Toho Liberal Arts Review No. 44 (2012)
東邦大学教養紀要　第 44 号 (2012)

What Do You Mean My Questions Are Too Hard?

Josef Messerklinger

Introduction

Knotty grammar questions are a staple of tests like the Eiken and many university entrance exams including Japan's Center test. A common rubric on these instruments asks examinees to find the grammar mistake. Ignoring the debate over whether getting students to find grammar mistakes is the best way to judge their language ability or not and disregarding the connection between testing and curriculum, this kind of question can easily be written as a multiple choice item and therefore makes a very attractive option for assessment, far easier than trying to determine a candidate's English proficiency through an interview, which is prohibitively time consuming and raises other issues such as rater reliability and definitions of linguistic competence.

Still, while multiple-choice grammar tests are easier to mark, writing appropriate questions can be perplexing especially if the test is a one-off event. And unlike commercially produced exams, our entrance test does not have the luxury of drawing its questions from a bank of items, which would involve pretesting and rigorous moderation. Item difficulty and discrimination, then, can only be inferred based on our experiences in the classroom and the reported averages of past exams. The objective of this exercise, therefore, is to improve the test writer's chances of constructing effective exams in future by giving him a better feel for which grammar problems are difficult and which are easy, what kinds of questions discriminate best, and what kinds of questions to write in general.

Procedure

For reasons of test security and because statistics from the tests themselves are not available this study relied on data collected from other sources. Students were given questions similar to those described above and basic item analysis was performed on the results. At first, one class was chosen as subjects, but after the second round of testing it was decided to use other groups as well since the first group's tolerance for the activity seemed to be stretching thin. While this makes comparisons between tests a bit more difficult, other groups were given the same two tests as the first group, so

Toho University, Faculty of Pharmaceutical Science

comparison can be made between groups. Still, keep in mind that it is the type of test question that is the real subject of this investigation and not the students (or their teachers, the curriculum, past exams or question moderators*, for that matter).

Another weakness of this evaluation is the number of subjects, which may have skewed the calculations of reliability, which are admittedly very poor. But assuming that the 20 or 30 students who took each set of questions are statistically representative of the whole—classes are divided by name and not ability—the data collected from them should be adequate for this study which is intended to give the item writer a rough but ready idea of which questions work and which do not.

Results

The table below gives general statistics for each set of questions including a calculation of the KR21 for each set, and in separate tables, item difficulty and discrimination indexes for each question along with the average index for each set.

Eyeballing the data shows that the sets of questions are indeed inconsistent, varying considerably from set to set and from group to group even. The results suggest that Set 3 is by far the most difficult and that Set 1 is easy for one group but difficult for another. Set 2 is perhaps the most consistent, or maybe the two groups are most similar, which is reflected in the better, but still not very good, KR21 score. Incidentally, questions of validity are moot since no one can expect these questions to accurately predict student ability with much reliability and especially since no other objective measure is available with which to compare.

Moreover, the target average for students who pass the entrance test is 60%, so these students should score, as a group at least, near that on these trial sets. However, we can see from its average of scores that Set 3 was totally inappropriate, although its discrimination index is good. Interestingly, group 4, second year students, performed worse than did first year students on Set 1, but this may have more to do

Set/group	1 / 1	1 / 4	2 / 1	2 / 2	3 / 3
Average (%)	4.77 (60%)	4.09 (51%)	5.92 (57%)	6.5 (65%)	4.28 (43%)
Variance (SD)	2.64 (1.60)	2.14 (1.47)	4.48 (1.82)	4.38 (2.09)	3.88 (2.01)
Items	8	8	10	10	10
Students	n = 27	n = 23	n = 27	n = 30	n = 25
KR21	r = 0.31	r = 0.08	r = 0.51	r = 0.53	r = 0.41

* My thanks to Andrew Valentine for his comments on the questions used for this investigation. All errors and mistakes, however, belong solely to the author.

with the fact that they are not English majors and may not be interested in improving their English abilities. And again without going into the relationship between testing and curriculum, use of such questions as an achievement test is invalid in any case even if we were to redesign our curriculum to focus on grammar study.

