levels and numbers of cognitive skills needed to answer a test item affect item difficulty. Problem complexity is an important factor in determining item difficulty.

8. The novelty of the item format to candidates will make test items more difficult.

9. Placement of the item in the test: items appearing late in a test are more difficult than when they appear early or in the middle of a test.

10. Closeness of the best distractors to the correct answer: (This is especially important.) an item with a near correct distractor is more difficult than a test item with a correct answer and three or four answer choices that can easily be distinguished from the correct answer.

Ideally, factors like those above should be incorporated into the training of panelists to estimate item statistics.

Field Test One

In view of the often disappointing results obtained from using test specialists to estimate item statistics, two new methods for estimating item difficulties were developed. These new methods are related to some advances in setting standards on tests and reporting test scores, and so there were good reasons for expecting improved results from the judgmental task. The ideas include (1) providing a frame of reference for panelists to complete their rating tasks, and (2) providing feedback to panelists and discussion to provide a basis for panelists to review their judgments prior to finalizing them.

One other innovation in our work was the use of an understandable scale on which to estimate item difficulties (the *p*+ scale).

Design

The main factors for inclusion in the training of panelists were identified (a summary is provided below). Basically, factors affecting item difficulty were organized around three major categories: the item itself (e.g., novelty of item format, reading difficulty, cognitive complexity, number of near-correct answer choices, item placement); administration (administrative mode, test speededness, etc.); and test takers (e.g., age, gender, ethnic group, test taker's test sophistication). Our literature review enabled us to train panelists on the factors that contribute to item difficulty. Panelists were made aware of the 10 specific factors above and others in estimating item difficulty levels.

Law School Admission Test Subtests

Two of the Law School Admission Test (LSAT) subtests were used in the field test: reading comprehension and logical reasoning. According to LSAC's book *The Official LSAT PrepTest® XVI* (Law School Admission Council, 1995, p. 2),

> The purpose of reading comprehension questions is to measure your ability to read, with understanding and insight, examples of lengthy and complex materials similar to those commonly encountered in law school work ... Reading comprehension questions require test takers to read carefully and accurately, to determine the relationships among the various parts of the passage, and to draw reasonable inferences from the material in the passage.

According to the same book (p. 4),

> Logical reasoning questions evaluate a test taker's ability to understand, analyze, criticize, and complete arguments. The arguments are contained in short passages taken from a variety of sources, including letters to the editor, speeches, advertisements, newspaper articles and editorials, informal discussions and conversations, as well as articles in the humanities, the social sciences, and the natural sciences. Each logical reasoning question requires the test

taker to read and comprehend the argument or the reasoning contained in the passage, and answer one or two questions about it. The questions test a variety of logical skills.

Panelists

The six panelists who participated in the training and item rating process were test specialists for the LSAC. Though they were generally familiar with all of the subtests on the LSAT, all were specialists with respect to only one of the three subtests. Some of the specialists indicated later that they would have preferred to rate only items from the subtest they normally addressed in their work.

Anchor-Based Methods

Anchor-based and mapping methods. Before gathering the judgments of item difficulty, a two-hour interactive training session was held to describe item attributes related to item difficulty. The purpose of this training session was to inform the panelists of item characteristics that previous research demonstrated were related to item difficulty. Examples were provided earlier.

Two procedures were used to gather the judgments of item difficulty. The first method, called the *anchor-based method*, featured a discussion of the attributes of items at three points along the item difficulty *p+* scale: 0.25, 0.50, and 0.75. After discussing the attributes of "difficult," "moderate," and "easy" items around these three scale points, a booklet of 27 LSAT reading comprehension items was distributed to the panelists. The empirical item difficulties (i.e., actual item *p+* values) for 6 of these 27 items were revealed to the panelists. These six items were chosen to be representative of items characterizing each of the three anchor points of the item difficulty scale.

The first task for the panelists was to place each item into one of the four categories delineated on the item difficulty scale (< .25, = >.25 but < .50, = > .50 but < .75, or = > .75; see Appendix A for the training materials). Next the panelists were asked to provide an exact estimate of the difficulty for each item, ranging from .00 (no one would get the item correct) to 1.00 (everyone would get the item correct). After the panelists had completed these ratings, the individual ratings were shared with the group, item by item, and the panelists were asked to explain their item difficulty ratings. The panelists were then given an opportunity to revise their initial difficulty estimates based on the group discussions.

The second rating method, called the *item mapping method*, presented the panelists with an entire booklet of 27 LSAT logical reasoning items. The empirical item difficulty (*p+*) for each item was included in the booklet. The panelists were told to use these actual item difficulties as a reference for estimating the difficulties of other logical reasoning items. The item difficulties associated with these items essentially spanned the item difficulty scale, and so more "anchor points" were provided than in the anchor-based method. The panelists were then presented with a booklet of 21 logical reasoning items and were asked to provide difficulty estimates for each item. The panelists first provided their ratings individually, then met as a group to discuss their individual ratings. They were encouraged to revise their initial ratings based upon the group discussions as they regarded necessary.

Brief descriptions of the methods and how they were implemented follow.

Anchor-Based Method Steps

1. Review the three anchor point descriptions (chosen to be 0.25, 0.50, and 0.75 on the item difficulty scale for this field test) in terms of content, cognitive skills, item format, and a sample item or two (in total, item difficulty estimates for six items were given during the training), and so on.

2. Read each set of items (associated with a common stimulus such as a passage or problem statement) and then sort each item into one of the four categories: category one, .00 to .24, category two, .25 to .49, category three, .50 to .74, and category four, .75 to 1.00. Record your ratings on the Round 1 rating form. Estimate item difficulty and place this estimate in the column provided on the rating form. (Round 1)

3. Receive feedback on panelists' placement of items and item difficulty estimates, discuss this information, and ultimately revise the category placement and item difficulty estimates. (Round 2)

Item-Mapping Method Steps

1. Review the six test items mapped onto the item difficulty scale for this field test. Try to determine what makes some items more difficult or easier than others. (In actual fact, panelists

were given item difficulty estimates on 27 items, and these were to serve as a framework for their judgments on 21 items.)

2. Read each item or set of items (associated with a common stimulus such as a passage or problem statement) and then decide whether individual test items are harder or easier than those with known item difficulty levels. Estimate item difficulty for these new items and place this item difficulty estimate in the column provided on the rating form. (Round 1)

3. Receive feedback on panelists' item difficulty estimates, discuss this information, and ultimately revise your item difficulty estimates if you feel revisions are in order. (Round 2)

Because of a shortage of time, the anchor-based method was carried out on the reading comprehension subtest only, and the item mapping method was applied only to the logical reasoning items. The anchor-based method was implemented during a one-day workshop by one of the researchers. The item mapping method was carried out by the panelists without the aid of one of the researchers a short time after the one-day workshop.

Results

Reading Comprehension

Table 1 presents some descriptive statistics summarizing the results of the item difficulty ratings for the LSAT reading comprehension items. As described above, these item difficulty ratings were gathered using the anchor-based method. The statistics presented in Table 1 include the mean and median estimates of item difficulty (calculated over panelists) for the first round (initial ratings) and second round (revised ratings), and the difference between the empirical item difficulty and the mean revised item difficulty estimate for each item.

TABLE 1

Descriptive statistics for reading comprehension items (anchor-based method)

	"True" *p*+*	Mean *p*+ Round 1	Median *p*+ Round 1	Mean *p*+ Round 2	Median *p*+ Round 2	"True" *p*+ Minus Round 2 Mean
	0.36	0.52	0.50	0.47	0.45	-0.11
	0.37	0.37	0.36	0.36	0.36	0.01
	0.39	0.47	0.48	0.48	0.46	-0.09
	0.39	0.49	0.50	0.48	0.45	-0.09
	0.43	0.64	0.69	0.59	0.58	-0.16
	0.44	0.57	0.54	0.52	0.52	-0.08
	0.48	0.57	0.57	0.60	0.57	-0.12
	0.49	0.61	0.63	0.62	0.63	-0.13
	0.51	0.58	0.58	0.57	0.55	-0.06
	0.54	0.41	0.38	0.43	0.40	0.11
	0.55	0.71	0.75	0.71	0.75	-0.16
	0.59	0.67	0.63	0.65	0.63	-0.06
	0.60	0.72	0.73	0.78	0.78	-0.18
	0.60	0.62	0.60	0.62	0.60	0.02
	0.60	0.49	0.47	0.49	0.47	0.11
	0.61	0.41	0.43	0.41	0.43	0.20
	0.62	0.73	0.75	0.73	0.75	-0.11
	0.64	0.66	0.73	0.69	0.73	-0.05
	0.64	0.71	0.70	0.71	0.70	-0.07
	0.68	0.66	0.68	0.66	0.68	0.02
	0.69	0.57	0.57	0.56	0.57	0.13
Mean	0.53	0.58	0.58	0.58	0.57	-0.04

*Note that items have been presented in ascending order by "True" *p*+ value.

The results in Table 1 illustrate that the panelists rated two-thirds of the items to be easier than the actual item difficulty estimates showed. The data also show that although the average panelists' ratings did not change very much from Round 1 to Round 2, there was a consistent improvement in average difficulty estimates from Round 1 to Round 2. The group discussions improved the estimates. The group discussions also brought the panelists closer to consensus. The inter-rater reliability index was .71 for Round 1 and .87 for Round 2. It was also apparent that panelists' means typically moved in the direction of the true item *p*+ values (this was the case for 7 of 8 items where shifts in excess of .02 were noted).

In terms of the precision of the estimated difficulties, 9 of the 21 items had mean (Round 2) estimates within .10 of the actual item difficulties. The largest discrepancy was for item number 15, which was one of the few (7) items estimated to be harder than its empirical difficulty. The correlation between the empirical item difficulties and the mean and median estimated difficulties improved from Round 1 to Round 2. The correlation between "true" and mean estimated difficulty was .50 for Round 1 and .55 for Round 2. The correlation between the true and median estimated difficulty was .47 for Round 1 and .59 for Round 2. These correlations, and the data presented in Table 1, indicate that the panelists were somewhat successful in estimating the item difficulties. The higher correlation observed for the median difficulty estimates from round 2 suggest that some panelists provided better estimates than others.

