

Can you beat guessing in multiple-choice testing?

Beth A. Mackey¹

Department of Defense

Introduction. The National Security Agency (NSA) tests applicants to ensure that their language levels are appropriate for the government's work. The Language Performance Test (LPT) is used to assess reading comprehension at the Interagency Language Roundtable (ILR) level 2/level 2+. Due to the volume of tests administered each year, many of the tests are in a multiple-choice format to allow for machine scoring. This paper is a result of our desire to better understand how our agency's Language Performance Tests work. Many of the LPTs are formatted to produce an effective, machine-scorable language test.

The machine-scorable LPT is made up of two distinct sections: the keyed completion section and the information identification section. A major goal in designing the keyed completion section was to attempt to overcome the effect of guessing, a main drawback of traditional multiple-choice testing. The potential exists that a testee would be able to improve his or her score by guessing the answer. The possible effects of guessing vary by the number of choices given to the test taker. With five options per item, the chance of getting the right answer is 20%, with four options 25%; and with three options 33.3%. The keyed completion model attempts to reduce the effect of guessing by increasing the number of available choices for each item from the often used four options to twenty.

The other component of the LPT is the information identification section, which is designed to determine how well the testee, given a full text, comprehends the semantic structure or content of texts. It uses a more traditional, four-choice response bank, which is exploited in a novel way to minimize guessing.

Developing the LPT. The Language Performance Test has been in use at NSA for the past twenty years, and in that period James R. Child, the agency's senior language testing researcher, has developed an effective process for test design. The test requires at least two complete text passages: one for the keyed completion section and one for the information identification section. It is often more practical to choose two passages for either or both sections to increase the variety of keyed items. The first stage in the design process is to select the

appropriate texts. They should be from authentic, in-country graphic material,² generated by a native writer for a native audience. Using Child's text typology (1987) as a guide, passages should be in the instructive mode.

The keyed completion section. As mentioned previously, the first part of the LPT is the keyed completion section, which tests knowledge of grammar in the context of communication; that is, the items are designed to test meaning first and form second. Once the passages have been selected, the test designer creates a one-to-one gloss for the keyed completion section before translating it in full. Many might regard this as an extra step, but it has proven useful for test design, in that it furnishes clues as to possible items that might prove difficult for the English reader.

The keyed completion section is based on the cloze format first introduced by Taylor (1953) in which every n-th word is chosen for deletion and the examinees are asked to predict the best word to fill in the blank. In the keyed completion model, the test designer selects the items to be tested, the "keys," by considering both the meanings as well as the forms (i.e., consideration of the text from the perspectives of subject matter and grammar).

In contrast to the original cloze format in which the examinee produced the answer based on his knowledge, the keyed completion model supplies the test taker with a list of options. The next step in the design process, therefore, is to build the lists of possible responses: the "keyed" answers and the "intended" distractors.³ "Intended" distractors are those possibilities that the test designer believes will be attractive to the test taker whose language skills are not sufficiently internalized,

Box 1. Instructive Mode

Language texts in the instructive mode are those that can be tied ultimately to the external world but are not dependent on immediate visual and auditory stimuli for their full interpretation. They are the products of speakers or writers who are conveying facts or exchanging information about situations and occurrences but are not analyzing or expressing personal involvement in the material conveyed. Texts at this level may be rationalized to include a variety of forms (other than news-item), such as extended instructions on how to assemble objects or direction to remote places, recounting of incidents in one's past, narration of historical events, or certain kinds of material where a supposedly factual treatment is strongly influenced by political theory. The content should neither be offensive or repugnant, nor should it contain information that could make it dated. The material should be at the high 2/2+ level of difficulty according to the Interagency Language Roundtable skill level descriptions for reading, and should yield approximately fifty items per section.