The following tables show item difficulty and discrimination for each set of questions.

Set 1 (group 1)

Item number	1	2	3	4	5	6	7	8
Difficulty	0.81	0.56	0.70	0.67	0.26	0.52	0.52	0.56
Discrimination	0.33	0.67	***0.11***	0.67	0.22	0.33	0.44	0.33

Average item discrimination = 0.39

Set 1 (group 4)

Item number	1	2	3	4	5	6	7	8
Difficulty	0.43	0.29	0.64	0.39	0.29	0.32	0.43	0.61
Discrimination	0.38	0.50	0.38	0.25	0.25	***−0.13***	0.38	0.50

Average item discrimination = 0.31

Set 2 (group 1)

Item	1	2	3	4	5	6	7	8	9	10
Difficulty	0.85	0.81	0.67	0.44	0.19	0.48	0.26	0.74	0.81	0.44
Discrimination	**0.11**	0.22	0.44	**0.11**	**0.11**	1.0	0.44	0.67	0.33	0.33

Average item discrimination = 0.38

Set 2 (group 2)

Item	1	2	3	4	5	6	7	8	9	10
Difficulty	**0.97**	0.83	0.87	0.60	0.30	0.47	0.50	0.83	0.70	0.50
Discrimination	***0.00***	0.50	0.50	0.50	0.63	0.75	1.0	0.50	0.38	0.38

Average item discrimination = 0.51

Set 3 (group 3)

Item	1	2	3	4	5	6	7	8	9	10
Difficulty	0.40	0.68	0.40	0.52	**0.00**	0.36	0.4	0.80	0.36	0.44
Discrimination	0.75	0.50	0.50	0.75	***0.00***	0.50	0.25	0.25	0.38	0.50

Average item discrimination = 0.44

For the problematic items and for the exemplary items, tallies of how often each distracter was chosen will be included in the discussion since they are an important consideration for test construction.

Discussion

So, how did we do? A quick scan of the results shows that only a few items on some administrations of the sets did poorly and only one did truly miserably: Set 1 group 4 item 6. Not bad for a group of questions that were not rigorously moderated, but of course, this kind of item needs to be avoided. Yet this is really what the test writer wants to know more than anything, more than simply that an item seems too easy or too difficult or that it is somehow unclear. We want to know: how will they perform in actual use, how well do they discriminate between students, how difficult are they really, what specifically is wrong with questions that perform poorly, why are some too difficult or too easy, how can they be improved, and what needs to be changed and how should they be changed? So, for the purpose of study what follows is a brief discussion of some notable items.

To begin, let's take a look at the most problematic item, Set 1 item 6, which is reproduced below.

Set 1 item 6: Just two generations [1]ago international travel was [2]accessible only [3]by the very rich or the very [4]adventurous.

1 since 2 accessed 3 to 4 adventurers

An analysis of the data shows that although this item did an acceptable job of discriminating in group 1, it discriminated negatively for group 4. The apparent mistake seems to be the preposition and its collocation with the verb, which is what the test writer seems to have intended. Students were expected to know that in the past travel was expensive and so out of reach for the poor and not accessible to them. On the other hand, a very strong argument can be made for answer 2 which would also solve the problem of finding a mistake and correcting it. Does it matter if the sentence says that only the rich *could* access international travel or that they simply *did?* The context does not make it clear since no mention is made about why only the rich could access international travel; they simply did. Looking at the number of students who chose each answer we see that while eight students chose 3, six chose 2, four chose 1, and three chose 4. Enough of the better students found answer 2 acceptable, causing an inverse in discrimination.

Obviously, questions with two answers must be avoided on multiple choice tests.