Table 2 presents some summary statistics for each of the six panelists. The results presented in Table 2 confirm the improvement in difficulty estimations from Round 1 to Round 2, and illustrate the relative success in item difficulty estimation for the panelists.

TABLE 2
Summary of panelists' ratings: Reading comprehension (21 Items)

Round	Rater	Mean	SD[a]	r_{AE}[b]	Average D[c]	Average |D|[d]	Max D[e]	No. of D>.10
1	1	0.57	0.18	0.52	0.04	0.13	0.32	12
	2	0.55	0.17	0.17	0.02	0.15	0.34	13
	3	0.59	0.16	0.33	0.06	0.11	0.49	10
	4	0.63	0.20	0.11	0.10	0.20	0.46	18
	5	0.53	0.18	0.56	0.00	0.11	0.31	9
	6	0.60	0.14	0.22	0.07	0.13	0.39	12
Average		0.58	0.17	0.32	0.05	0.14	0.39	12
2	1	0.58	0.15	0.51	0.05	0.11	0.27	10
	2	0.55	0.16	0.30	0.02	0.13	0.34	11
	3	0.59	0.14	0.54	0.05	0.09	0.31	10
	4	0.61	0.17	0.20	0.08	0.17	0.41	17
	5	0.54	0.15	0.62	0.01	0.09	0.22	7
	6	0.59	0.12	0.46	0.05	0.10	0.26	12
Average		0.58	0.15	0.44	0.04	0.12	0.30	11
Average *p*+-value		0.53	0.11					

[a]Mean and SD refer to the estimated item difficulties for a panelist on a particular round. (In the Average rows, they are compiled over the panelists.)
[b]r is the correlation between actual and estimated item difficulties for a panelist on a particular round.
[c]Average D is the average of the differences between actual and estimated item difficulties. (We use this column to address bias in individual raters.)
[d]Average |D| is the average of the absolute deviations of differences between actual and estimated item difficulties for a panelist on a particular round.
[e]Max D is the maximum of the item differences between actual and estimated *p*+ values.

Logical Reasoning

Table 3 summarizes the results of the item difficulty ratings for the logical reasoning items. As described above, these item ratings were gathered using the item mapping procedure. The predominance of estimating items to be easier than their empirical difficulties did not occur. About half of the items (10) were judged to be easier than their empirical difficulties; the other half (11) were judged to be harder than their empirical difficulties. As with the reading comprehension ratings, the estimates improved from Round 1 to Round 2. The Round 2 mean difficulty estimates for 11 items were within .10 of their empirical difficulties; however, two items (numbers 3 and 10) were judged to be greater than .20 easier than their empirical difficulties. And again, it was noted that when there were shifts in the panelists' mean estimate (.02 or more), the shifts were consistently toward the actual item *p*-values (10 of 12 cases).

TABLE 3

Descriptive statistics for logical reasoning items (item mapping method)

	"True" *p*+*	Mean *p*+ Round 1	Median *p*+ Round 1	Mean *p*+ Round 2	Median *p*+ Round 2	"True" *p*+ Minus Round 2 Mean
	0.20	0.54	0.47	0.45	0.45	-0.25
	0.24	0.40	0.37	0.35	0.35	-0.11
	0.28	0.42	0.43	0.41	0.43	-0.13
	0.42	0.58	0.55	0.56	0.55	-0.14
	0.45	0.58	0.59	0.51	0.51	-0.06
	0.47	0.63	0.65	0.61	0.62	-0.14
	0.48	0.62	0.66	0.60	0.64	-0.12
	0.48	0.40	0.39	0.36	0.35	0.12
	0.50	0.57	0.58	0.57	0.58	-0.07
	0.51	0.53	0.55	0.49	0.50	0.02
	0.54	0.61	0.62	0.60	0.59	-0.06
	0.56	0.78	0.78	0.78	0.78	-0.22
	0.61	0.66	0.73	0.59	0.65	0.02
	0.72	0.68	0.70	0.68	0.69	0.04
	0.72	0.70	0.72	0.71	0.74	0.01
	0.77	0.65	0.64	0.63	0.64	0.14
	0.81	0.75	0.75	0.78	0.80	0.03
	0.81	0.79	0.78	0.80	0.81	0.01
	0.82	0.70	0.69	0.68	0.69	0.14
	0.83	0.81	0.83	0.83	0.83	0.00
	0.83	0.72	0.74	0.74	0.74	0.09
Mean	0.57	0.62	0.63	0.61	0.62	-0.03

*Note that items have been presented in ascending order by "True" *p*+ value.

The correlations between the empirical item difficulties and the average estimated item difficulties were much higher for the logical reasoning items than for the reading comprehension items. The correlation between the mean estimated difficulties and empirical difficulties was .81 for Round 1 and .84 for Round 2 (compared to .55 for the reading comprehension items). The correlations between the median estimated difficulties and the empirical difficulties were .83 for Round 1 and .84 for Round 2.

The inter-rater reliabilities were also higher for the logical reasoning ratings in both Round 1 and Round 2. The inter-rater reliability was .84 for the Round 1 ratings and .95 for the Round 2 ratings. Once again, the group discussions brought the panelists closer to consensus.

The results of the individual panelists' logical reasoning difficulty estimates are summarized in Table 4. The data reflect the higher correlation between the true and estimated difficulties across all panelists and reveal the relative successes of the panelists in terms of their difficulty estimates.

TABLE 4
Summary of panelists' ratings: logical reasoning (21 items)

Round	Rater	Mean	SD[a]	r_{AE}[b]	Average D[c]	Average |D|[d]	Max D[e]	No. of D>.10
1	1	0.64	0.14	0.75	0.07	0.12	0.32	12
	2	0.65	0.16	0.78	0.08	0.11	0.36	8
	3	0.62	0.14	0.33	0.04	0.18	0.46	18
	4	0.55	0.20	0.51	-0.02	0.13	0.65	10
	5	0.68	0.14	0.68	0.10	0.13	0.47	10
	6	0.60	0.18	0.59	0.03	0.15	0.32	12
Average		0.62	0.16	0.61	0.05	0.14	0.43	12
2	1	0.62	0.15	0.81	0.05	0.10	0.27	9
	2	0.62	0.15	0.88	0.05	0.09	0.23	6
	3	0.59	0.15	0.64	0.02	0.14	0.27	14
	4	0.55	0.17	0.68	-0.02	0.12	0.30	10
	5	0.64	0.14	0.81	0.07	0.11	0.30	8
	6	0.60	0.18	0.76	0.03	0.11	0.29	10
Average		0.60	0.16	0.76	0.03	0.11	0.28	10
Average *p*+ value		0.57	0.20					

[a]Mean and SD refer to the estimated item difficulties for a panelist on a particular round. (In the Average rows, they are compiled over the panelists.)
[b]r is the correlation between actual and estimated item difficulties for a panelist on a particular round.
[c]Average D is the average of the differences between actual and estimated item difficulties. (We use this column to address bias in individual raters.)
[d]Average |D| is the average of the absolute deviations of differences between actual and estimated item difficulties for a panelist on a particular round.
[e]Max D is the maximum of the item differences between actual and estimated *p*+ values.

For the logical reasoning items, the empirical difficulty estimates and the deviations between mean difficulty estimates and true difficulties were highly correlated, but this correlation was reduced from Round 1 (.82) to Round 2 (.72). This finding indicates that the panelists' estimates of the easier logical reasoning items were relatively poorer than their estimates of the harder items.

Comparing the data across Tables 1 through 4, it is evident that more precise difficulty estimates were provided for the logical reasoning items. However, because different procedures were used for the reading comprehension (anchor-based procedure) and the logical reasoning items (item mapping procedure), these results do not tell us if the improved results are due to item type (reading comprehension vs. logical reasoning) or estimation procedure (anchor based vs. item mapping).

Evaluation

After completing the item mapping-procedure, the panelists were given an evaluation form (similar to the form shown in Appendix B and used in the second field test). This evaluation form was primarily open-ended and addressed questions about panelist confidence in the procedures; that is, whether they had enough training time and rating time, and asked them for suggestions for improving the procedures. Five of the six panelists completed the evaluation form.

In evaluating the amount of time devoted to training, three of the five panelists reported that the amount of training time was sufficient, and two reported that more time was needed. The panelists found the discussion of factors that influence item difficulty to be very helpful. They suggested that more time be spent on describing cognitive features of the items that contribute to difficulty as well as the nature and positioning of the keys and distractors. In terms of the anchor-based method, some panelists felt that more anchor points would have been helpful. Given the results of the item-mapping method, it appears increasing the number of anchor points would be helpful to panelists.

The panelists also provided information regarding the item characteristics they used to make their difficulty judgements. Some factors listed were: length and ambiguity of the item stems and item options, placement and level of abstraction of the key, overall difficulty level of the passage associated with an item, and the presence of negations in the item. The panelists drew on their differential experience with reading comprehension and logical reasoning items in making their difficulty ratings. However, the panelists reported that they had not reviewed the items used in this study previously, so familiarity with the specific items used in this study did not confound the results.

The panelists were consistent in praising group review of the initial item ratings. However, there was consensus that the time devoted to the group discussions was too short. (The time devoted to group review may be another factor affecting the differences between the anchor-based and item-mapping methods.)

A final question presented to the panelists was "Based on your experiences today, how confident are you that test specialists can provide useful estimates of item difficulty?" The panelists were asked to respond to this question using a six-point scale where the lowest point of one represented "not at all confident" and the highest point of six indicated "very confident." The panelists' ratings and comments reflected moderate-to-strong confidence in the procedures. The mean rating was 4.4 and the median rating was 4. One panelist reiterated that the validity of the procedures would be improved with additional training and experience.

What follows is a detailed summary of panelist comments to each open-ended question.

Question 1: For the anchor-based method, do you think the amount of training was adequate?

1. I did feel rushed, but I think I got the basic idea.

2. The examination of the anchors was useful, but I might have preferred less time on the general introductory material before that and more on anchor items and the actual estimating of difficulty.

3. Though the training we received was of quality, the exercise as a whole suffered due to lack of sufficient training time.

4. The amount of training time was sufficient.

5. Skimming passages to answer the questions made the items seem much harder than they were. The way I keyed the items for the actual rating was quite different from how I did it during the training.

Question 2: Please describe your opinions of the content of the training sessions. Was the description of factors that influence item difficulty helpful? Were the reviews of specific items helpful?