Box 2. Interagency Language Roundtable Skill Level Descriptions for Reading

Reading 2 (Limited Working Proficiency). Sufficient comprehension to read simple, authentic written material in a form equivalent to usual printing or typescript on subjects within a familiar context. Able to read with some misunderstandings straightforward, familiar, factual material, but in general insufficiently experienced with the language to draw inferences directly from the linguistic aspects of the text. Can locate and understand the main ideas and details in material written for the general reader. However, persons who have professional knowledge of a subject may be able to summarize or perform sorting and locating tasks with written texts that are well beyond their general proficiency level. The individual can read uncomplicated but authentic prose on familiar subjects that are normally presented in a predictable sequence which aids the reader in understanding. Texts may include descriptions and narrations in contexts such as news items describing frequently occurring events, simple biographical information, social notices, formulaic business letters, and simple technical material written for the general reader. Generally the prose that can be read by the individual is predominantly in straightforward or high-frequency sentence patterns. The individual does not have a broad active vocabulary (that is, which he or she recognizes immediately on sight) but is able to use contextual and real-world cues to understand the text. Characteristically, however, the individual is quite slow in performing such a process. He or she is typically able to answer factual questions about authentic texts of the types described above.

Reading 2+ (Limited Working Proficiency, Plus). Sufficient comprehension to understand most factual material in nontechnical prose as well as some discussions on concrete topics related to special professional interests. Is markedly more proficient at reading materials on a familiar topic. Is able to separate the main ideas and details from lesser ones and uses that distinction to advance understanding. The individual is able to use linguistic context and real-world knowledge to make sensible guesses about unfamiliar material. Has a broad active reading vocabulary. The individual is able to get the gist of main and subsidiary ideas in texts that could only be read thoroughly by persons with much higher proficiencies. Weaknesses include slowness, uncertainty, inability to discern nuance or intentionally disguised meaning.

whether in meaning or in form. These “intended” distractors and correct responses are then arranged in four or five columns. Items involving nominal, verbal, or other categorical entity are grouped with like entities. This is probably the most difficult stage in the design, as the “intended” distractors must be valid in some context, but must not be correct answers for the item or any other keys in the text.

Simultaneously, the designer completes an English rendition of the passage(s). She can manipulate the English to clue or mask any item, depending on the degree of difficulty needed at any point of the test. Clueing gives a hint through the English rendition and makes an item easier, masking on the other hand renders meaning more indirectly through the English and, therefore, increases its difficulty.

As the test designer builds the Keyed Completion test, three components emerge: (a) the skeleton text, (b) the columns of distractors, and (c) an English rendition. The test designer builds in only one or two “intended” distractors into the columns of possible responses, suggesting that the keyed completion test will function as a traditional multiple choice test. As we will see, however, the columnar arrangement generates more responses than the designer projects. Weaker test takers, in particular, will be attracted not only to the “intended” distractors but to other responses in the column.

The information identification section. The second part of the LPT is the information identification section, which attempts to determine how well the testee, given a full text, comprehends the semantic structure or content of texts. The information identification section uses a traditional multiple choice format with a four-choice item bank, thus increasing the possibility that students could improve their scores by guessing. Students are permitted dictionaries on this section as it is a performance test, which “relates to a person’s skill in performing a language task typical of or similar to one required in the workplace” (Child, Clifford, and Lowe 1993: 20). Child has found that “the solution to the problem (of guessing) lies in the skillful linguistic design of the exercise, for example, avoidance of ‘code matching’ responses (such as matching the syntax of the original language) or development of plausible ‘counter plots’ across a set of items” (Child 1995). A counterplot suggests an alternate, but incomplete, scenario. It is yet another means of generating attractive distractors.

I have outlined the nature of the information identification section for the purposes of contrasting its use of four multiple-choice options with the twenty options in the keyed completion section. We will return to the information identification model for this purpose shortly, but first we must focus on the results of our research question in the keyed completion model.

Does the Keyed Completion model reduce the effect of guessing? Compared to the information identification section, the keyed completion section is not easy to develop. It takes considerable skill to select the key items, determine

Box 3. Sample Keyed Completion in German

This sample in German, although illustrative of the general principles of the keyed completion format, represents an earlier model of the test. Newer tests have a more sophisticated arrangement of possible answers, wherein like entities such as nominals or verbals are grouped together.

