Use of Bloom's Taxonomy in Developing Reading Comprehension Specifications

Stephen Luebke
James Lorié

Law School Admission Council

Abstract

This article is a brief account of the use of Bloom's Taxonomy of Educational Objectives (Bloom, Engelhart, Furst, Hill, & Krathwohl, 1956) by staff of the Law School Admission Council in the 1990 development of redesigned specifications for the Reading Comprehension section of the Law School Admission Test. Summary item statistics for the test items developed based on these specifications from 1991 to the present are also presented. These statistics offer evidence that this application of Bloom's Taxonomy to the development of test specifications was useful and helped achieve testing goals, although with limitations.

Keywords: Bloom's taxonomy, Reading comprehension, LSAT

When the specifications of the Law School Admission Test (LSAT) were redesigned in 1990, the goal was to provide a coherent and justifiable hierarchical structure for the development and assembly of test questions assessing postgraduate-level reading skills. Ideally the hierarchical structure would correspond to levels of reading skills and link the content and statistical specifications for a Reading Comprehension section; a balanced assembly of question types would ordinarily provide an appropriate range of difficulty of questions, and vice versa. Law School Admission Council (LSAC) staff used ideas, concepts, and terminology from Bloom's Taxonomy of Educational Objectives (Bloom, et al., 1956) to devise these new Reading Comprehension specifications.

The statistical characteristics of the items developed, assembled, and administered based on these specifications indicate that the specifications do reflect a hierarchy of difficulty, although practical considerations in the development of multiple-choice items for certain subtypes of items result in some exceptions to the ordered hierarchy. LSAC's use of an educational "theory"—in the loose sense that the term applies to Bloom's Taxonomy—to develop item specifications was not as systematic or complete as it might have been if the test had been designed from scratch. However, the use of Bloom's Taxonomy does represent an early and limited attempt at something akin to assessment engineering. (Luecht, 2013) The statistical results presented in this paper, as well as the stability and consistency of the LSAT over more than two decades, testify to the fruitfulness of such an approach.

Context

The LSAT is a paper-and-pencil test consisting of five 35-minute sections of multiple-choice questions. Four of the five sections contribute to the test taker's score. These sections include one Reading Comprehension section with 26–28 questions, one Analytical Reasoning section with 22–24 questions, and two Logical Reasoning sections with 24–26 questions. The unscored section is typically used to pretest new questions or pre-equate new test forms. A 35-minute unscored writing sample is administered at the end of the test.

In 1990, LSAC significantly redesigned the LSAT to address psychometric and content issues. This redesign involved the development of new statistical and content specifications for the test and the introduction of a new score scale. The development of the new specifications was, however, constrained by the continued use of the existing item types for the LSAT—Analytical Reasoning, Logical Reasoning, and Reading Comprehension—and the need to utilize an existing item pool. Also, the redesign needed to be accomplished and made operational in a little over a year.

The goal of the Reading Comprehension item type on the LSAT is to assess high-level reading comprehension skills that are fundamental to success in law school. In 1990, the LSAC staff did not find much, if any, literature on the assessment of skills at this level. The literature on reading assessment that was found focused on lower skill levels, primarily K–12. The composition of the LSAC Reading Comprehension item pool at the time reflected the earlier judgments of test developers about what sorts of questions were appropriate for assessing reading skills at a postgraduate level. The questions addressed the specific content of the reading passage, its main point, inferences that could be drawn, the author's attitude, and other areas. These question types provided a starting point for the development of the new specifications.

The goal of the new specifications was to provide an organizational structure for both existing items and newly developed items that would place them into a plausible and intellectually defensible cognitive framework, and that would also be useful in addressing practical testing needs.

The new specifications would be used to provide appropriate structure and consistency in developing new items and in assembling items into sections. Ideally the specifications would provide a structure that would reflect the difficulty of the reading tasks being organized so that the content specifications for a Reading Comprehension section would be positively related to the statistical specifications. To achieve this, the new specifications needed to do the following:

- Clearly define subtypes of items in terms of particular reading skills
- Organize the subtypes into categories reflecting cognitive skill levels and corresponding levels of difficulty
- Provide for an appropriate balance among the item subtypes that would (a) represent a reading skill set fundamental to success in law school and (b) provide consistency between LSAT Reading Comprehension sections and comparability between LSAT test forms.

In the absence of available useful research on assessment of postgraduate-level reading skills, the LSAC staff wanted the new Reading Comprehension specifications to be based on theoretical assumptions about cognitive processing that were widely accepted. Bloom's Taxonomy of Educational Objectives (Bloom, et al., 1956) was a good candidate because it was widely known and discussed, and it reflected the consensus of the broad group of educators who developed it. Moreover, it purported to reflect a hierarchy of skills that could be used to organize reading comprehension questions into categories that reflected different levels of difficulty.

However, Bloom's Taxonomy was not designed to apply specifically to reading comprehension, so it did not provide a ready-made categorization of reading skills. To develop useful Reading Comprehension specifications, the educational objectives described in Bloom's Taxonomy had to be adapted to apply to high-level reading skills in general and the skills assessed by LSAT items in particular. There was no way of knowing for sure that the specification structure developed would actually correspond to a hierarchy of difficulty of test items. Nevertheless, it seemed a fruitful way to approach the task of developing new specifications within the constraints LSAC was working with and the amount of time available.

Bloom's Taxonomy

Bloom's Taxonomy was developed as a classification of levels of intellectual behaviors (Bloom, et al, 1956). Bloom divided these behaviors into three domains: cognitive, psychomotor, and affective. What has come to be known as Bloom's Taxonomy most often refers to the classification of behaviors in the cognitive domain. The taxonomy of the affective domain was not as widely discussed, and work on the psychomotor domain was never completed. A revision of Bloom's Taxonomy was published in 2001 (Anderson, Krathwohl, & Bloom, 2001) based on additional work undertaken in the 1990s. However, because that work was not available to LSAC staff in 1990 the revision of Reading Comprehension test specifications was based on the original version of the taxonomy.