1. More time could be devoted to review of specific items with reference to the factors; that is, how does the particular item exemplify (or not exemplify) various of the factors.

2. Yes to both questions. I think that the two factors that are perhaps the most important in determining difficulty are skills, placement, and quality of distractors. I'd like to see more analysis of those in future sessions.

3. The factors we covered were immensely helpful, but, due to the nature of the differences between item types, it would have been useful to separate the factors and rank them according to their particular relevance to individual item types.

4. There were many factors presented. It's difficult to consider each factor for each item. Moreover, most of the factors seem relatively minor. It's the difficulty of the basic cognitive task that primarily determines item difficulty, I suspect. Most of the factors presented would probably make an item slightly more or less difficult.

5. It is hard to gauge the usefulness of the description of factors without knowing how well our estimates match the actual levels of difficulty. I did find them helpful in estimating the difficulty of items that I did not have strong intuitions about.

Question 3: Please indicate other factors or item characteristics that were NOT included in the training that you think would be helpful in the future for training panelists to estimate item difficulties.

1. For reading comprehension(RC): More about the passage itself; for example, Is it a science passage? How much of the passage must be comprehended to answer the question? I'd guess that the less you need, the easier the item.

2. I didn't think of any additional characteristics, just that more attention should be given to the skills and the kind of task required, as well as to the distractors. These affect items more than negatives or the level of vocabulary.

3. Subject matter of the particular item; that is, Is it a science item? A literature item?, and so on.

4. Item subtype. Closeness or accuracy of key. Position of close distractors relative to key position. Whether one can guess key without reading options.

5. Though we did not work on AR items, I thought it may be useful to include some factors we believe influence item difficulty (for any future work with that subtest):

   - number of conditions
   - complexity of conditions (number of logical connectives; i.e., and, or, if ...)
   - use of qualifiers
   - use of types vs. tokens (e.g., items that force one to distinguish between things true of a type and things true of specific tokens of that type)
   - the fit of the logic of the item and everyday contexts

Question 4: For the anchor-based method (RC items) do you think describing three scale points (i.e., .25, .50, .75) was sufficient? Would you like to see more scale points described?

1. Fine at start, perhaps as we get better at ratings, the scale could become finer.

2. The .50 anchors were useful, but given that very few items are harder than .25 or easier than .75, some additional anchor points between those extremes would be useful. Actually, there are some points higher than .75, but there were very few lower than .25, so something between .25 and .50 would be helpful.

3. I think more explanation is needed, especially in terms of discrimination between .25 and .24—borderline items.

4. Probably a few more scale points (about five in total) would be good. But really, I think we need five scale points for each item subtype. (There are 9 LR subtypes and 12 RC subtypes.)

5. I felt confident with three scale points. Any more would make this exercise equivalent to the item mapping task.

Question 5: Please indicate your opinion of the usefulness of the group discussions following the first round of item ratings. Did these discussions lead you to change many of your initial item ratings?

1. Very useful, elucidated other factors about the items that I had missed or discounted. I changed most of my ratings slightly (about .05 or so).

2. I think these discussions are quite useful. They give people with outlier estimates an opportunity to revise their estimates if they considered inappropriate factors. I did not change many of my initial estimates, but I still think the discussions are useful in confirming or disconfirming my initial estimates.

3. Unfortunately, due to interests of time, we did not engage in meaningful discussion, which I'm sure accounted for my only changing a few of my initial ratings.

4. They were useful and changed a few of my item ratings.

5. The group discussions were very useful. They reminded me to consider factors that I had dismissed in favor of subjective "gut" reactions. Discussion caused me to change 25% or so of my initial ratings.

Question 6: Was the amount of time allocated to group discussions sufficient? Did you feel rushed?

1. I felt rushed, not your fault though, the whole day was behind schedule.
2. Sometimes rushed, more time would have been useful.
3. I felt rushed. For some items, we even skipped discussion.
4. A little more time would have been good.
5. Yes, Yes.

Question 7: What characteristics of the items did you use to make your judgments of item difficulty?

1. RC: complicated stem—harder

   distractors easy to accommodate—easier

   need whole passage—harder

   goes beyond passage—harder

   harder for me to key—harder

   reading load of stem and options

2. I don't remember what I considered the first time (RC items), but the second time I used primarily "the skills required," "the nature of the task," and the difficulty of distractors. After these three characteristics, I considered presence of negations, abstract text, length and vocabulary levels of the item; that is, reading difficulty.

3. It was hard for me to separate my own biases from the factors we were given. Factors that I considered:

   - reading task of stem (word count, ambiguity, multiple levels of understanding, layering of cognitive tasks)
   - difficulty of passage (subject matter, word count, multiple ideas/arguments, amount of new terminology introduced)
   - placement of key
   - subtype category of item (inferences, analogies tend to be more difficult)
   - length/ambiguity of options

4. The characteristics I used were:

   - stimulus strength
   - level of abstraction of key
   - subjective feel
   - attractiveness of distractors subtype

- can figure out key without reading options
- comparison to exemplar items of the same subtype

5. I relied, perhaps too heavily, on intuition. After that I considered reading load, number of negatives, clarity of question stem, and naturalness of key.

Question 8: Please indicate your opinion regarding the utility of the item *p*-values (proportion-correct statistics) in the item mapping method (LR items). Was having more item *p*-values helpful in anchor-based method?

1. Yes, in fact I ended up using almost all of the items in the "known" section, not just the "anchors." This was a learning experience for me because I was familiar with IRT statistics.

2. It was useful to have more *p*+ values. On the other hand, I mostly used them ahead of time to establish a general sense of the range of *p*+ values in my head. I did not compare items to the items with known *p*+ values very often.

3. I'm not sure if I can categorize the *p*+ values as being helpful, but the LR anchor-based method was of no use. I must admit that I have very little familiarity with LR items, thus no sense of determining a pattern of difficulty across subtypes.

4. Item *p*+ values were not very helpful. It's very difficult to compare a new item to one for which I know the *p*+ value. When the items are very different, I felt like I needed a range of *p*+ values for each item subtype.

5. Again, because we did not know how well we did, it is hard to answer this to my satisfaction. Subjectively, it did seem very useful.

Question 9: Were the item-judgment tasks more difficult for one section of the test than the other?

1. I found it much harder to evaluate difficulty of RC than LR items. I've never done any significant recurring work in RC, while I've been doing LR test items for over five years.

2. I don't think so, though I think I got better with more experience.

3. The LR items were more difficult to gauge.

4. Yes. More difficult for RC (I have almost no experience reviewing RC test items).

5. I felt more confident about rating the RC items than LR items, not because of the difference in estimation formats, but because I am more familiar with RC test items.

Question 10: How familiar were you with the specific items used in this study? Had you reviewed these items previously? Were you able to remember some of their difficulty estimates from previous exposure to these items?

1. RC—no familiarity

   LR—I reviewed these items at some point, perhaps five years ago. I had not reviewed them since; I didn't remember their difficulty though.

2. I had never seen any of the test items before.

3. None of the test items were seen or reviewed by me previously.

4. These items were all completely new to me.

5. Not very familiar to me. I had reviewed only two in the LR and did not remember their difficulty estimates.

Question 11: What variations in the procedures used today would you recommend for the future?

1. I took some index cards, one for each known item, and wrote their number and point for each. Then I took cards of a different color, one for each unknown item, and put them into the deck, starting from the easy end, and I went through them until I could say "harder than that one but easier than the next one." It didn't work perfectly, but it was helpful.

2. Just more time analyzing known $p+$ values and doing actual estimates. (I realize, of course, that we did not have an unlimited amount of time).

3. I'm not comfortable enough with what was gained from the exercise to make recommendations for future use.

4. A range of anchors for each subtype would be helpful. Also, provide test specialists with average difficulty levels for each subtype (I suspect that there's significant variation between subtype). Ask panelists to estimate IRT-b rather than $p+$ values.

5. I would expose participants to many more anchor points; especially in LR, where these exemplars for each item subtype would have helped a lot. Also giving us feedback on how well we did is essential for training us to give useful estimates.

Question 12: Based on your experiences today, how confident are you that test specialists can provide useful estimates of item difficulty?

1. Most of us thought after doing the LR exercise that it would not be helpful to draw conclusions comparing the anchor vs. item mapping method. There are too many confounding factors. Another reason not to compare the methods based on this experience: We all seemed to use the known LR in different ways, not all of which were what you probably had in mind. Some people did not use them at all, feeling that their relative inexperience of LR would make these items basically useless.

2. As I said earlier, even a little bit of experience made me more comfortable. I think with more training and more experience, we could provide some very useful estimates.

3. I think the personal biases would be a hindrance. And, of course, we all judge items based on our own strengths and weaknesses in solving an item.

Conclusions

Many things were learned from the first field test. First, we were encouraged by the results. Despite the hurried and less-than-ideal training and the speed with which panelists were expected to provide their ratings, the evidence seemed clear that panelists were able to provide item difficulty estimates that might be useful in a Bayesian item parameter estimation process. We were also encouraged by the role of discussion. For the reading comprehension subtest, seven of the eight items showing shifts (of 0.02 or greater) were in the direction of the actual item difficulty estimates (see Table 1). For the logical reasoning subtest, 10 of 12 items showing shifts (of 0.02 or greater) were in the direction of the actual item difficulties (see Table 3). Clearly, the discussions and feedback were having their effect.

We also felt that there was no clear separation between the two methods. In the course of introducing the anchor-based method, statistics for a small sample of items were given. This amounted to a modest attempt at item mapping. This was especially important because the anchors we attempted to develop were not very sophisticated for the panelists who were experts on one of more of the three major components of the test.

In sum, the main suggestions for improving the item judgmental process included: extending the time for training, discussing the role of cognitive complexity on item difficulty during training, emphasizing the role of item distractors in item difficulty, and retaining and emphasizing group discussion. Nearly all of the panelists felt that with good and extended training, and with discussions and feedback during the process, they would be capable of judging item difficulties fairly accurately. The results in Tables 1 to 4 suggest that their perceptions on this point are accurate.

Field Test Two

Field test two was a modest revision of the first field test. The main goals of the second field test were (1) to improve the training, (2) slow down the rating process to provide panelists with additional time to complete their ratings, and (3) to compile some data on the third component of the LSAT—the assessment of analytical reasoning.