Yet, despite careful consideration, this error in test construction was not discovered. Why? Perhaps it is the way the question is read by students and the way it is read by the test writer. Students are primarily looking for a mistake while the meaning is incidental. The item writer, on the other hand, thinks about the expected answer and overlooks other possible interpretations of the sentence assuming that all readers will apply the same background knowledge when deciding which answer is better: *Of course in the past travel was expensive and difficult so only people with money or courage were able to while the poor and timid could not or simply did not.* Unfortunately, it is such assumptions that blind the test writer to the way the question will actually perform when given to students, yet the only criteria for whether a question is good or not is how well it discriminates between actual test-subjects regardless of how well the test writer thinks it will work in theory and regardless of how interesting the sentence might be to the moderators.

Ideally, questions should discriminate perfectly between examinees or at least with a high degree of accuracy, say above 80%. But if the answer is obvious this is just as unlikely as when there are two possible answers. In fact, on Set 2, we find such an item. The better and weaker students in group 2 had equal difficulty with it.

Set 2 item 1: [1]Most [2]people [3]do not enjoy [4]to get up early in the morning.

1 Almost 2 person's 3 are not enjoying 4 getting

As with Set 1 item 6, the answer seems rather obvious, at least to most any native speaker and to any English teacher worth her salt, and a look at the difficulty index for this question we see 85% and 97%, showing that almost all students agree: the question is really too easy even for the average Japanese high school student and so it is not a very good discriminator of their ability. But how were we to know without pretesting and proper item analysis? Again, despite item moderation, this issue was overlooked. On the other hand, while it is hard to justify wasting even one question on such a short test, it is nice to start and end a test with easy items as a way of giving students confidence and helping them do their best.

Although that is a possible justification for including easy items on a test, at the other end of the spectrum and harder to justify are overly difficult questions, which pose similar problems for discrimination, as was the case with this item:

Set 3 item 5: [1]All of the [2]difficulties that [3]are facing us many feel [4]this is the biggest.

1 Of all 2 difficulty 3 is facing 4 these are

This item was also subjected to moderation, yet none of the moderators imagined how it would perform. But a look at the difficulty of this item, 0%, tells us clearly: it is too difficult. In other words, no one got this question right, which means that it does not discriminate at all between the better and worse students, at least not among the students of the class it was given to. Moreover, it can serve no real purpose on an exam, unlike Set 2 item 1 which can be used as a warm up. Most students, eighteen, chose answer 4, yet not only is the singular here not incorrect but also in context the plural does not make sense. Why didn't they get it? Six students chose 3 creating agreement with the apparent subject *all*, and one chose 2, an answer that obviously does not make sense given the subject and in any case would only work if the verb number were also changed. Answer 3 does make some sense, given what appears to be the subject of the sentence, however, only students who were able to analyze the structure of the sentence, see that as it stands there are two nouns that could serve as the subject, and notice that one of the answers solves this problem could get this item correct. Sounds simple enough to the item writer, but to students the sentence may have been just a confusing chain of words.

Getting difficulty right and finding questions that discriminate well usually means proper moderation of items, pretesting and item analysis, and then discarding items that do not work very well, which is how professionally written tests are made. But we sometimes get lucky and make questions such as questions 6 and 7 on Set 2 which managed to discriminate perfectly. Although with the very small numbers involved, this probably has more to do with chance than any outstanding quality these questions might possess. Still, it is worth looking at them to get an idea of what sorts of questions perform well.

Set 2 item 6: [1]<u>If</u> they [2]<u>arrive</u> late to work, the boss will probably [3]<u>be angry</u> [4]<u>to</u> them.

1 Unless 2 arrived 3 angry 4 with

Only two students chose answers 1, which makes nonsense of the sentence and suggests that most students maybe are thinking about the meaning of the sentence, and two others chose 3 making them weak distractors. Otherwise, nine students chose 2, and fourteen chose 4. So, even though two distractors were wasted, having one other attractive choice helped make this question successful, at least for this administration of the set of questions.