The taxonomy is often represented as a pyramid (Figure 1):

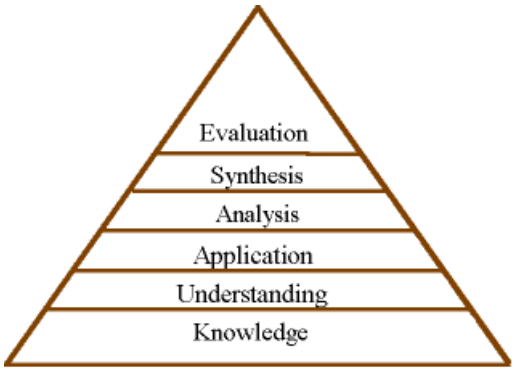


Figure 1. Bloom's (1956) Taxonomy

The pyramid represents a hierarchy, with each level of cognitive behavior regarded as in some sense dependent on those below it. Thus, the hierarchy purportedly represents types of tasks that are of increasing levels of cognitive complexity:

Level 1—Knowledge: recognizing or recalling facts, terms, generalizations, etc.

Level 2—Comprehension: understanding meaning, interpreting, translating

Level 3—Application: applying what is known to a new situation or problem

Level 4—Analysis: separating into parts so that organizational structure can be understood

Level 5—Synthesis: putting parts together to form a new idea or thing

Level 6—Evaluation: making judgments about value

Associated with each level of the hierarchy are terms describing intellectual behaviors that are characteristic of that level. More specifically, the terms displayed in bold play an important role in defining the LSAT Reading Comprehension specifications.

Level 1—Knowledge: select, label, list, **identify**, name, locate, define, recite, **describe**, **state**, memorize, **recognize**

Level 2—Comprehension: match, explain, **restate**, defend, **paraphrase**, **distinguish**, rewrite, summarize, give examples, interrelate, express, **interpret**, illustrate, defend

Level 3—Application: organize, sketch, **generalize**, **apply**, dramatize, solve, prepare, draw, produce, show, choose, paint

Level 4—Analysis: **compare**, **differentiate**, analyze, subdivide, **classify**, **infer**, point out, survey, distinguish, select, categorize, prioritize

Level 5—Synthesis: compose, construct, originate, produce, hypothesize, plan, develop, create, design, invent, combine, **organize**

Level 6—Evaluation: judge, consider, relate, critique, weight, recommend, criticize, summarize, support, **appraise**, **evaluate**, **compare**

LSAT Reading Comprehension Categories

The LSAT Reading Comprehension Specifications divide reading comprehension questions (items) into four categories—(1) Recognition, (2) Understanding and Analysis, (3) Inference, and (4) Application. These categories are intended to represent a hierarchy of reading skills, with the later categories representing higher levels of reading skill that are based on the skills in the earlier categories. The governing principle is ascent from mere recognition of the ordinary meanings of words and sentences to higher levels of critical thinking and application.

- **Category 1—Recognition**: The first and most basic category is Recognition. Items in this category test the ability to recognize what is and is not said in a passage. Most items involve the ability to recognize paraphrases or restatements of what the passage does or does not say. This includes the details of the passage, general claims, and the points being made, including the main point. The emphasis is on recognition of what the sentences of the passage say, given the ordinary meanings of their words. The skills involved in this category are primarily found in the first two levels of Bloom's Taxonomy. Questions ask the test taker, using terms from the taxonomy, to identify, describe, state, and recognize restatements, paraphrases, and basic interpretations of what the passage says.
- **Category 2—Understanding and Analysis**: The second category includes items that test the ability to more fully understand a text by determining the meaning and purpose of terms and phrases from the context in which they are found, and to analyze the parts of the passage, understand their argumentative or rhetorical roles, and grasp the relationship of those parts and their roles to each other. These skills would seem to be at a higher level than just basic recognition of what the sentences of a passage say given the ordinary meaning of their words.

  This category involves skills found primarily in the Comprehension and Analysis levels of Bloom's Taxonomy (Levels 2 and 4 above). These include interpretation, analysis, comparison, and classification. Items might ask a test taker to identify what a term refers to, what the purpose is of a phrase or reference, what the role of a paragraph is in the passage, or how the argument in the passage proceeds.

- **Category 3—Inference**: Items in this category ask what can be inferred from the passage. While Recognition items ask about what is explicitly stated in the passage, inference items ask about what is implicit in the passage. They ask the test taker to "read between the lines." Drawing justified inferences depends on both recognizing what the passage says and understanding the text, so the skills tested in this category depend on those in the previous two categories. Moreover, drawing appropriate and justified inferences from a text is an important high-level reading skill.

  In Bloom's Taxonomy, inference is a term in the Analysis category (Level 4). However, drawing justified inferences from passages involves elements of both Analysis and Synthesis (Levels 4 and 5), such as comparing, differentiating, and classifying facts and ideas; combining and organizing facts and ideas; and formulating hypotheses.

- **Category 4—Application**: Items in this category ask the test taker to apply what is in the passage to the world outside the passage. This includes questions that ask what the author's view might be about something not mentioned in the passage, what might be analogous to something in the passage, what general principle might be suggested by the passage, how other facts or ideas not mentioned in the passage might bear on the passage (i.e., strengthen or weaken), and what the author might intend the passage to accomplish in the world. They ask the test taker to apply, generalize, and evaluate. These tasks require recognition of what the passage says, understanding of its text, and recognition of its implications. Therefore, the skills involved in the Application category depend on those skills in the other categories.

  Application is itself a category in Bloom's Taxonomy (Level 3), but it is at a middle level of the hierarchy (i.e., below Analysis, Synthesis, and Evaluation). However,

in the context of reading comprehension, the application of the comprehension or understanding of a passage to outside issues, contexts, or problems seemed to LSAC staff to presuppose or involve the supposedly higher-level cognitive activities. A category that includes extending the ideas of the passage to new contexts, drawing analogies, developing principles, evaluating additional evidence, and determining what the author of the passage intends to accomplish is possibly a richer and more complex category than Bloom's Application category. What holds it together is the relation of the passage to a broader, outside context.

Hierarchy and Item Difficulty

It has been argued that Bloom's Taxonomy is not a true hierarchy because the categories are interdependent (Paul, 1993). Other researchers have offered revisions of the hierarchical order; for example, Anderson et al. (2001) suggest that the first three levels are hierarchical but the last three are on par with one another. However, the use of Bloom's Taxonomy in the development of the LSAT Reading Comprehension Specifications supposes only that its levels represent, to some extent, increasing levels of complexity in cognitive tasks. The LSAT Reading Comprehension categories do not strictly follow the order of Bloom's levels; instead, they represent different levels of interaction with a text that roughly correspond to the general relationships of Bloom's categories, although with some disordinality.