Design

Six test specialists were available for a day to participate in training and item difficulty estimation. The item difficulty estimations were made of items from the analytical reasoning subtest. Analytical reasoning items, according to *The Official LSAT PrepTest™ XVI* (Law School Admission Council, 1995, p. 3),

> are designed to measure the ability to understand a structure of relationships and to draw conclusions about the structure. The test taker is asked to make deductions from a set of statements, rules, or conditions that describe relationships among entities such as persons, places, things, or events. They simulate the kinds of detailed analyses of relationships that a law student must perform in solving legal problems. For example, a passage might describe four diplomats sitting around a table, following certain rules of protocol as to who can sit where. The test taker must answer questions about the implications of the given information; for example, who is sitting between diplomats X and Y.

Training was similar to the anchor-based method described in field-test one. In total, six practice items were used, then ratings were collected on 18 items. Similar materials as in the first field test were used (see Appendix A). Panelists were asked to identify the subtests they worked on for the LSAC.

Results

Analytical Reasoning

Descriptive statistics for the 18 analytical reasoning items are presented in Table 5. This table lists the empirical $p+$ value, the mean and median difficulty estimates for each round of ratings, and the difference between the $p+$ value and Round 2 mean rating for each item. Several findings are notable. First, there was little change in the mean and median difficulty estimates from Round 1 to Round 2. Second, the Round 2 mean difficulty estimates were within .10 of the 'true' empirical p-values for 10 of the 18 items. Those items with the largest discrepancies between the $p+$ values and the mean estimates tended to be relatively difficult items (i.e., $p+$ values less than .40). In fact, the three items with the largest discrepancies were the three most difficult items. Overall, the panelists tended to underestimate the item difficulties. That is, they tended to estimate items to be easier than they performed on the test. Two-thirds of the items were judged to be easier than their actual item difficulty estimates. Nevertheless, the finding that the majority of the difficulty estimates were close to the corresponding $p+$ values is encouraging.

TABLE 5

Descriptive statistics for analytical reasoning items (anchor-based method)

	"True" *p*+*	Mean *p*+ Round 1	Median *p*+ Round 1	Mean *p*+ Round 2	Median *p*+ Round 2	"True" *p*+ Minus Round 2 Mean
	0.25	0.50	0.53	0.48	0.44	-0.23
	0.25	0.43	0.40	0.46	0.45	-0.21
	0.30	0.52	0.50	0.52	0.51	-0.22
	0.32	0.43	0.42	0.47	0.45	-0.15
	0.38	0.56	0.59	0.57	0.54	-0.19
	0.39	0.57	0.67	0.58	0.61	-0.19
	0.39	0.49	0.45	0.46	0.45	-0.07
	0.39	0.43	0.38	0.43	0.40	-0.04
	0.41	0.57	0.66	0.57	0.60	-0.16
	0.44	0.42	0.44	0.43	0.44	0.01
	0.47	0.48	0.45	0.48	0.47	-0.01
	0.56	0.59	0.57	0.62	0.58	-0.06
	0.62	0.50	0.49	0.49	0.49	0.13
	0.62	0.64	0.64	0.64	0.66	-0.02
	0.65	0.61	0.67	0.65	0.68	0.00
	0.65	0.62	0.70	0.64	0.68	0.01
	0.70	0.62	0.68	0.62	0.64	0.08
	0.73	0.69	0.73	0.70	0.74	0.04
Mean	0.47	0.54	0.55	0.54	0.55	-0.07

*Note that items have been presented in ascending order by "True" *p*+ value.

The Round 1 and Round 2 item difficulty estimates provided by the panelists exhibited high positive correlations with the actual item difficulty values. However, there was little change in this correlation from Round 1 to Round 2 (r =.73 for Round 1 and r =.74 for Round 2). Although the revisions made by the panelists from Round 1 to Round 2 tended to be small, and were not always in the correct direction, the inter-rater reliability among the panelists increased substantially from .47 to .75 from Round 1 to Round 2. The difficulty estimates for one panelist (panelist #1) exhibited negative correlations with four of the five remaining panelists after Round 1, and with three of the five panelists after Round 2. If the ratings for this panelist are eliminated, the inter-rater reliabilities increased to .54 and .79 for rounds one and two, respectively. However, there is virtually no change in the magnitude of the mean difficulty estimate/actual item difficulty estimate discrepancies (e.g., only the same 10 items have estimates within .10 of the actual *p*+ values). Thus, the degree of consensus among the panelists does not appear to be related to the precision of the difficulty estimates. This observation is also evident from the similarity among the mean and median difficulty estimates. The largest mean-median discrepancy is .04 for Round 2.

Table 6 presents some summary statistics for each of the six panelists. The results presented in Table 6 confirm the small change in difficulty estimates from Round 1 to Round 2. However, for all panelists, the correlation between the item *p*+ values and their estimated difficulties increased from Round 1 to Round 2. The data in Table 6 also illustrate the relative successes among the panelists in accurately estimating the item difficulties. Panelists number 2, 4, and 6 provided item difficulty estimates that were relatively closer to the item *p*-values than those estimates provided by the other three panelists.

TABLE 6
Summary of panelists' ratings: analytical reasoning (18 Items)

Round	Rater	Mean	SD[1]	r_{AE}[a]	Average D[b]	Average |D|[c]	Max D[d]	No. of D>.10
1	1	0.62	0.13	-0.01	0.15	0.20	0.46	12
	2	0.51	0.16	0.32	0.04	0.15	0.35	9
	3	0.57	0.18	0.30	0.10	0.18	0.44	13
	4	0.49	0.18	0.73	0.01	0.10	0.28	7
	5	0.58	0.15	0.24	0.11	0.19	0.35	14
	6	0.45	0.16	0.63	-0.03	0.11	0.35	7
Average		0.54	0.16	0.37	0.06	0.19	0.37	10
2	1	0.62	0.12	0.05	0.14	0.19	0.46	13
	2	0.51	0.12	0.61	0.04	0.10	0.30	8
	3	0.57	0.15	0.40	0.10	0.15	0.40	11
	4	0.50	0.15	0.84	0.03	0.07	0.18	5
	5	0.57	0.13	0.31	0.10	0.17	0.32	13
	6	0.50	0.10	0.78	0.03	0.08	0.24	6
Average		0.55	0.13	0.50	0.07	0.13	0.32	9
Average p-value		0.47	0.16					

[a]r is the correlation between actual and estimated item difficulties for a panelist.
[b]Average D is the average of the differences between the estimated and actual item difficulties (i.e., estimate minus p-value).
[c]Average |D| is the average of the absolute deviations of differences between actual and estimated item difficulties for a panelist.
[d]Max D is the maximum of the item differences between actual and estimated p-values.

Panelists' Evaluation of the Second Field Test

The panelists participating in the second field test were asked to complete an evaluation form. This form contained 19 questions: 13 were open-ended and 6 were questions that could be answered on a five-point rating scale. The questionnaire is reproduced in Appendix B. Only four of the six panelists completed the evaluation form. For those questions inquiring about specific aspects of the field test, the panelists' evaluations suggested that: (1) the general orientation to the procedure was somewhat useful, (2) more training time should be given to the anchor-based method, and (3) the group discussions were useful. Two of the four panelists thought more time for group discussion should be given.

The panelists were asked to provide suggestions for improving the process. In addition to requests for more training and discussion time, the panelists suggested to: (1) provide more criteria for evaluating item difficulties (e.g., use linguists to develop a metric of syntactic complexity, consider number of sorting tasks involved in an item, consider nature and number of cognitive tasks involved in an item), (2) spend less time on discussion of general factors affecting item difficulty and more time on the methods used to rate the items, and (3) use fewer items for training and discussion to allow for more discussion time per item. When asked what criteria they used in making their item-difficulty ratings, the panelists reported using a mix of overall impressions of difficulty and more complex rules based on analysis of the items. One panelist reported that she/he primarily used "gut feeling" regarding item difficulty.

The other panelists reported using more intricate factors such as the cognitive complexity of the item, number of steps involved in answering the item, the complexity of the rules required for answering the item (one-dimensional versus relational), number of fixed assignments, nature and familiarity of the task required, plausibility of the distractors, and clarity of the item stimulus.

The panelists were also asked to provide ratings of how confident they were that their ratings would be useful for predicting item difficulty. On a six-point scale where "6" represented "very confident" and "1" represented "not at all confident," the responses were 2.5, 5, 2, and 3. When asked "how confident are you that test specialists can provide useful estimates of item difficulty?," the confidence ratings (on the same six-point scale) increased for three of the four panelists. Theses responses were 4.5, 5, 5, and 3. The discrepancy between these two sets of ratings may be caused by the fact that three of the four panelists spent relatively less time with the analytical reasoning section of the LSAT. When asked to rank-order the three LSAT sections in terms of "your involvement with each section," three of the four panelists ranked the analytical reasoning section third. These results suggest that the panelists are confident that test specialists can provide accurate ratings of item difficulty when they are rating items related to the LSAT sections with which they are most familiar.

Conclusions

The findings from this study were not quite as encouraging as those from the first field test. Panelists did a reasonable job of estimating item statistics, though the findings were not as positive as those from the first field test. We learned, too, that panelists definitely felt that they would do a better job of rating items on the subtest they work with on a daily basis.

In conclusion, we were encouraged by the panelists' predictions; we learned a bit more about how panelists went about the task of judging item difficulty; and, again, we learned that the training and rating tasks take more time than we expected. Finally, it was clear that panelists were not interchangeable across the subtests; panelists did not feel comfortable judging items they were not working with on a regular basis.

Field Test Three

The second field test was not completely successful. Again, despite adjustments in the time allocations for training and the item ratings process, several of the panelists felt that they were hurried. In addition, we were never completely successful in offering up anchor-point descriptions that met the needs of the panelists. In this final field test, we focused more on panelists developing the anchor descriptions for themselves (with analytical reasoning) and shifted our emphasis to the use of anchors based on single items. We provided more training of panelists (for the reading comprehension test items). Appendix C provides the details of our third field-test experience.

