Shiken: JALT Testing & Evaluation SIG Newsletter. 6 (2) April 2002 (p. 6-9)

Reading complexity judgments: Episode 3

Gholam Reza Haji Pour Nezhad (Tehran University, Iran)

The first part of this article, online at http://jalt.org/test/haj_1.htm, introduced several factors thought to influence reading comprehension. The second part, available at http://jalt.org/test/haj_2.htm, showed how a test investigating judgments about reading complexity was developed to explore various aspects of non-expert (candidate) judgments of complexity. This section focuses on the last two questions of the study.

Question 4: *How do complexity ratings by students differ from those made by teachers?*

First, it should be made clear that this study involved four different populations: (1) Ninety-nine university English majors in Tehran comprised the student population, (2) Thirty-two English teachers with M.A. degrees comprised the teacher population, (3) Ten English Teachers with Ph.D. degrees and some background in language testing comprised the testing expert population, and (4) Ten native speakers with post-graduate degrees comprised the panel of judges for correctness/incorrectness decisions.

In this study, the panel of judges evaluated 99 paired stem-response statements in terms of the criteria appearing at http://jalt.org/test/haj_p3_0.htm. Of the 99 original items, 44 items were eliminated because less than 9 of the 10 experts agreed on the ratings for those items. The student and teacher populations were then asked to rate those remaining 55 items. The raw data for the student ratings is online at http://jalt.org/test/haj_p3_7.htm and the teacher ratings appear at http://jalt.org/test/haj_p3_7T.htm.

How do the ratings of these two groups differ? Table 1 offers a clue. Both students and teachers judged the complexity and factuality/inferentiality of many items differently from the expert panel.

Table 1. *Student and teacher judgments of item general complexity based on factuality/inferentiality judgment responses*

Factuality/ Inferentiality Judgments	General Complexity Judgments	% of Correct Student Judgments	% of Correct Teacher Judgments
Completely Factual	very easy (1)	65.5%	41.3%
	fairly easy (2)	65.5%	38.2%
	average (3)	67.3%	55.4%
	fairly difficult (4)	55.9%	32.0%
	very difficult (5)	72.7%	30.8%
	Average	65.4%	39.54%
Mostly Factual	very easy (1)	52.5%	55.1%
	fairly easy (2)	57.7%	41.3%
	average (3)	58.8%	50.2%
	fairly difficult (4)	67.5%	45.3%
	very difficult (5)	62.5%	43.2%
	Average	57.3%	47.2%
Evenly Mixed	very easy (1)	55.0%	60.2%
	fairly easy (2)	56.0%	51.6%
	average (3)	45.4%	59.8%
	fairly difficult (4)	54.4%	52.1%
	very difficult (5)	50.0%	58.3%
	Average	53.3%	56.4%
Mostly Inferential	very easy (1)	52.3%	40.5%
	fairly easy (2)	63.6%	32.1%
	average (3)	59.0%	58.1%
	fairly difficult (4)	62.5%	43.1%
	very difficult (5)	50.0%	37.0%
	Average	59.2%	42.16%
Completely Inferential	very easy (1)	66.0%	36.0%
	fairly easy (2)	69.0%	38.8%
	average (3)	63.3%	49.1%
	fairly difficult (4)	60.8%	27.0%
	very difficult (5)	65.9%	19.7%
	Average	65.6%	34.12%

There are clear differences between the student and teacher responses. Whereas student respondents assigned mixed factuality/inferentiality to the most difficult items, teacher respondents assigned mixed factuality/inferentiality to the simplest items. Let us illustrate this with an example. The following stem-response item was rated as "completely inferential" by at least nine of the ten experts, yet "completely inferential" by only 31% (N=31) of the student respondents and "completely inferential" by 82% (N=26) of the teacher respondents.

STEM: The fat hens and chickens in the box beyond the fence were what the fox looked at.
RESPONSE: Because the fox was hungry, it stopped when it saw them.

Furthermore, the most difficult items on the students' ratings continuum were located on the rating "3" on the item general complexity scale cross-tabulated with the "mixed" category on factuality/inferentiality, while in the case of teachers, this is located on the rating "5" on the item general complexity scale cross-tabulated with "completely inferential" on the factuality/inferentiality scale.

Question 5: *How do stem-response combinations influence perceived complexity order rankings?*

In order to find out whether informants assigned a significantly differentiated complexity order to items on the basis of the statement/restatement combination (item) kinds, a two-way ANOVA was utilized to find the interactions between statement kinds and restatement kinds while item complexity rating was taken as the dependent variable. Table 2 and Figure 1 summarize this analysis.