So, are the LSAT Reading Comprehension categories truly hierarchical? LSAC concedes that there is interdependency between the LSAT categories. That is, some of the tasks in the "lower" categories also involve or incorporate some tasks in higher categories. Consider, for example, the task of recognizing what a text actually says— as opposed to what it does not say. Often—but not always—such a task involves understanding the meanings of terms in context or

other higher level tasks. In contrast, the tasks placed in the category Understanding and Analysis would seem always to include or depend on some of the skills in the Recognition category. The same relationship holds for the other categories. On the basis of that understanding of hierarchy, the LSAT Reading Comprehension categories are hierarchical and represent different and increasingly more complex skill levels. From that understanding one can derive the hypothesis that LSAT test questions in the four LSAT categories should, on average, increase in difficulty in order from Recognition to Application.

There are reasons, however, why that hypothesis might not actually hold. First, there are many factors that can influence the difficulty of a question other than the cognitive complexity of the task it requires. These factors include difficulty of vocabulary, subject matter familiarity, test speededness, and many others. Second, LSAT questions are multiple-choice questions, which might constrain the cognitive tasks involved in answering LSAT questions in ways that weaken or undermine the relationship between item difficulty and the cognitive hierarchy of question types. Finally, LSAT questions must have one and only one correct answer, which might eliminate complexity and ambiguity inherent in higher-level cognitive tasks and thereby reduce the difficulty of the questions.

Summary Statistics

LSAC hypothesized that the four categories of Reading Comprehension questions modeled on Bloom's Taxonomy would represent increasingly complex and sophisticated classes of reading tasks. In operational terms, the expectation was that the mean difficulty of the four categories would ascend from the first through the fourth. Has that hypothesis been borne out by operational experience? The short answer to the question is yes, more or less, but with some notable exceptions that are discussed in more detail below.

The following charts summarize the data for all Reading Comprehension items pretested over the past 20 years (except for items rejected for statistical reasons). The total number of items is exactly 4,900. The data used is pretest data. The number of items in each of the four categories is: Category 1—Recognition = 1,452; Category 2—Understanding and Analysis = 951; Category 3—Inference = 1,503; and Category 4—Application = 994.

The following is an overview of the mean difficulty of the four categories, presented first as a chart of mean IRT-b values (Figure 2) and next as a chart of mean percentage correct values (Figure 3) for all LSAT Reading Comprehension items administered since June 1991.