Here again, as with Set 1 item 6, students must know the verb/preposition collocation. Answer 1 does not really make much of a change, so was probably ignored by most. Verb tenses, especially in conditionals and reported speech, are usually difficult for students, so choice 2 may have caught a few off their guard. Likewise, choice three confused students, perhaps because angry can be both an adjective and a verb, as in:

a) He angered the boss.

The data analysis itself is unable to reproduce what the examinees were thinking; we only get numbers and have to imagine why one item worked and why another did not. Likewise, that is the weakness of item moderation without pretesting: we can only imagine how the item will work. The test writer in this case assumes that examinees will ask themselves, how does this verb work? Is it anger or be angry? Is it be angry with or be angry to?

Similarly, Set 2 item 7, which required students to know which word fits and which word does not, also seems to have worked very well.

Set 2 item 7: There [1]maybe a long line of [2]people who [3]want to speak [4]with them.

1 may be 2 persons 3 wants 4 to

The students in group 2 who got this question wrong may simply have been careless, which may be exactly what we should be testing for: attention to detail. Difficulty was 50%, which is ideal for this kind of assessment, not too hard and not too easy. Breaking this down by answers chosen, fifteen students chose 1, only one chose 2, five picked 3, and nine went for number 4. And here again, one distracter is wasted since so few chose it.

Looking at the analysis of the same item for group 1, though, shows that these students found the question more difficult, only about a quarter of them got it right. This may also have affected the question's ability to discriminate between students. Seven chose 1, three 2, two 3 and fourteen 4.

The thinking behind this question seems to have been that answer 2 is a common error for Japanese learners of English; the words *people* and *persons* have similar meanings, but slightly different uses. Not many seem to have fallen for this trap. A few students mistook the subject of the verb *want* perhaps thinking that it was *line*. That so many chose answer 4 was a bit of a surprise, but interchangeable

answers sometimes make the best distracters. And as with other successful questions, verb/particle or verb/preposition collocations are a very good indicator of language knowledge.

Finally, it may be instructive to look at one of the most successful items, Set 1 item 2, to see how it was constructed.

Set 1 item 2: We [1]tend to complain [2]that airline seats [3]being too small and intercontinental flights [4]lasting too long.

1 trend 2 about 3 are 4 last

When writing these kinds of questions, the test constructors might try to anticipate how the examinees will tackle the item. In this case, we can imagine that better students know the meanings of the two choices *tend* and *trend* and will disregard answer 1. In fact, in both classes to which this test was given, only one student chose answer 1. On the other hand, answer 2 should cause the better students to pause since it is possible to replace *that* with *about* in this case. But looking at the last two choices we see that these also can be placed into the sentence; however, if we change one, we must also change the other and since only one answer can be correct, we can eliminate these two choices. Still, in group 1, seven students chose 3 and four more chose 4 and in group 4 over half, thirteen students, chose 3 and two chose 4 while only eight, five of the top scores but none of the weaker students, chose 2. These students were perhaps simply looking at the answers to find one that could be replaceable and did not really read the question carefully, which again points to the success of the strategy of using interchangeable distracters.

Conclusion

Why this question works so well is perhaps a bit of a mystery; even if students do not know the difference between complain that (which takes a clause) and complain about (which takes a noun/noun phrase) it can be solved using some test smarts and the process of elimination using rather basic grammar knowledge: parallel construction.

Although some may argue that students do not really need to understand the meaning and communicative value of the sentence but can use grammar and a bit of test knowledge to solve the problem and that we are not testing English ability at all, this is a clever way to test general intelligence, something that any student needs if they are to be successful in university. Therefore, it seems that a key to writing these questions is to make all of the possible answers plausible and get the examinee to use more than grammar skill but also common sense to answer correctly.