Skeleton text

_____ 1* _____ 2Δ Regierungsrates _____ 3• Nikaragua, Daniel Ortega, _____ 4 die
 USA _____ 5• Wochenende erneut _____ 6• einer _____ 7Δ Intervention gewarnt.
 _____ 8* Fernsehansprache _____ 9Δ _____ 10 das Ausmaß der von _____ 11* gesteuerten
 und unterstützten _____ 12Δ, _____ 13• die Nikaragua gegenwärtig _____ 14 Souveran-
 ität und _____ 15 Integrität ernsthaft bedroht _____ 16* _____ 17* Gefahr _____ 18 sich für
 Nikaragua _____ 19* Beginn der gemeinsamen Kriegsmanöver der USA _____ 20Δ
 Honduras. Weiter informierte Ortega _____ 21Δ zunehmende Sabotageakte auf
 Industrieanlagen, _____ 22* er _____ 23 Bestandteil der Destabilisierungspläne der
 USA bezeichnete.

English rendition. Daniel Ortega, the coordinator of Nicaragua's ruling council, renewed his warning to the U.S. over the weekend against military intervention. In a television address he pointed to the dimensions of the provocations directed and supported by the U.S., by which Nicaragua currently sees its sovereignty and territorial integrity seriously threatened. Nicaragua sees a particularly dangerous development in the start of joint U.S.-Honduran war maneuvers. Ortega went on to announce increasing acts of sabotage against industrial installations, which he characterized as part of the U.S. destabilization plans.

Box 3. (Continued)				
Possible Responses				
	*	Δ	•	
A	An	Aber	am	als
B	Befinden	Als	Bei	auf
C	Den Koordinator	Auf	Dafür	aus
D	Den USA	Das	Damit	entwickele
E	Der Koordinator	Des	Daß	entwickeln
F	Der USA	Dich	den Flugzeugen	entwickelte
G	die	Ein	Dessen	haben
H	Eine besondere	Errichtung	durch	hat
I	Einem	Er sah	für	hatte
J	Einer	Er verwies	Seiner	hatten
K	Eines	Es	Seines	hätte
L	Eins	militärische	Sich	ihre
M	in deren	militärischen	sie	sandisten
N	Indessen	Provokation	So	sandinistische
O	in einem	Provokationen	Über das	seine
P	In einer	Sah	von	sollen
Q	mit	Sah er	von Flugzeugen	territoriale
R	sah	über	vor	zu
S	sehe	und	Wurden	zum
T	sind	verwies er	Würden	zur

plausible distractors, arrange the choices into columns, and translate the passage(s) into English. The keyed completion section also has some face validity problems. Test takers who are unfamiliar with this model are often frustrated by the hand-eye coordination it requires to match their responses with the correct block on the machine-scorable form. It is also difficult for some testees to follow the three

Box 4. Sample Information Identification

Der Damm des bösen Geistes Asi Abo Aus der südwestchinesischen Provinz

1

Yunnan ist eine Sage überliefert, die die Herkunft eines der schönsten Natur-

2

denkmäler Chinas bezeugen soll. Vor langer Zeit, so heißt es dort, lebte in

3

4

5

dieser Gegend der böse Geist Asi Abo, der alle Menschen vertreiben wollte.

6

7

Als sich die Jäger und Bauern des Stammes der Sani und Axi weigerten ihr

8

Land zu verlassen, beschloß er, sie zu vernichten. Von weit her schleppte der

9

10

11

12

Geist gewaltige Felsbrocken heran, um einen Damm am Nanppanjiang- Fluß

13

zu errichten. Städte und Dörfer und die fruchtbaren Felder sollten in den

14

Wassermassen versinken.

15

Response sheet

- A. from China's southwestern province
- B. south of China's western province
- C. except for China's southwestern province
- D. from the Chinese provincial southwest

- A. an old proverb has it
- B. a certain sage has handed down
- C. a saga has been transferred
- D. there is a traditional tale

- A. is intended to explain
- B. should testify to
- C. should convince
- D. will convince

- A. too long a time
- B. for a longer time
- C. for a long time
- D. a long time ago

Box 4. (Continued)

- A. as hot as it was there
- B. as it says there
- C. as it is called there
- D. as he calls it there

- A. of the wicked ghost Asi Abo
- B. the wicked spirit Asi Abo
- C. of the friendly spirit Asi Abo
- D. the chief spirit Asi Abo

- A. who wanted to chase all the people out
- B. who would have chased out all men
- C. who would drive all men
- D. whom everyone wanted to drive out