Table 2: *Most/least complex items based on statement-restatement combinations*

Statement Type	Restatement Type	Least Complex	Most Complex
Lexically complex	Syntactically complex	*	
Lexically complex	Non-complex	*	
Syntactically complex	T-unit Complex	*	
Abstract complex	Abstract complex		*
Abstract complex	Doubly complex		*

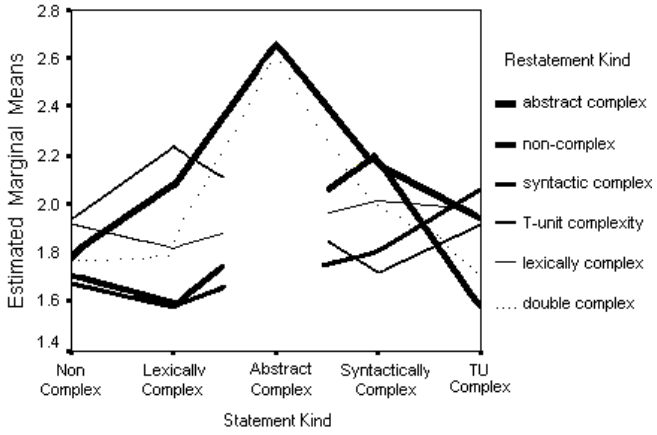

Figure 1. Estimated marginal means of staterestate items for student respondents

To make sure that Figure 1 is clear, let us point out that "mean difference" refers to the difference between the mean of one variable from that of another. For example, the mean difference between non-complex statements and abstract complex ones was -.84. This means that most non-complex statements were .84 points less complex than abstract complex statements.

Moreover, the * mark in Table 2 denotes statistically significant differences at a $P< 0.05$ level, meaning that the result cannot be ascribed to random occurrence in 95% of the cases.

As Table 2 and Figure 1 suggest, it is not necessarily the case that informants consider a combination of non-complex statements and restatements as the least complex. However, they have considered abstract elements as making the item very complex. Table 2 clearly suggests that statement-restatement types are not the only sources of item general complexity, but that informants also pay attention to how statement types and restatement types interact in order to decide how complex items are. Furthermore, as the results of other analyses not mentioned here suggest, informants also paid attention to the pragmatic interaction between item statement and restatement, which is clearly not manifested in statement type-restatement combinations. For instance, they paid attention to the demanded world knowledge between the statement and the restatement and to the factuality/inferentiality demand of the combination. Results showed that, for instance, with combinations demanding high levels of world knowledge, respondents decided that the item was more complex and vice versa.

Nevertheless, the hierarchy displayed in Figure 1 is substantial as it suggests a perceived hierarchy of the complexity caused by statement-restatement combinations. This hierarchy means that, for instance, a combination of a non-complex statement and a doubly complex restatement is considered easier than a combination of a syntactically complex statement and a doubly complex restatement, and this is easier than a combination of an abstract complex statement and a doubly complex restatement. The details of this hierarchy are consistent with the hierarchies presented earlier for statement and restatement complexity ratings. However, as mentioned earlier, respondents have not considered a combination of a non-complex statement and a non-complex restatement as the easiest form. This discrepancy is due to the fact that respondents also pay attention to the world knowledge demand and the factuality/inferentiality interaction between the statement and the restatement as another source of complexity.

Conclusion

The results of the investigations into the questions of the study were revealing in terms of their implications for determining test item difficulty. Analyses performed on statement, restatement, and item perceived complexity led to a systematic hierarchy of perceived complexity where double complexity was considered as the most complex type with the other types following it. This hierarchy suggests that non-expert informants' judgments of complexity are by no means random, but are a very systematic manifestation of their evaluation of factors producing text and test item difficulty so that we observe the same hierarchy again and again in the case of statements, restatements, and items. This hierarchy of perceived complexity is not to replace objective measures of complexity but to complement the existing tools for determining test item difficulty, as comprehensibility is to be considered a measure of text and test item development.

"non-expert informants' judgments of complexity are by no means random, but are a very systematic manifestation of their evaluation of factors producing text and test item difficulty."

Analyses on factuality/inferentiality revealed that there exist different patterns of judgments among students and teachers on this scale. The results showed that students tend to have the poorest performance on test items which they judge as "mixed" on the factuality/inferentiality scale, and the "medium (3)" level on the item general complexity scale. This suggests that what causes students to have the weakest performance on items is not a misclassification of them as factual or inferential, but an overt uncertainty about their factuality/inferentiality. However, in the case of teachers' judgments, the pattern is the exact opposite, where this category is the easiest among the various groups of items.

Alderson (1993) observed differences among teachers and students in determining item difficulty, and concluded that students' and teachers' judgments of item complexity are not reliable sources of information in determining item difficulty. However, he ignored one aspect of this diversity of judgments: consistency. The present study focused on judgments to see whether there is any consistency in the manner students and/or teachers judge complexity, and found meaningful patterns of consistency in the judgments of each group. However, it ignored another important aspect of complexity judgments: why we observe consistency. I believe a fruitful line of further research would be to ask why there is consistency in complexity judgments, what factors give rise to this consistency, and whether the same amount of consistency in judgments is present in standardized proficiency tests.

Reference

Alderson, J. C. (1993). Judgments in language testing. In D. Douglas & C. Chapelle (Eds.)., *A new decade of language testing.* Alexandria, CA: TESOL.

HTML: http://jalt.org/test/haj_3.htm / **PDF**: http://jalt.org/test/PDF/Haji3.pdf