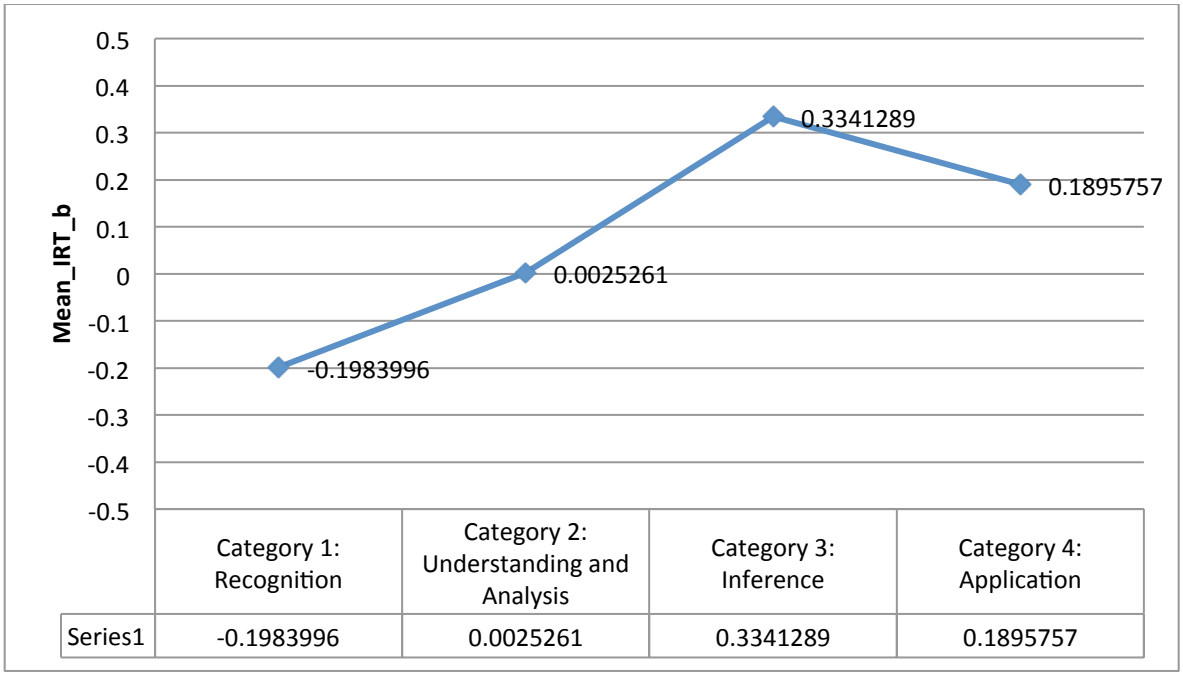


Figure 2. Mean IRT-b values by category

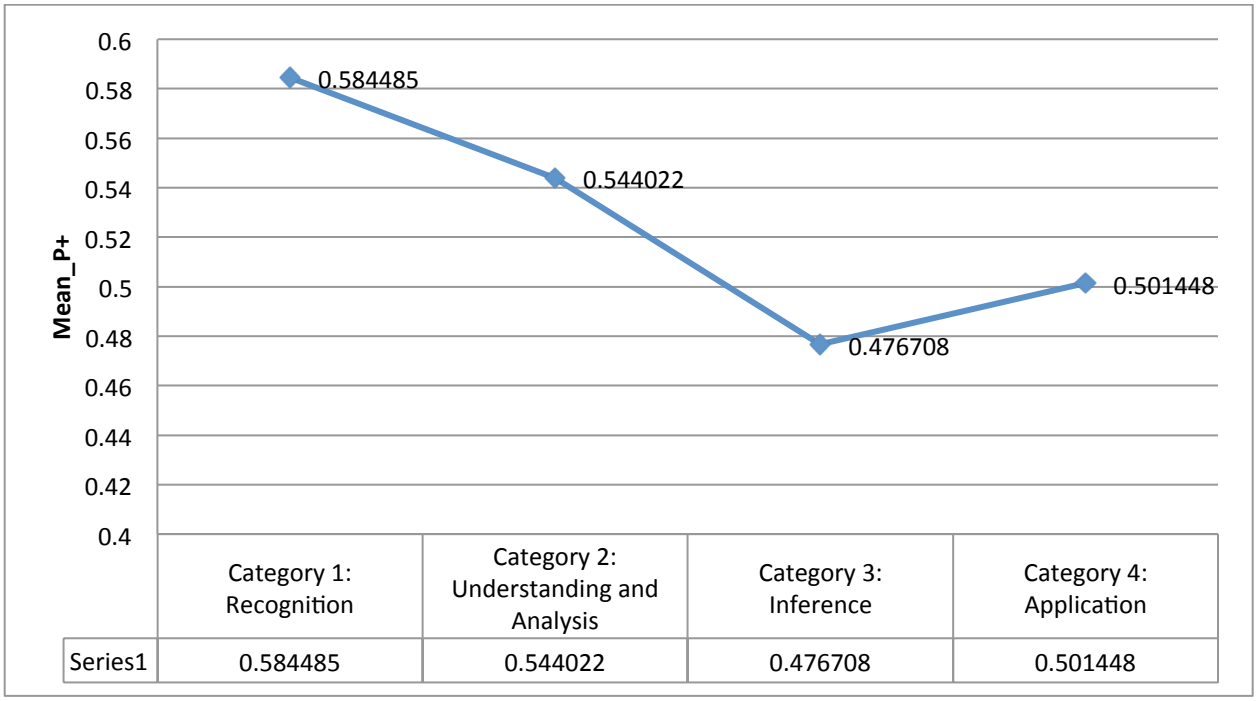


Figure 3. Mean percentage correct (P+) values by category

The following box plots represent the same data, but show that there is substantial overlap in item difficulty across the four categories.