Design

This field test was conducted over two days. On day one, five panelists worked through a set of 17 analytical reasoning items. The original plan had been to have them work through a revised anchor-based procedure with the panelists producing the anchor descriptions based upon reviews of sets of test items. Unfortunately, this task took a considerable amount of time and therefore the work during the day was limited to producing the descriptions. Three of the panelists indicated that analytical reasoning was the test with which they were most familiar. The other two panelists indicated that the subtest was the second most familiar to them.

On day two, five panelists working with the reading comprehension subtest were trained on six practice items and they were then asked to rate 13 test items. The specific details on the training are contained in Appendix C.

Results

Analytical Reasoning

During the course of reviewing test items and estimating item difficulties and receiving feedback, the panelists produced the following descriptions of items at three anchor points.

- At $p+ = 0.75$—items that can be answered by checking or through elimination of distractors are found at this level; other items have distractors that can easily be eliminated. Items with fixed assignments or limited numbers of conditions would be found in this region of the p-value scale.
- At $p+ = 0.50$—items that are not completely straightforward might be found at this level. The conditions of these items "would not be too difficult." They might contain complex stimulus material but the questions would be easier. These items too might contain explicit or simple combinations of conditions (if/then...).
- At $p+ = 0.25$—items that are novel or include an element of abstraction would be located at this level. Other items would contain many possible permutations. "Complex mapping" would be used to describe these items. These items, too, would require candidates to work through a large number of cognitive tasks.

This group of five panelists was asked to read the first set of questions and then estimate item difficulties. Discussions followed the presentation of the item difficulties, and the goal was to construct the anchor descriptions and determine the basis for the miscalculations in item difficulty estimates. The results of the item difficulty estimates follow.

Group Estimate	Actual *p*+ value*
0.51	0.20
0.41	0.30
0.63	0.33
0.58	0.48
0.56	0.55
Average Absolute Difference = .166	

*Items have been presented in ascending order by Actual *p*+ value.

Results from the first five items were disappointing. The panel decided that they had underestimated item difficulty because of a failure to consider the facts that the first problem was (1) asymmetric and (2) not realistic.

The panel moved to the second stimulus and associated test items, and the following results were obtained.

Group Estimate	Actual *p*+ value*
0.35	0.30
0.36	0.33
0.50	0.39
0.47	0.40
0.49	0.44
0.56	0.45
0.59	0.73
Average Absolute Difference = .08	

*Items have been presented in ascending order by Actual *p*+ value.

Clearly the panelists did a considerably better job of estimating item difficulties with this second set of test items despite the fact that they felt the stimulus was not very well written and remained a bit ambiguous. Again, the panelists tended to underestimate the difficulty of the items. For five of the six items, they judged items easier than they actually functioned on the test.

With respect to the third and final stimulus and associated five test items, the following results were obtained.

Item	Group Estimate	Actual *p*+ value*
	0.57	0.39
	0.53	0.56
	0.53	0.58
	0.56	0.60
	0.54	0.63
Average Absolute Difference = .074		

*Items have been presented in ascending order by Actual *p*+ value.

Again, the ratings were better than either the first or second set of items and could suggest the value of practice and feedback. On the other hand, this particular set of items had middle-difficulty items, and

previous work in this study showed panelists tended to be more accurate in their estimates with middle-difficulty and easier items. Another possible explanation is that panelists were reluctant to stray too far from middle-difficulty estimates since (1) middle-difficulty estimates reduced the likelihood of a large error in their estimates, and, in the discussions, panelists preferred not to be outliers in the distribution of panelists based on the initial ratings.

With these ratings, we were aware of each panelist's major responsibilities. Three identified the analytical reasoning subtest as their primary responsibility, and the other two panelists indicated that analytical reasoning was their second area of expertise. Of interest was whether panelists most familiar with analytical reasoning would be more accurate in their ratings. The statistical information is shown below:

	Panelist				Panelist		
Item Set	1	2	5	Average	3	4	Average
1 to 5	0.122	0.172	0.205	0.166	0.180	0.196	0.188
6 to 12	0.107	0.116	0.103	0.109	0.109	0.124	0.117
13 to 17	0.068	0.124	0.072	0.085	0.044	0.080	0.062

From this analysis, there is no clear evidence that there were differences in the size of the prediction errors made by panelists who were most familiar with the analytical reasoning items versus those panelists who were not. On the other hand, panelists 3 and 4 indicated that they had seen the test items before (along with panelist 1). This may have influenced their results so the question about the role of subtest familiarity in item difficulty estimation remains unanswered.

Evaluation

The panelists were asked to identify factors that influenced their judgments of item difficulty but which were not part of the training. Two suggestions were offered: (1) more use of the "asymmetry" factor. This factor itself could be further broken down into subfactors that could be identified fairly mechanically, and (2) stimuli (problems) that utilize type-token distinctions or require mathematical calculations tend to be harder for candidates.

Three of the panelists indicated that they had seen these problems previously (panelists 1, 3, and 4). This was an important finding because it may explain the failure to find that panelists who worked on the analytical reasoning subtest did not make more accurate estimates than those panelists who worked principally on a different subtest. Both panelists in the second group (panelists 3 and 4) indicated that they had seen the test items before, which may have improved their item difficulty estimates.

Panelists were asked to offer variations on the procedures for estimating item difficulties that they had seen in the training: One panelist indicated that he/she thought that panelists would need to look at hundreds of items over many weeks to have the context for rating items successfully. Another felt that knowing the typical range of $p+$ values in the item bank would be helpful in providing a useful framework. Another panelist felt that the procedure used (looking at many items during the day) was exactly what was needed to improve item difficulty estimation.

A major criticism of the two previous field tests was the lack of time. On this day, four of the five panelists felt the pacing was about right. The remaining panelist felt he/she was rushed.

On the question about training, four of the five panelists indicated that they were pleased. Clearly the pacing was better and the discussions more relevant to the tasks. One panelist felt that each panelist should develop their own template for judging item difficulty, though it would be influenced by group discussion during the training.

Panelists were asked to comment on the discussions of item difficulty following the completion of first ratings. To a person, panelists felt this activity was useful (four of the five indicated that the discussions were very useful). As for the amount of time in discussion, three of the panelists felt the time was about right. The other two panelists felt the time for discussion was too little.

Panelists were also asked to comment on the influence of group discussions on their confidence of item difficulty estimates. The panelists were highly confident about their estimates as predictors of the actual item difficulty, except for one panelist who was somewhat confident. This panelist indicated that he/she was more familiar with the logical reasoning subtest.

Finally, the panelists indicated that the most influential factor in their ratings of item difficulty was their experience in working with the items on the job. The least important factor was the descriptors that were developed.

Reading Comprehension

Descriptive statistics for the 13 reading comprehension items are presented in Table 7. This table lists the actual (empirical) *p*+ values, the mean and median difficulty estimates for each round of ratings, and the difference between the *p*+ value and Round 2 mean rating for each item. The inter-rater reliabilities were as follows: Round 1 inter-rater reliability = .69, Round 2 inter-rater reliability = .86.

Again, the discussion phase of the process had the effect of improving the reliability of the judgments of item difficulty.

TABLE 7
Descriptive statistics for reading comprehension items (third field test)

	"True" *p*+*	Mean *p*+ Round 1	Median *p*+ Round 1	Mean *p*+ Round 2	Median *p*+ Round 2	"True" *p*+ Minus Round 2 Mean
	0.22	0.42	0.43	0.42	0.43	-0.20
	0.32	0.47	0.47	0.47	0.47	-0.15
	0.38	0.49	0.48	0.49	0.48	-0.11
	0.40	0.57	0.57	0.57	0.57	-0.17
	0.40	0.48	0.50	0.48	0.50	-0.08
	0.45	0.51	0.59	0.51	0.51	-0.06
	0.50	0.55	0.55	0.55	0.55	-0.05
	0.51	0.49	0.49	0.49	0.49	0.02
	0.51	0.49	0.49	0.51	0.49	0.00
	0.53	0.39	0.40	0.39	0.40	0.14
	0.61	0.43	0.43	0.42	0.61	0.19
	0.62	0.55	0.55	0.55	0.55	0.07
	0.63	0.39	0.37	0.36	0.37	0.27
Mean	0.47	0.48	0.49	0.48	0.49	-0.01

*Note that items have been presented in ascending order by "True" *p*+ value.

The results in Table 7 reflect the fact that only modest revisions in item difficulty estimates were made between the two rounds. There was also an unusual pattern of results. For 6 of the 13 items, the estimates were excellent. For 6 other items, the estimates were very poor. There was no obvious pattern in the good or poor estimates as a function of item difficulty.

The statistics in Table 8 seem to suggest that none of the panelists were especially good at estimating difficulties. First-round and second-round estimates were about 0.13 apart (in an absolute sense). In fact, for a number of panelists, the correlations between estimated item difficulties and actual item difficulties were negative. This was surprising and inconsistent with earlier findings.

TABLE 8
Summary of panelists' ratings for field test 3: reading comprehension (13 items)

Round	Rater	Mean	SD	r_{AE}[a]	Average D[b]	Average |D|[c]	Max D[d]	No. of D>.10
1	1	0.53	0.08	-0.09	-0.06	0.14	0.23	7
	2	0.49	0.09	-0.39	-0.02	0.14	0.36	7
	3	0.49	0.12	-0.46	-0.02	0.17	0.33	8
	4	0.45	0.10	0.21	0.02	0.11	0.35	6
	5	0.43	0.11	0.48	0.04	0.09	0.28	4
	6	0.55	0.11	0.04	-0.08	0.15	0.33	9
Average		0.49	0.10	-0.04	-0.02	0.13	0.31	
2	1	0.52	0.07	-0.15	-0.05	0.14	0.23	6
	2	0.48	0.09	-0.30	-0.01	0.13	0.31	7
	3	0.49	0.10	-0.44	-0.02	0.16	0.29	9
	4	0.46	0.08	0.06	0.01	0.11	0.33	6
	5	0.41	0.11	0.41	0.05	0.10	0.30	5
	6	0.51	0.10	-0.04	-0.04	0.14	0.33	9
Average		0.48	0.09	-0.08	-0.01	0.13	0.30	
Average p-value		0.47	0.12					

[a]r is the correlation between actual and estimated item difficulties for a panelist.
[b]Average D is the average of the differences between the estimated and actual item difficulties (i.e., estimate minus p-value).
[c]Average |D| is the average of the absolute deviations of differences between actual and estimated item difficulties for a panelist.
[d]Max D is the maximum of the item differences between actual and estimated p-values.