- A. of the stem of the Sani and Axi
- B. the Sani and Axi system
- C. stemming the Sani and Axi
- D. of the Sani and Axi tribe

- A. refused to leave their land
- B. refused to allow her land
- C. refused to rely on their land
- D. refused to trust their land

- A. did he complete
- B. he decided
- C. he completed
- D. did he decide

- A. to destroy it
- B. to annihilate her
- C. to annihilate them
- D. to deny them

- A. from far away
- B. from right here
- C. far from here
- D. long ago

Box 4. (Continued)

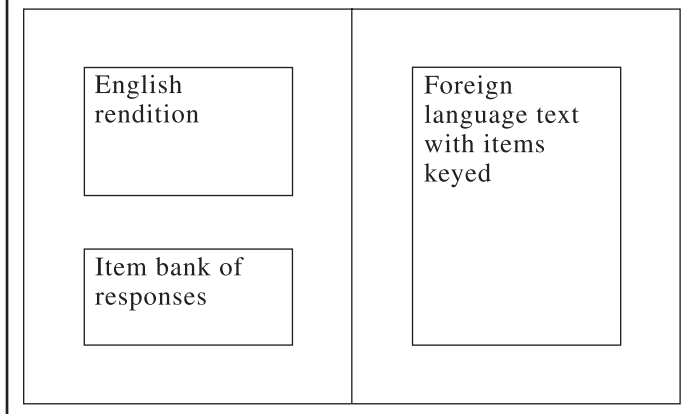
- A. to reach the Nanppanjiang dam
- B. to put a curse on the Nanppanjiang
- C. to build a dam on the Nanppanjiang
- D. to reach the Nanppanjiang river

- A. the barren fronts
- B. the fertile fields
- C. which were fertile fields
- D. the barren fields

- A. sink in the backed up waters
- B. sink the watery mass
- C. deep in the waters
- D. sink the masses of water

separate, but related, components of the test. We have tried to overcome these difficulties by allowing examinees to write in their test booklets and by using page layout to keep the three components of the section aligned at any one point in the test. In the future, computer-assisted testing may address these face validity issues.

As mentioned previously, the test designer builds in a few “intended” distractors; so does it really prove to be any different from a traditional multiple-choice test? In other words, does the Keyed Completion model reduce the effect of guessing?

Box 5. Page layout (facing pages throughout the selection)

Research Methodology. In order to conduct the necessary research, answer sheets from 1997 and 1998 Language Performance Tests were evaluated. Each of the three tests had been previously validated under contract by the Educational Testing Service in Princeton, New Jersey. For this study the data was rescanned into our local system and manipulated in Excel 97 and SPSS 7.0. The research goal was to investigate whether or not the keyed completion format had any effect on the number of answers chosen for each item when compared to the four-choice response bank in the information identification section.

Results. There are at least two ways to reduce the effect of guessing. The first is by increasing the number of choices given to the test taker. The keyed completion model does this by providing the test taker a twenty-item bank of possible responses. As explained previously, however, the designer only builds in two or three "intended" distractors. As table 1 shows, on average, test takers are attracted to almost double that number of possible responses, including the correct response.

One aspect of reducing guessing is the number of times weaker test takers are attracted to less correct or totally incorrect answers. Our study has shown that this can be done by increasing the number of possible distractors on the keyed completion section from four distractors to twenty. But it can also be accomplished by making the distractors more attractive. The test data were used to study the attractiveness of wrong answers. An attractive distractor was defined as one that was selected by at least 10% of the test population. The Russian test produced, by far, the most attractive distractors, but the results were mixed and further research is indicated (see table 2).