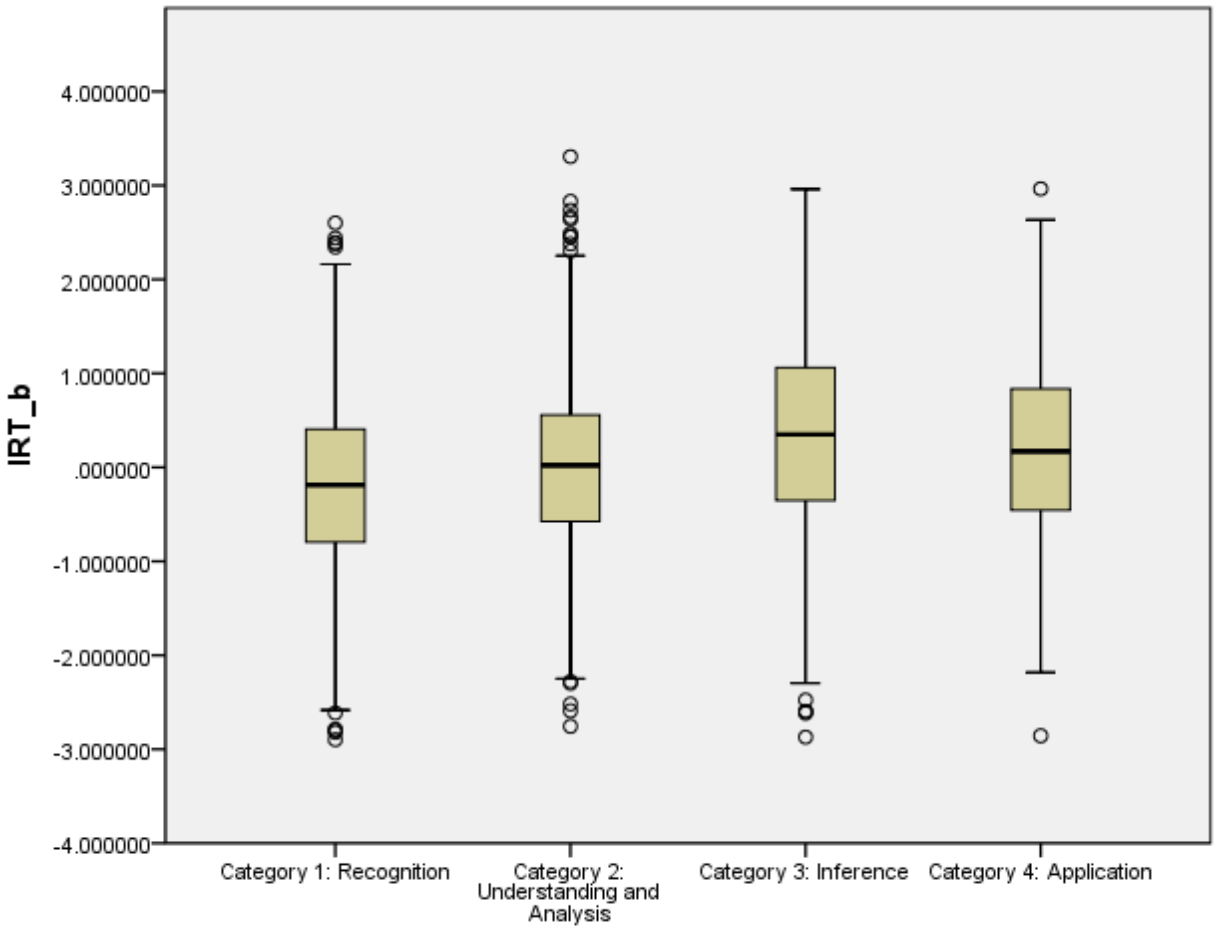


Figure 4. Box plots of IRT-b values by category

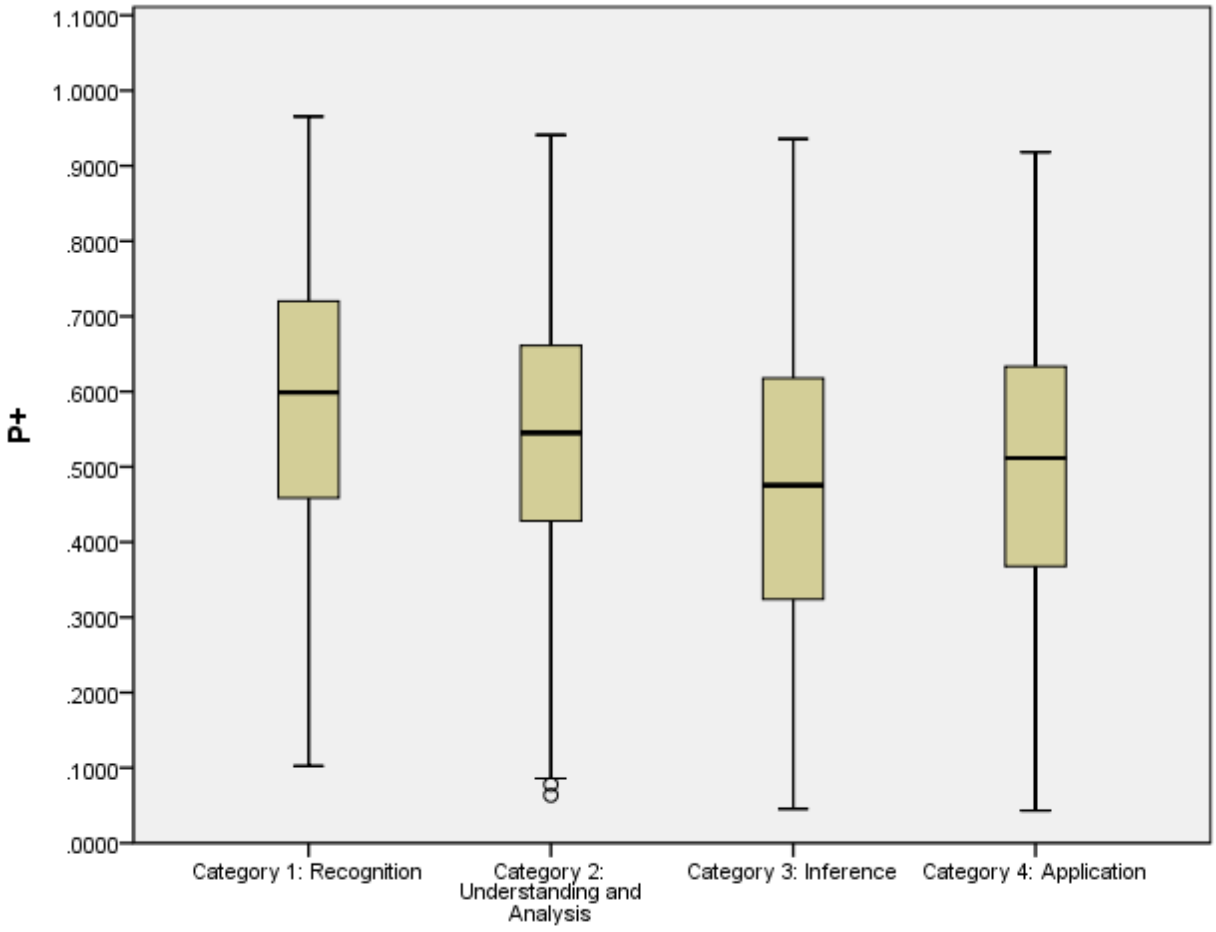


Figure 5. Box plots of percentage correct (P+) values by category

.

This overlap is not surprising in that item difficulty in Reading Comprehension is driven to a large extent by the differing degrees of difficulty and complexity of the various reading passages, as well as by item-level factors. This overlap does affect the usefulness of the hierarchy of categories. The goal of this analysis, however is to explore the extent to which the mean difficulties of the four categories conform, or fail to conform, to the hypothesis that overall difficulty increases across the four categories.

As seen clearly in Figures 2 and 3, the trend across the first three categories matches the predicted progression from less difficult to more difficult. But the fourth category, Application,

does not conform to the hypothesis. The Application category is easier, on average, than the Inference category is. In part this can be explained by the fact that there is an outlier item subtype in that fourth category, the Primary Purpose subtype. Items of this subtype ask test takers to identify the main purpose for which the passage is written, and these items tend to be considerably easier than items of the other subtypes in the Application category. Refining the Application category by subtracting that easier subtype yields a progression of difficulty that conforms more closely to the hypothesis. But, Category 3, the Inference category, also contains an outlier item subtype—called Author's Attitude—that is much easier than the rest of that category. Items of this subtype ask test takers to draw inferences about the attitude displayed by the author of the passage based on indications like tone and word choice. Like Primary Purpose items, these items tend to be fairly easy. Subtracting the Author's Attitude subtype thus makes Category 3 quite a bit harder on average as well, which yields a progression of difficulty across the four categories that resemble the progression described above. In other words, if both outlier subtypes are subtracted, Category 4 remains easier, on average, than Category 3, though both are harder on average than are Categories 1 and 2. The main aspect that's changed when these outlier item subtypes are removed from Categories 3 and 4 is that that the difference in mean difficulty between the first two categories and the last two categories is larger. Figures 6 and 7 show the means after the categories are adjusted in this way