Evaluation

In general, the panelists felt that more time might have been used for training given the large number of item difficulty factors that could be discussed. Specific factors suggested include item subtypes, position of the key, the difficulty level of the passage, and the extent to which items depend on the global features of the passage. One of the panelists didn't have much to say about the training, but two others felt that training should begin with items and their statistics to help panelists determine a baseline. Also, more time could be allocated to these initial items to ensure that the relevant factors affecting item difficulty are estimated. One panelist was particularly interested in having anchor items for each subtype of item on the test. The idea was that generalizability of insights about item difficulty on one subtype of item is low for items from other subtypes.

As has been the case in all of our work, the group discussions were rated highly; however, two of the panelists felt that the discussions had modest influence on their ratings. The same two panelists thought that the group discussions may have even had a slight negative effect on their ratings, though not on their confidence in the ratings. For the majority of the panelists, the discussions had a positive effect on their confidence in the ratings. As expected the same two panelists thought that they were less confident about their estimates being useful in predicting actual item difficulty. Another rater who thought that he/she was less confident in this respect was the one who indicated in the beginning that he/she was not sure why his/her ratings didn't match with the group. It seemed that this panelist was a little unclear about the process and wanted some concrete examples of how items are affected by the various factors.

The most commonly used factors in making estimates of item difficulty were the general and specific factors provided by the trainer and also the factors that are experienced in editing LSAT items on the job. The use of anchors and descriptors was the least considered factor.

In response to some background checks on the panelists, it appeared that none of the panelists were working, principally, with the reading comprehension subtest. Five of the six panelists were most familiar with the logical reasoning subtest. One of the panelists indicated that he/she was most familiar with the reading comprehension subtest.

Conclusions

As in the previous field tests, a considerable amount was learned about judging item difficulties. From the work with the analytical reasoning items we learned that anchor descriptions can be produced by panelists, though considerable time is needed. It remains to be seen whether these descriptions would improve the judgmental ratings of item difficulty. Also, the use of training in a sequential fashion (train, rate items, receive feedback, rate items again, discuss the process, and then repeat through additional sets of items) was valued by panelists. Basically, panelists felt that more training and practice was the key to accurate judgments of item difficulty.

Again, we learned that panelists tend to overestimate item difficulty (or estimate that items are easier than they actually are). Presumably this point could be highlighted in training and practice sessions. Perhaps panelists who work with these test items on a regular basis (and hence, have lots of practice) forget how difficult these items might be for test candidates, or possibly panelists fail to consider the speededness of the LSAT subtests. In any case, this is a good area for further investigation. If the systematic bias in the estimation process could be removed, the estimates would be considerably improved.

There is no clear basis for answering the question about the role of experience with a subtest in estimating item difficulty. It seems reasonable to speculate that experience is importance, but our evidence was too limited to answer the question one way or the other. Again, this is a good area for further research.

Finally, panelists provided a number of promising suggestions for improving the training and for judging item difficulty. These are summarized in the results section above.

Conclusions

The purposes of this study were to develop and field test two methods for panelists to use in estimating the difficulties of LSAT items. Over the course of four months, three field tests were carried out. Each produced some very useful information, at least for future research. Additional time, of course, was spent conducting background research, planning the field tests, developing materials, and analyzing results.

It would be incorrect to leave the impression that this research study produced a set of validated methods for estimating item difficulty. The three field tests revealed that there was still very much to learn about the process of training panelists to judge the difficulties of test items. At the same time, some of the results, especially those from field test one, were very encouraging. Panelists indicated that they thought they could be trained to complete the item difficulty estimation process with accuracy, and they demonstrated this, at least to some degree. Panelists also demonstrated that they would benefit from discussion. Almost always, estimates after discussion were more accurate than estimates before discussion. Panelists indicated too that they found the discussions useful.

In field test three we were finally able to conduct a meeting without major criticism from panelists about the time allotted to complete the various activities. Failure to estimate times required to complete various activities in the process meant that meetings were rushed, and panelists felt uncomfortable.

We were able to demonstrate that panelists were capable of producing anchor descriptions that more closely met their needs than those produced by the researchers. Perhaps this point should have been obvious to us, but we missed it. The use of a sequential process involving the review of many items and the continued refinement of the descriptions seemed to work.

Another important point learned was that panelists brought to the task their own experiences and ideas about how to judge item difficulties. In future studies, these experiences might be incorporated directly into the process of judging items (rather than collected during the evaluation period or in a haphazard way during the course of the meeting). Research findings are perhaps a good starting point for identifying factors to consider in judging item difficulty, but what we learned in this study is that because the LSAT subtests were sufficiently different from each other and from other research in the field, a unique set of factors would almost certainly be needed to guide panelists through the estimation process. General factors gleaned from previous research are helpful, but much more needs to be done with specific subtests.

We began with the idea that we had two methods to study: one based on anchor descriptions and the other based on item mappings to define the $p+$ scale. In the end (the third field test), what we had was a mixture of the two methods. In part, this was because of the difficulty of producing usable anchor descriptions (though considerable progress was made in field test three), but also because panelists appeared to greatly benefit from the item-mapping information as well. Even when we applied the anchor-based descriptions in field tests one and two, we found it useful to provide sample items to further articulate the $p+$ scale. At this time, our conclusion is that both defining anchor points (such as 0.30, 0.50, and 0.70) along the $p+$ scale and providing exemplar test items either at the anchors or other points along the scale will be valuable information for panelists. Item characteristic curves (Hambleton & Swaminathan, 1985) like those shown in Appendix D (for items in each subtest—across three tests) could be valuable in helping to clarify the descriptions and selecting test items.

With respect to providing exemplary items, many panelists felt that consideration needs to be given to the subtypes of items that are found in each subtest. It is difficult, perhaps even impossible, to apply item difficulty information about items in one subtype to items in other subtypes. Clearly then, there are additional issues in selecting test items besides adequate coverage of the item difficulty scale.

We continue to believe that panelists can be trained to provide accurate estimates of item difficulty. In this study we were somewhat successful, though we were never completely pleased with our implementation, in part because of a failure to allow sufficient time to complete the training, ratings, and discussions. Also, because of the high cognitive level of the test items, we were never completely

comfortable in training panelists to make the ratings, since we were never sure ourselves about the cognitive makeup of the items. In any future efforts, researchers and test specialists should work together (and perhaps with a cognitive psychologist) to produce the descriptions, select the exemplary items, and work collaboratively on the training until panelists feel comfortable with the process.

References

Bejar, I. I. (1983). Subject matter experts' assessment of item statistics. *Applied Psychological Measurement, 7,* 303-310.

Chalifour, C. L., & Powers, D. E. (1989). The relationship of content characteristics of GRE analytical reasoning items to their difficulties and discriminations. *Journal of Educational Measurement, 26* (2), 120-132.

Freedle, R., & Kostin, I. (1993). *The prediction of TOEFL reading comprehension item difficulty for expository prose passages for three item types: Main idea, inference, and supporting ideas items* (ETS Research Report 93-13). Princeton, NJ: Educational Testing Services.

Hambleton, R. K., Bastari, & Xing, D. (1998). *Estimating item statistics* (Laboratory of Psychometric and Evaluative Research Report No. 298). Amherst, MA: University of Massachusetts, School of Education.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications.* Boston: Kluwer Academic Publishers.

Law School Admission Council (1995). *The Official LSAT PrepTest™ XVI.* Newtown, PA: Author.

Lewis, C., & Sheehan, K. (1990). Using Bayesian decision theory to design a computerized mastery test. *Applied Psychological Measurement, 14,* 367-386.

Mislevy, R. J. (1986). Bayes modal estimation in item response models. *Psychometrika, 49,* 359-381.

Mislevy, R. J. (1988). Exploiting auxiliary information about items in the estimation of Rasch item difficulty parameters. *Applied Psychological Measurement, 12* (3), 281-296.

Mislevy, R. J., & Bock, R. D. (1990). *BILOG 3* (2nd ed.). Mooresville, IN: Scientific Software, Inc.

Mislevy, R. J., Sheehan, K. M., & Wingersky, M. (1993). How to equate tests with little or no data. *Journal of Educational Measurement, 30* (1), 55-78.

Sheehan, K., & Mislevy, R. J. (1990). Integrating cognitively and psychometric models to measure document literacy. *Journal of Educational Measurement, 27*(3), 255-272.

Sheehan, K., & Mislevy, R. J. (1994). *A tree-based analysis of items from an assessment of basic mathematics skills* (ETS Research Report 94-14). Princeton, NJ: Educational Testing Services.

Swaminathan, H., & Gifford, J. (1982). Bayesian estimation in the Rasch model. *Journal of Educational Statistics, 7,* 175-191.

Swaminathan, H., & Gifford, J. (1985). Bayesian estimation in the two-parameter logistic model. *Psychometrika, 50,* 349-364.

Swaminathan, H., & Gifford, J. (1986). Bayesian estimation in the three-parameter logistic model. *Psychometrika, 51,* 589-601.

Swaminathan, H., Hambleton, R. K., Sireci, S. G., Xing, D., & Rizavi, S. M. (1999). *Small sample estimation in dichotomous item response models: Effects of priors based on judgmental information on the accuracy of item parameter estimates* (Laboratory of Psychometric and Evaluative Research Report No. 322). Amherst, MA: University of Massachusetts, School of Education.

Thorndike, R. L. (1982). Item and score conversion by pooled judgment. In P. W. Holland, & D. B. Rubin (Eds.), *Test equating* (pp. 309-317). New York: Academic Press.