Future Research. Another research question connected with the concept of attractiveness is a study of whether the test designer's "intended" distractors proved to be truly attractive, how often, and for what types of words. This is not the only possible area of future investigation. Although preliminary research confirms that the keyed completion test effectively discriminates between test takers,⁴ a more thorough analysis of the relationship between the number of

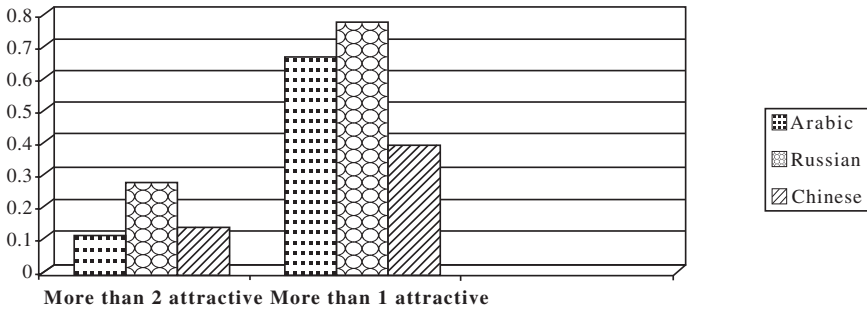
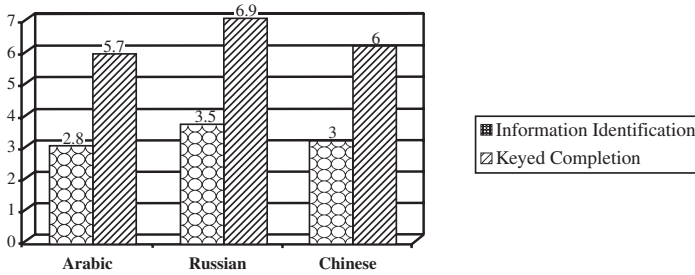
Table 1. Average number of responses

Language	Average number of answers, Keyed completion	Average number of answers, Information identification
Arabic	5.7	2.8
Russian	6.9	3.5
Chinese	6	3

Table 2. Attractiveness of responses

Language	At least two attractive distractors	At least one attractive distractor
Arabic	5 out of 53 questions	34 out of 53 questions
Russian	13 out of 53 questions	40 out of 53 questions
Chinese	6 out of 54 questions	20 out of 54 questions

Note: I define “attractive” as a response by 10% or more of the testing population.



responses chosen and item discrimination would be beneficial. Other research issues include how the language being tested impacts on the results, how the “key” items are selected, what makes the intended distractors attractive, and how this format can be adapted to computer-based testing. We hope that continued research will continue to improve on the Language Performance Test design.

Conclusions. It appears then, from this initial study, that the keyed completion model does reduce the effect of guessing by increasing the number of available responses and, to a certain extent, by including attractive distractors. An unanticipated research result is that the more traditional, four-item multiple choice information identification section of the test is so strong, especially on the Russian LPT, and the causes of this should be investigated further.

REFERENCES

- Child, James R. 1987. Language Proficiency Levels and the Typology of Texts. In Heidi Byrnes and Michael Canale (eds.), *Defining and Developing Proficiency: Guidelines, Implementations and Concepts*. Lincolnwood, IL: National Textbook. 97–106.
- Child, James. 1995. "Description of graphic level 2 testing system." Unpublished manuscript available from author.
- Child, James, Ray Clifford, and Pardee Lowe, Jr. 1993. Proficiency and Performance in Language Testing. *Applied Language Learning* 4(1 and 2): 19–54.
- McIntyre, Sandra S. 1986. Paper presented at the Eighth Annual Language Testing Research Colloquium, February.
- Taylor, W. L. 1953. Cloze Procedure: A New Tool for Measuring Readability. *Journalism Quarterly* 30: 414–438.

NOTES

1. The thoughts expressed herein are the author's own and in no way reflect those of the Department of Defense. I would like to express my gratitude to four colleagues, Mr. James R. Child, my mentor and advisor; Mr. Pardee Lowe, Jr., for his encouragement to pursue this paper and for his considerable feedback on the draft versions; Mr. Reginald Lee Heefner, for helpful suggestions and comments; and Mr. Christopher Bean, for his patience with statistics.
2. In U.S. government work, graphic material is defined as any written format, which would include news items, literary materials, handwritten correspondence, etc.
3. The quotation marks around "intended" distractor(s) reflect the fact that these options are the test designer's estimation that these distractors will prove "attractive" prior to any pilot testing. Further research should investigate the "attractiveness" of the "intended" distractors.
4. A study by the Defense Language Institute concludes that "[O]n the positive side, the keyed completion format seems almost unique in its ability to discriminate with a very high reliability and power across a wide range of reading proficiency levels" (McIntyre 1986).