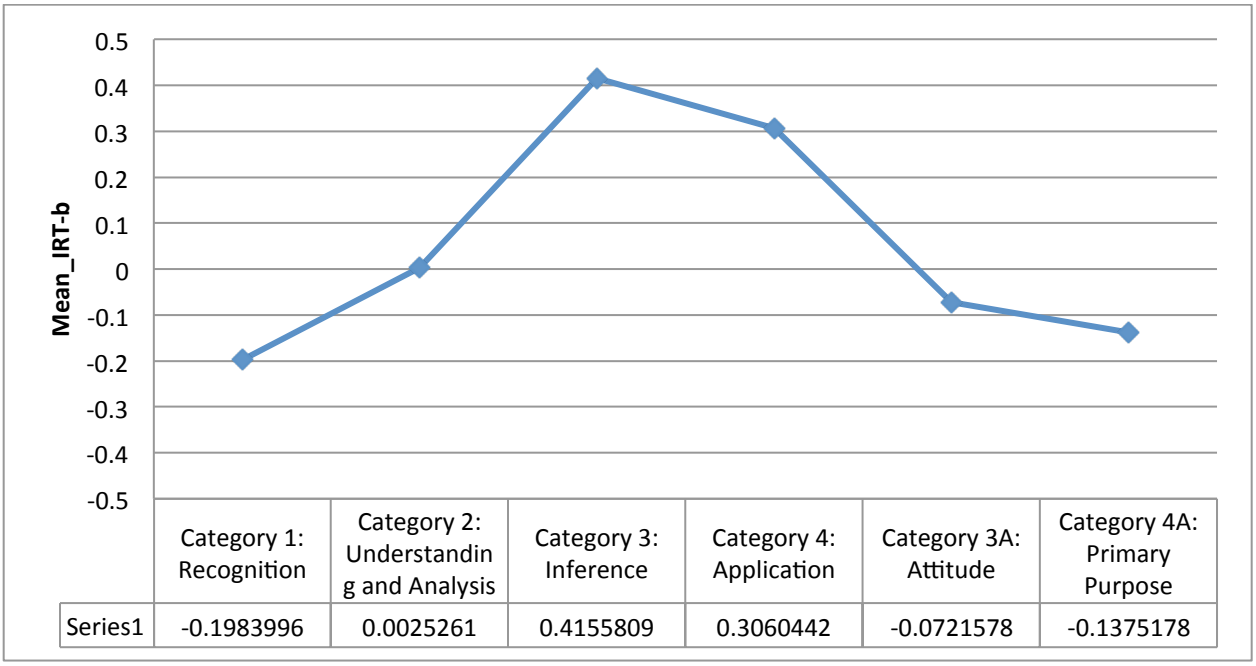


Figure 6. Mean IRT-b values by category (modified). *Note*: The outlier subtypes are the last two values on the right.

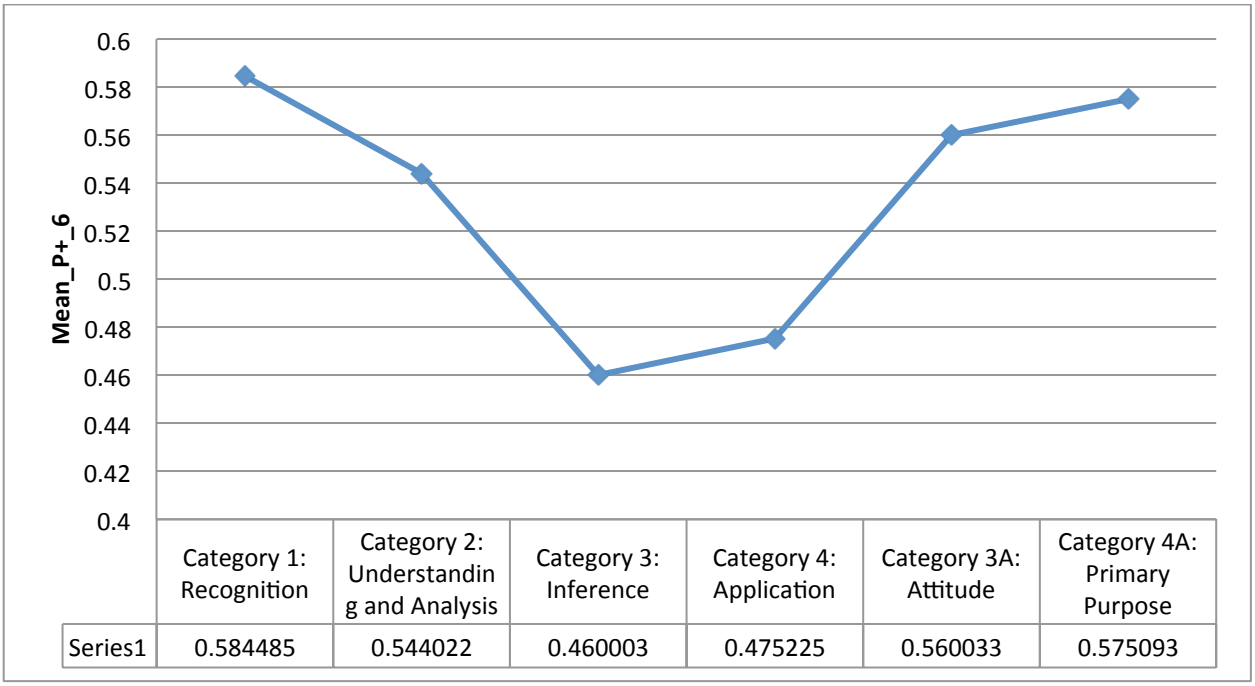


Figure 7. Mean percentage correct (P+) values by category (modified). *Note*: The outlier subtypes are the last two values on the right.

Both of the outlier subtypes are fairly easy. As indicated by both mean IRT-b values and mean percentage correct values, they fall between Categories 1 and 2 in mean difficulty. Consequently, subtracting those subtypes from Categories 3 and 4 makes these two categories harder on average. The mean IRT-b values rise by about 0.09 for Category 3 and by 0.12 for Category 4. Category 4 is still easier than Category 3, but the gap between them is somewhat smaller (0.11, as opposed to 0.14 on the original chart). Once the outlier subtypes are removed, both Categories 3 and 4 are considerably harder on average than Categories 1 and 2.

The Outliers—A Case Study

Two questions must be addressed: (a) Why are the outlier subtypes so much easier than the other items in their respective categories? and (b) Why is the Application category easier than the Inference category (with or without subtraction of the outliers)? Certain characteristics of the skills measured by these subtypes explain why they are easier than the other subtypes in their respective categories. Moreover, the explanations reveal interesting facts about the limitations and constraints faced by test developers measuring verbal skills and reasoning skills with a multiple-choice instrument. Finally, this analysis suggests that the same dynamic that explains the outlier subtypes might help to explain why Category 4 fails to be harder than Category 3.

These questions can be answered by means of a case study of one illustrative item. This item, drawn from the outlier subtype in the Inference category, is an easy item, but it is not the easiest in the subtype. This item (shown below) was selected because it exemplifies the dynamic that explains the outlier subtypes.