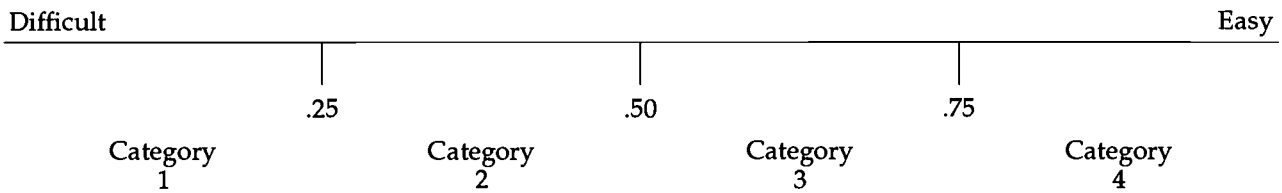
Appendix A

■ Descriptions of the Item Judgmental Tasks in Field Test One

ITEM JUDGEMENT TASK 1 (Anchor-based Method)

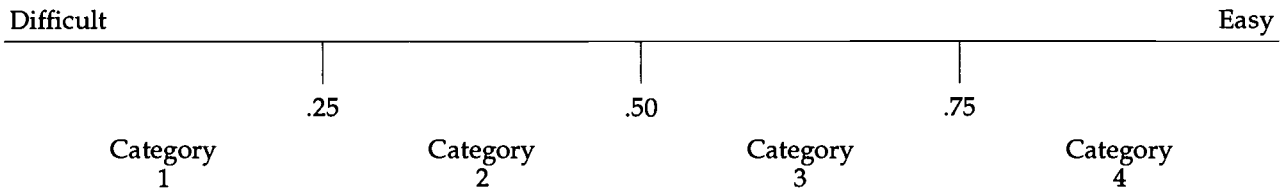
The horizontal line below represents a scale of perceived item difficulty. Three points along this scale are characterized as "more difficult" (.25), "average difficult" (.50), and "less difficult" (.75). The numbers .25, .50, and .75 correspond to the proportion of test takers who are expected to answer items of this type correctly. Earlier today, we discussed characteristics of LSAT items that are located around these three points of the item difficulty scale. Your task here is to review the items in your packet and place each item into one of the four categories demarcated on the scale. For example, if you believe an item is more difficult than those representing the .25 (more difficult) point on the scale, you would place the item in the first category, which is to the left of .25. If you thought an item was extremely easy, you would place it in the category located to the right of .75. Items that you think are between these two extremes should be place to the left or right of .50.

Item Difficulty Scale



This scale is repeated on the following page, which also provides a rating sheet for you to enter your item ratings. For each item, please enter the difficulty category in which you located the item on the scale.

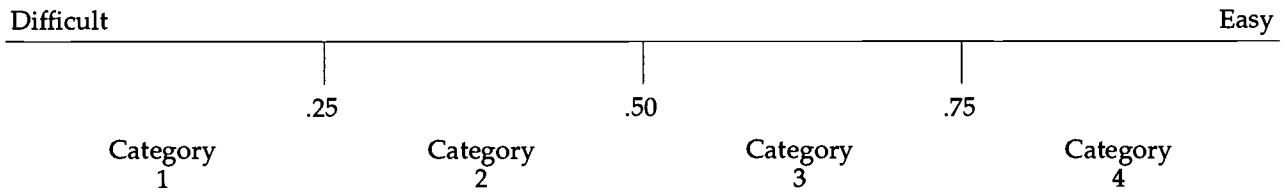
Please review each of the LSAT items in your packet. For each item, decide where you think it should be located on the difficulty scale presented below. Place a "✓" in the appropriate box to indicate your rating. Remember, Category 1 should comprise the items that are most difficult in your opinion, and Category 4 should Comprise the items you believe are the least difficult. If you have any questions about entering these ratings, please ask one of the facilitators.



Item Number	Category 1	Category 2	Category 3	Category 4
1				
2				
3				
4				
5				
6				
7				
8				
9				
10				
11				
12				
13				
14				
15				
16				
17				
18				
19				
20				

Round 2: Opportunity to Revise Ratings

Thank you for completing the first round of ratings. Based on the group discussions and statistical information provided, you may want to revise some of your ratings. If so, please indicate your revised ratings in the table below by placing a "✓" in the appropriate box. If you do not want to revise your original classification of an item, leave that row of the table blank. Please do not change any of your ratings on the previous sheet. If you have any questions about entering these ratings, please ask one of the facilitators.



Item Number	Category 1	Category 2	Category 3	Category 4
1				
2				
3				
4				
5				
6				
7				
8				
9				
10				
11				
12				
13				
14				
15				
16				
17				
18				
19				
20				

ITEM JUDGEMENT TASK 2

Now that you are experts at classifying LSAT items in terms of their difficulty, we would like you to try a new rating task. There are twenty LSAT items located on the table in front of you. These items have been ordered in difficulty from the most difficult (on the left) to the easiest (on the right). Each item on the table is labeled by a number from one to twenty, where one signifies the most difficult item and twenty signifies the easiest item. Your task here is to match each of the items in your packet to one of the twenty difficulty levels represented by these exemplar items. Please enter your rating for each item in the table below. For each item, enter the number representing the difficulty category that you think best represents the difficulty of the item. Your item classifications do not have to exhaust the difficulty categories. **You can place more than one item in any difficulty category.** If you have any questions about completing this task, please ask one of the facilitators.

Item Number	Difficulty Rating	Item Number	Difficulty Rating
1		11	
2		12	
3		13	
4		14	
5		15	
6		16	
7		17	
8		18	
9		19	
10		20	

Round 2: Opportunity to Revise Ratings

Thank you for completing the first round of ratings for task. You are almost finished! Based on the group discussions and statistical information provided, you may want to revise some of the difficulty ratings you gave to some of the items. If so, please use the table below to revise your ratings. You do not need to enter a rating for an item if you believe your original rating was appropriate. If you need assistance, please ask one of the facilitators.

Item Number	Difficulty Rating	Item Number	Difficulty Rating
1		11	
2		12	
3		13	
4		14	
5		15	
6		16	
7		17	
8		18	
9		19	
10		20	

Thank you for completing these ratings!! Thank you for your hard work today!!

Appendix B

■ Field Test Two Evaluation Form

LSAT Item Judgment Participant Questionnaire (Field Test No. 2)

Thank you for completing the item difficulty estimation tasks. We would now like to give you an opportunity to provide feedback to us about your thoughts on this project. Your answers to the questions below will be confidential, and will help us evaluate and improve the process of gaining accurate judgments of item difficulty. If there is insufficient space provided to answer the questions, feel free to use the back of the pages, or attach additional sheets.

General

1) How useful was the general orientation to the purposes of the project and the presentation of factors that influence the level of item difficulty? (Circle one)

a) very useful b) useful c) somewhat useful d) not at all useful

2) Now that you have completed the item rating tasks, can you think of other factors or item characteristics that may influence item difficulty that were **NOT** covered in the training sessions? If so, please describe these factors.

3) How familiar were you with the specific items used in this study? Had you reviewed these items previously? Were you able to remember some of their difficulty estimates from previous reviews?

4) What variations in the procedures used today would you recommend for the future?

Anchor-based method

5) What is your opinion about the amount of *training time* for the anchor-based method?

a) about right b) too much time c) too little time

6) What is your opinion about the training for the anchor-based method? Do you have any specific suggestions for improving this training?

7) For the anchor-based method, how useful were the group discussions following the initial, individual ratings of item difficulty? How did the group discussions affect your confidence in making the item difficulty ratings?

8) For the anchor-based method, what is your opinion about the amount of time allocated for *group discussion*? (circle one)

a) about right b) too much time c) too little time

9) For the anchor-based method, what item characteristics or factors did you use to make your ratings of item difficulty?

Item Mapping Method

10) What is your opinion about the amount of *training time* for the item-mapping method? (circle one)

a) about right b) too much time c) too little time

11) What is your opinion about the training for the item-mapping method? Do you have any specific suggestions for improving this training?

12) For the item-mapping method, how useful were the group discussions following the initial, individual ratings of item difficulty? How did the group discussions affect your confidence in making the item difficulty ratings?

13) For the item mapping method, what is your opinion about the amount of time allocated for *group discussion*? (circle one)

a) about right b) too much time c) too little time

14) For the item-mapping method, what item characteristics or factors did you use to make your ratings of item difficulty?

15) How useful were the item p-values in the item-mapping method? Do you think having these *p*-values led to more accurate ratings in comparison to the anchor-based method?

Confidence Ratings

16) How confident are you that your ratings from the *anchor-based* session will be useful for meaningful predictions of item difficulty? (Please circle a number on the scale below.)

1	2	3	4	5	6
Not at all Confident					Very Confident

Comment:

17) How confident are you that your ratings from the *item mapping* session will be useful for making meaningful predictions of item difficulty?

1	2	3	4	5	6
Not at all Confident					Very Confident

Comment:

_____

_____

_____

_____

_____

_____

18) Based on your experiences today, how confident are you that test specialists can provide useful estimates of item difficulty?

1	2	3	4	5	6
Not at all Confident					Very Confident

Comment:

_____

_____

_____

_____

_____

_____

19) Please rank-order the three LSAT sections in terms of your involvement with each section. A rank of one indicates the section on which you spend the most time, and a rank of three indicates the section on which you spend the least time.

_____ Reading Comprehension

_____ Analytical Reasoning

_____ Logical Reasoning

Appendix C

■ Field Test Three Materials

Training Materials

Estimating the Difficulty of LSAT Reading Comprehension Items

Workshop Handout (August 27, 1997)

Goals of the One-Day Meeting

There are two goals for this one-day workshop with LSAT test specialists:

1. They will receive training on the estimation of test item difficulties. (morning)

2. They will estimate item difficulties on one section of the LSAT. (afternoon)

Item Difficulty

For our purposes, *item difficulty* is defined as the proportion of candidates in the national sample of candidates taking the LSAT who answer the item correctly. For example: an item difficulty of .65 means that 65% of the candidates answered the item correctly. Easy test items have high numbers (say, .70 and .80) and hard test items have low numbers (say, .25 and .40). Over the course of this workshop, the expectation is that test specialists will be trained in the accurate estimation of item difficulty and then apply their skills to some new test items.

Factors Which Affect the Difficulty of Test Items

In helping test specialists make estimates about item difficulty, two sets of factors have been identified: (1) those factors which are general and are often cited in the measurement literature, and (2) specific factors concerning reading comprehension test items.

General Factors

There are several general factors in the measurement literature which can help in the prediction or estimation of the difficulty of test items. Probably the ones most relevant to reading comprehension test items are the following:

1. Negations: the greater the number, the more difficult the test item. For example, "Except-type" test items are generally harder than "Which statement is true...".

2. Vocabulary: the more multisyllabic words used, the more difficult the test item.

3. Sentence and paragraph length affect test item difficulty (because length often introduces more points of complexity, rules, and so on.).