As discussed earlier, Category 3 (Inference) measures test takers' ability not only to understand what a text says on the literal level, but also to detect implications that are not spelled

out and draw conclusions that are not part of the text but supported by the text. Clearly, these are high-level skills, particularly when the new conclusions drawn by the reader have as their starting point difficult, dense texts such as those used on the LSAT.

The outlier subtype in this category asks test takers to draw conclusions about the author's attitude toward his or her central topic, or toward some narrower subject discussed in the passage. These conclusions are based on subtle, non-explicit indications such as tone and word choice. Thus one can say that this subtype is not quite like the others in the category; the others focus on conclusions that can be drawn using the information presented in the passage, or on conclusions about what views the author is likely to hold, given the views expressed in the passage. Although the attitude items are not cut from precisely the same cloth, it is reasonable to classify them as inference items because they require conclusions about matters that are not explicitly stated in the passage. A conclusion about an attitude expressed in a passage is an inference.

The particular attitude question examined here is based on a passage that discusses the United Nations' Universal Declaration of Human Rights (UDHR), which was approved by the General Assembly in 1948. The first half of the passage describes some historical background, as well as the process by which the UDHR was drafted, revised, and approved. The author does not reveal much regarding attitude in this part of the passage; however, the latter half of the passage provides several indications of the author's stance:

> The document as it was finally approved set forth the essential principles of freedom and equality for everyone—regardless of sex, race, color, language, religion, political or other opinion, national or social origin, property, birth or other status…While the UDHR is in many ways a progressive document, it also

has weaknesses, the most regrettable of which is its nonbinding legal status. For all its strong language and high ideals, the UDHR remains a resolution of a purely programmatic nature. Nevertheless, the document has led, even if belatedly, to the creation of legally binding human rights conventions, and it clearly deserves recognition as an international standard-setting piece of work....

The attitude inference item based on this passage reads as follows[1]:

Example Attitude Item

The author's stance toward the Universal Declaration of Human Rights can best be described as

A. [ ]

B. qualified approval

C. absolute neutrality

D. reluctant rejection

E. strong hostility

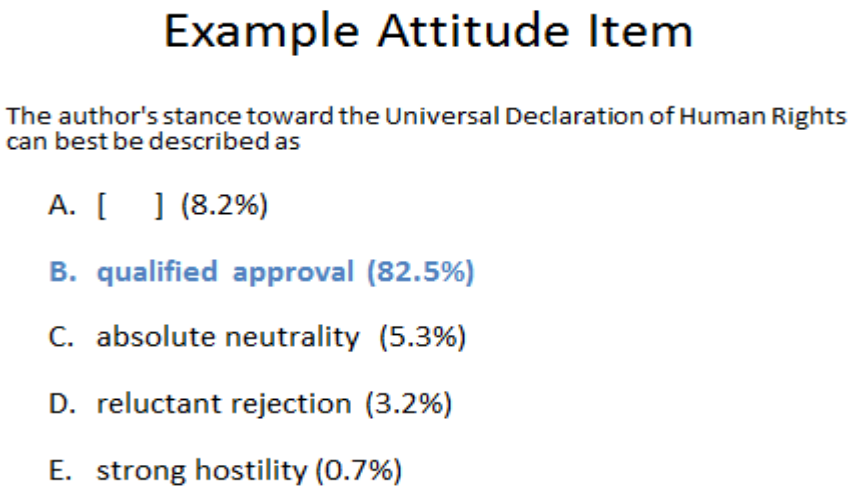
Based on the material quoted from the passage above, it is not difficult to eliminate options D and E. While the author expresses some reservations about the UDHR, his or her overall position is a positive one. There is obviously nothing resembling hostility (E) in the passage, nor does the author ultimately reject the UDHR, reluctantly or not (D). Option C is perhaps slightly more difficult to eliminate, but not by much. Again, the stance taken by the author involves some reservations, but he or she does take a stance, and therefore the passage is

[1] Option A is left out for now; it bears directly on the point being made in this section, but it is not the correct response. Note that some LSAT items belonging to the attitude inference subtype, such as this one, use the term "stance" in the question stem.

not absolutely neutral (C). Option A is also not the correct response, which obviously leaves option B. Indeed, option B captures the positive, though mixed, stance seen in the passage: the author clearly approves of the UDHR (i.e., "it clearly deserves recognition as an international standard-setting piece of work"), but the author also expresses certain regrets about the UDHR (i.e., "it also has weaknesses, the most regrettable of which is its nonbinding legal status").

It is important to note that the stance captured by the correct response is reasonably complex. Because it involves both approval and some criticism, correctly identifying the stance requires the integration of more than one element in the text—elements, moreover, that are in tension with each other. If this were an open-ended constructed response question asking test takers to describe the stance taken by the author, the task of synthesizing these might be moderately difficult—perhaps even quite difficult.

As it turned out, however, this item was quite easy despite the complexity of the task. The following shows the percentage of test takers who chose each response:



The Reading Comprehension section of the LSAT uses only multiple-choice questions. In writing difficult multiple-choice questions, the plausibility and attractiveness of the distractors is all important. In math items, for example, there are often many avenues to take for writing plausible and attractive distractors. These usually involve mistakes that someone who is either careless or confused about the correct approach might make. Test takers often think such distractors are "tricky" or "mean." For example:

A Hypothetical Algebra Item

$(a + b) \times (a + b) =$

A. $a^2 + 2ab + b^2$

B. $a^2 + b^2$

C. $2(a + b)$

In contrast, the distractors in the example attitude inference item above are all negative or neutral on balance. None of these distractors is positive—only the correct response is positive. The task of identifying the correct option among those considered thus reduces to the question, "Is the author's stance positive or not?" The complexity involved in perceiving and synthesizing the elements in tension drops out of the picture, and the task becomes considerably simpler.

In other words, to make this item more difficult, and to capture the complexity that is present in the passage, this item needs *more positive distractors.* However, finding responses that are positive (yet demonstrably incorrect) is quite difficult to do. The best evidence for this comes from what happened with option A (the only positive distractor in the example item):

Example Attitude Item

The author's stance toward the Universal Declaration of Human Rights can best be described as

A. unbridled enthusiasm (8.2%)

B. qualified approval (82.5%)

C. absolute neutrality (5.3%)

D. reluctant rejection (3.2%)

E. strong hostility (0.7%)

As a possible description of the author's stance, "unbridled enthusiasm" is extreme, almost silly. The problem that this distractor illustrates is that while the reading skill involved in identifying an author's attitude is fairly complex, our ability to write the kind of solid, plausible distractors that would make for a hard multiple-choice attitude item is limited by the inherently "soft" nature of the question type: attitudes and stances are entities that are, by their very nature, hard to articulate with great precision. The result is that distractors that are close enough to the correct response to be attractive are probably unsound.

In the item being discussed, almost any distractor that combines a positive overall attitude with some sort of qualification is liable to be a problem. For example, try "restrained enthusiasm" instead of "unbridled enthusiasm": is that a demonstrably worse answer than "qualified approval?" It is arguably not any less accurate as a description of the author's stance. In fact, even responses that do not directly indicate that the author has reservations about the UDHR, but merely leave open the possibility—for example, "general enthusiasm," or "overall enthusiasm"—are hard to defend. They are not demonstrably incorrect. And if it is that difficult

to write one distractor that is both positive and close enough to the correct response to be attractive, imagine trying to write four.

An element of LSAC policy bears directly on this issue. Three of the four main LSAT administrations each year are disclosed to test takers. Test takers are sent a PDF file of the test they took along with their scores, and they have 90 days after they receive their scores to submit a written challenge to any question, or questions, on the test they took. LSAC policy is to answer every challenge, in writing, with an argument defending the credited response and showing why the challenger's response was incorrect. Challengers who are not satisfied with LSAC's explanations have the option of appealing the matter to a panel of independent outside experts. Thus, almost all of the items on the LSAT are subjected to very close scrutiny. Therefore, LSAC must be able to defend every correct response, and every distractor, with solid, reasoned arguments. These arguments must be solid enough to be convincing to independent experts if a challenger chooses that option. The result is that LSAC cannot use distractors that are possibly, or arguably, as good as the correct response.

The impact of this policy is felt throughout LSAC's test development process, but it has an especially pronounced effect on items such as attitude inferences. While the skill being measured is complex, such that a constructed-response question about an author's attitude or stance could actually be fairly difficult, in the multiple-choice context item writers are hindered by the difficulty of writing attractive distractors that are demonstrably incorrect. The resulting items are easier than the skill might lead one to expect. To return to the math analogy, consider what an item with a hard task, but extremely implausible distractors, might look like:

Another Hypothetical Algebra Item

$12x^2 + 9x - 228 = 0$; solve for x:

A. London
B. Paris
C. 4, -4.75
D. Sydney
E. Tokyo

This is an absurd example, but it illustrates the point. Test takers who have no idea how to use the quadratic formula to solve for x, or who have never even heard of the quadratic formula, would still be able to guess the right answer. The general lesson is this: no matter how sophisticated the skill being tested, multiple-choice items cannot adequately capture the difficulty of that skill unless reasonable, plausible distractors can be written.

In the case of reading comprehension, certain reading skills—even though such skills are complex—do not readily lend themselves to good, close distractors for multiple-choice questions. Therefore, attitude inference items are easier than one might have expected based on the hierarchy of cognitive skills represented in the LSAT Reading Comprehension specifications. The same argument applies to the outlier subtype in the Application category. That subtype focuses on the author's overall purpose in writing his or her text—and a purpose is another type of entity that is hard to articulate with great precision.

A case could be made that the same dynamic applies to the Application category as a whole, and that this is why LSAC's hypothesis regarding increasing difficulty across all four categories is not borne out precisely by the data. As mentioned earlier, items in this category ask test takers to apply concepts from the passage to the world outside the passage. This is arguably a more sophisticated class of tasks than drawing inferences from the material in the passage, but moving to the world outside the passage is difficult in a multiple-choice context. The items in the first three categories are constrained by what is actually on the page, but bringing in the broader context outside the passage creates many possible connections that are difficult to rule out definitively. This phenomenon may explain why Application items are easier on average than LSAC anticipated them to be.

Conclusion

The summary item statistics for LSAT Reading Comprehension questions over the past 20 years indicate that the LSAT Reading Comprehension categories by which items and item subtypes are organized do represent a hierarchy of reading skills and item difficulty, albeit an imperfect one. It has been the experience of LSAC that the use of Bloom's Taxonomy in the development of these specifications was fruitful and had useful operational impact, although the relationship between levels in the skill hierarchy and item difficulty has not been as strong was we might have initially hoped. The deviations from the hierarchy of item difficulty found in the LSAT Reading Comprehension categories and subtypes reveal some of the difficulties inherent in developing items that conform to hierarchies of cognitive complexity or skill level. This suggests that there are practical limits to the usefulness of utilizing cognitive or skill hierarchies in the design of test specifications.

The cognitive hierarchy used in LSAT Reading Comprehension has been useful on a general level, but less so on an individual item level. The level of the skill tested by an item in a hierarchy of skills cannot serve as a simple proxy for the difficulty of the item. However, the general differences in the mean difficulty of the items in the hierarchy of categories are useful in form assembly and item development. The proportion of items from the different categories assembled on a test form is generally related to the difficulty of the form. Maintaining the proportions of items in different hierarchical item categories from one test form to another helps with test form comparability and equating. Moreover, the general relationship between the hierarchy of skill categories and item difficulty provides guidance for item development. The difficulty of the LSAT item pool can be affected by directed item writing in certain categories.

The LSAC experience suggests that the use of cognitive or skill hierarchies in the design of item and test specifications may very well be useful for other testing programs in similar ways. However, LSAC experience also indicates that the usefulness of such cognitive hierarchies can be limited by other characteristics of test items and the nature of tests and test programs.

References

Anderson, L. W., Krathwohl, D. A., & Bloom, B. S. (2001). *Taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives.* New York, NY: Longman.

Bloom, B. S., Engelhart, M. D., Furst, E. J., Hill, W. H., & Krathwohl, D. R. (1956). *Taxonomy of educational objectives; the classification of educational goals; Handbook I: Cognitive domain.* New York, NY: Longmans, Green.

Luecht, Richard. (2013, this issue) "Assessment Engineering Task Model Maps: Task Models and Templates s a New Way to Develop and Implement Test Specifications"

Paul, R. (1993). *Critical thinking: What every person needs to survive in a rapidly changing world* (3rd ed.). Rohnert Park, CA: Sonoma State University Press.