4. Abstract (or non-intuitive) texts/concepts affect test item difficulty.

5. In general, the level of cognitive skills needed to solve a problem affects the test item difficulty.

6. The closeness of distractors to the correct answer affects item difficulty. When distractors and the correct answer are close in meaning or features, the test item is more difficult.

Specific Factors for Reading Comprehension Test Items

In reviewing the measurement literature (e.g., papers by Freedle and Kostin, and Boldt and Freedle) there were a number of factors suggested which can help in the prediction of the difficulty of reading comprehension test items:

1. Look for surface features such as number of words in the stimulus, item stem, answer choices; number of hard words.

2. Watch for connective propositions such as: and, but, however, since, because. These are rhetorical devices which tend to influence item difficulty.

3. Negations influence item difficulty.

4. The use of referential expressions influences item difficulty.

5. Main ideas expressed early in a passage (first paragraph) tend to be easier than when they appear in the middle of passages.

6. Vocabulary level, sentence length, passage length, number of paragraphs, paragraph length, and abstraction of text have all been found to influence item difficulty.

7. Inference test items, too, are harder when the relevant material appears in middle paragraphs of the passage.

8. Overlap of words in the passage and the correct answer reduces item difficulty.

Steps in the Item Difficulty Estimation Process

The plan to be followed today consists of nine steps:

1. General Orientation (very brief, 10 minutes maximum, describe purpose of study, and p-values).

2. Practice Test (test specialists will work through about 20 items, 45 minutes)

3. General Rules for Estimating Item Difficulty (these will be specific to the section of the test, 15 minutes)

4. Practice in Estimating Item Difficulty (5 to 7 items, first ratings, discussion, second ratings, discussion of factors affecting difficulty, 30 minutes)

5. Continued Practice in Estimating Item Difficulty (repeat Step 4 with the next 5 to 7 items in the Practice Test, 30 minutes)

6. Development of Anchors or Descriptions (based on the first 10 to 14 items, 30 minutes)

7. Continued Practice in Estimating Item Difficulty (repeat Step 4 with the next 5 to 7 items in the Practice Test, focusing special attention on the Anchors and sample items—we want the test specialists to be using the Anchors and the item statistics for the first 10 to 14 items, 30 minutes)

8. Final Revisions to the Anchors and Review of Sample Item Statistics (30 minutes)

9. Implementation of the Estimation of Item Difficulty Process (review the items, provide initial ratings, discuss the ratings, provide second ratings, 120 minutes).

We think the nine steps should ensure that test specialists understand the procedure for judging item difficulty, have confidence in applying the procedure (because of the practice and feedback received), and have sufficient time to complete the procedure in an unhurried way.

August 25, 1997

Evaluation Form

LSAT Item Difficulty Estimation Workshop (Field Test No. 3)

Thank you for completing the item difficulty estimation tasks. We would like to give you an opportunity to provide feedback to us about your thoughts on this workshop. Your answers to the questions below will be confidential, and will help us evaluate and improve the process of gaining accurate judgments of item difficulty. If there is insufficient space provided to answer the questions, feel free to use the back of the pages, or attach additional sheets.

General Questions

1) Now that you have completed the item rating tasks, can you think of other factors or item characteristics that may influence item difficulty that were **NOT** covered in the training? If so, please describe these factors.

_____

_____

_____

_____

_____

_____

2) How familiar were you with the specific items used in Set 2: Had you reviewed these items previously? (circle one)

Yes No Unsure

3) What variations in the procedure used today for estimating item difficulty would you recommend for the future?

_____

_____

_____

_____

_____

_____

Item Difficulty Estimation Procedure

4) What is your opinion about the amount of training time that was allocated for the procedure? (circle one)

a) about right b) too much time c) too little time

5) What is your opinion about the nature of the training you were given for estimating item difficulties? Do you have any specific suggestions for improving this training?

_____

_____

_____

_____

_____

6) How useful were the group discussions following the initial, individual ratings of item difficulty?

_____

_____

_____

_____

7) What is your opinion about the amount of time allocated for *group discussion*? (circle one)

a) about right b) too much time c) too little time

8) How did the group discussions affect your confidence in estimating the item difficulties?

_____

_____

_____

_____

9) What item characteristics or factors did you use to make your estimates of item difficulty? (circle **all** that apply)

a. Anchors or descriptors developed
b. Practice item difficulties (Set 1)
c. My own experiences in working the items today
d. General factors provided by the facilitator
e. Specific factors provided by the facilitator
f. My experience in editing LSAT items in my job

Additional factors:

_____

_____

_____

_____

10) How confident are you that your estimates will be useful for predicting actual item difficulty? (Please circle a number on the scale below.)

	1	2	3	4	5	6
Not at all Confident						Very Confident

Comment:

_____

_____

_____

11) Please rank-order the three sections of the test in order of your familiarity with them. (1= most familiar and 3 = least familiar)

_____ Analytical Reasoning
_____ Logical Reasoning
_____ Reading Comprehension

Thank you for taking the time to answer these questions!!

Appendix D

■ Sets of Item Characteristic Curves for the Three LSAT Subtests

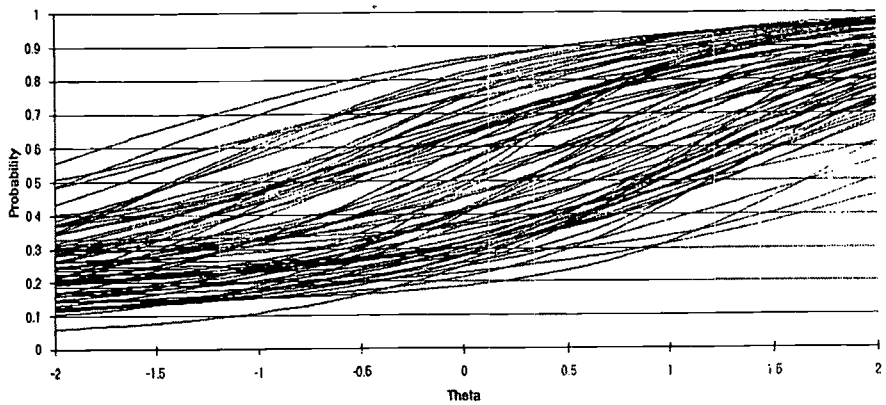


FIGURE D1. *Analytical Reasoning*

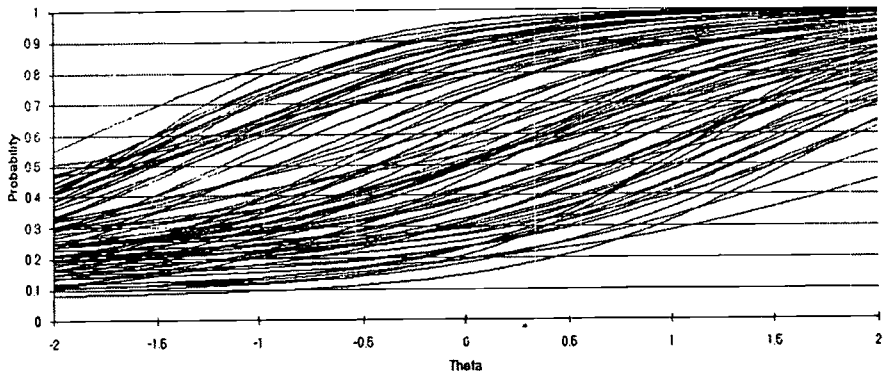


FIGURE D2. *Logical Reasoning (A)*

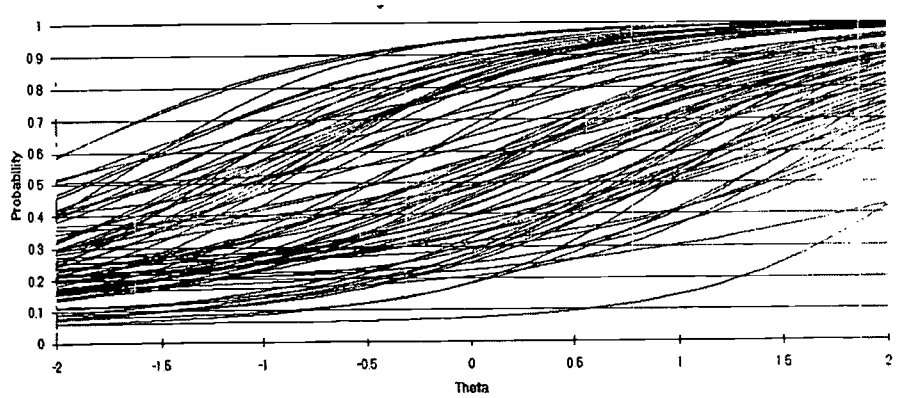


FIGURE D3. *Logical Reasoning (B)*

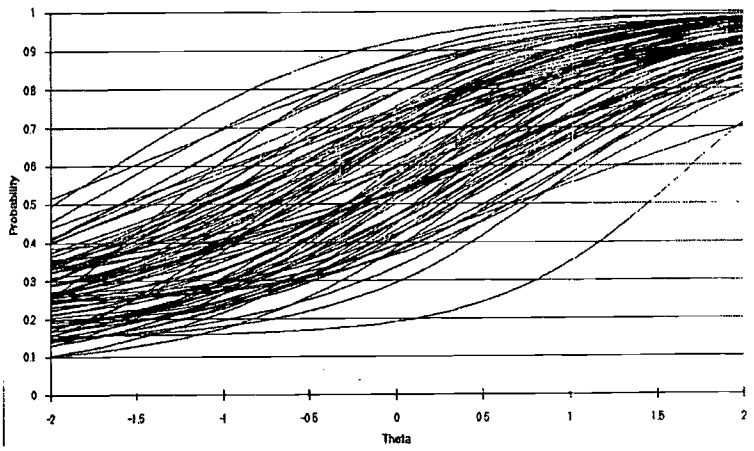


FIGURE D4. *Reading Comprehension*

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

NOTICE

Reproduction Basis

[X] This document is covered by a signed "Reproduction Release (Blanket)" form (on file within the ERIC system), encompassing all or classes of documents from its source organization and, therefore, does not require a "Specific Document" Release form.

[ ] This document is Federally-funded, or carries its own permission to reproduce, or is otherwise in the public domain and, therefore, may be reproduced by ERIC without a signed Reproduction Release form (either "Specific Document" or "Blanket").