

Developing the GSTEP Reading and Writing Test

Eve Ribeiro, Marciana Lobo, Tony Becker

Second Language Assessment

AL 8550

The aim of this paper is to design, pilot test, and, as a result of it, make any necessary changes to the reading and writing sections of the GSTEP (Georgia State Test of English Proficiency) in order to have a valid, reliable test. After administering it, we shall make a global analysis of the test development process, its' outcome, rating consistency and reliability. Finally, we shall present a plan on how we would assess the reliability and validity of the test if it were to be used in real testing conditions. Ultimately, we shall demonstrate our ability to apply the concepts and theories covered during this course, and apply it in a realistic test development situation.

The GSTEP is a proficiency test aimed at aspiring undergraduate and graduate students already accepted into Georgia State University, and whose native language is not English. It is used for predicting language performance in their courses by assessing reading, writing, listening and speaking. However, for the purposes of this paper, we shall only concentrate on the reading and writing sections.

The GSTEP was developed to test students who self-identify as non-native speakers. The writing section was based on the previously established Regents until 2000, a writing test geared towards native speakers. However, major changes were made after the faculty responsible for scoring and developing the test raised doubts about its appropriateness. The test started to be prepared by the Department of Applied Linguistics & ESL, when it underwent several modifications. One of them was to integrate the reading with the writing sections on the grounds that it was inauthentic to require students to write a timed impromptu essay about an unseen topic based solely on their background knowledge. Normally in the academic setting, writing is the product of readings, critical

thinking and discussions. Thus, a source text was included in order to ensure that all examinees are provided with the same information, which would activate their prior knowledge for the topic development. Another strong argument for using a source text is that college students must often engage in responding to others` academic texts and provoke reactions to texts of their own (Weigle, in press).

Reading and writing plays a fundamental role in the academic life and there are many aspects that need to be considered when designing a test that will tap into these skills. The issue becomes especially critical when the test takers` native language is not English. We shall make an attempt to make a thorough analysis of the finding in this paper, which will be divided into three sections: I-Test Development, II- Test Development, III- Test Piloting, IV- Test Administration and Analysis , V- Conclusion and VI- Appendix A, B and C.

In following the established test specifications (see *Appendix A*), we encountered several issues. The largest problem we faced was finding two articles which related to the same topic, but would also provide opposing views. We needed to ensure that each of the articles was free of cultural bias, and that the issue did not cause overly strong reactions, such as abortion, capital punishment and so forth. The list of possible topics we carefully considered included alternative medicine, organ donation and space exploration.

Another issue was our attempt to find passages that were authentic. Space exploration is a topic that somehow affects most societies and is often part of academic writings and discussions. Even those students whose country does not participate in space programs, are likely to have some knowledge about it through various media-types. As a

result, we tried to find readings that would activate test takers' schemata, as well as supply them with supporting arguments with which they could base their argument.

Initially, we found the articles, but they far exceeded the 350 word maximum required by the specifications. We made sure to manage the content of the article, so that the last paragraph gave some closure to the overall text, despite the fact that some paragraphs from Passage A were left out. Also, we had concerns about preparing questions that looked for similar or overlapping responses. Therefore, we had to make sure each question was properly constructed, in order to avoid any overlap or repetition.

When developing the questions for the short answer questions (see *Appendix C*), our first step was to formulate two questions that would elicit the main arguments, from both passages. Then, those questions were followed by four questions that would elicit detailed information about the passage. Finally, we had to develop two synthesis questions that would effectively allow the test takers' to use information from each passage to offer suggestions for resolving the issue. However, we initially failed to create a second synthesis question because of concerns about time constraints, but then added the eighth question later. Feedback from the peer discussion conference held in class shed some light on this fact, and other members of the class convinced us that a second synthesis question was needed, as instructed by the test specifications.

Keeping the questions devoid of ambiguity or low-frequency words was yet another source of concern for us. We were concerned about using words and phrases that might not be familiar to students, and this type of issue was touched upon several different times. For examples, we found it was better to use "space programs" consistently in questions *three, four, and five*, rather than confusing the issue by using

other possible subject names. Also, we gave preference to “effects” rather than “drawbacks” in question number *four*, since there were possible positive and negative effects of the space programs, depending upon how one perceived the passages. In question *seven*, however, even though the word “debris” may be rather challenging to a NNS, it was used most of the time and we deemed that there was enough contextual information for students to still understand the question. In reality, we hoped that the test piloting would indicate if the word choice would be an impediment to clarity.

It was quite difficult make question *seven* simply worded, since the nature of the question is itself complex. Our original idea was that, “If you were the author of passage A, what arguments would you use to convince the author of passage B that space debris should be solved?” After much thought and several suggestion, we eventually came to the conclusion that the question failed to successfully state that the test taker had read passage B, and it was too long for students to keep track. These could be probable causes for misunderstanding. Thus, we decided that “imagine,” as an imperative verb, coupled with a period at the end of the sentence, would separate the “reading stage” from the “argument formulation” stage.

Overall, the articles we chose for the final test were challenging to find, as well as difficult to adapt to the GSTEP test specifications and scoring rubrics (see *Appendix D*). We encountered such issues as, authenticity, cultural bias, overlapping, etc., which are all very important to consider when developing a test. The issue was further complicated by the fact that this is a major test, used for the placement of many thousands of students. As a result, we had to be sure that the test was relevant and effective for testing students, especially at the post-secondary levels.

In order to gain some feedback on the development of our test, we administered a pilot test to three non-native English speakers. Afterwards, we asked each of the participants to provide us with feedback on specific questions that we had about the entire reading/writing test. This procedure was performed so that we could find questions that were possibly weaker than others, or to gain more information about the essay topics we had chosen for the test. This process is very important, since native English speakers can also inform us about whether there was adequate time for the test, and whether they thought the test questions (both short answer and essay) were relevant to the topic of space exploration.

One of our most important concerns for the test was the practicality of administering this test. We were particularly interested in learning whether the time given was adequate to complete the test, and if the directions were clear enough. Time becomes an important consideration when taking a test, since if a native speaker feels there is not enough time, then a non-native speaker will certainly be at a disadvantage. Therefore, it was important to make sure there was sufficient time for the native speakers, and this would determine whether we change parts of the test, or add more time. Also, it is important to make sure directions are clear because non-natives would most likely have a more difficult time understanding confusing directions, than would a native English speaker. From the feedback of the pilot study, we were informed that the time length allowed seemed very generous, and that the directions were easy to read and understand.

Although the format of the test appeared strong for test-takers in the pilot study, there were some issues from the pilot study that were brought to our attention. First, there were some remarks about Question #6 [ In Passage A, what is the author referring to

when he writes about “close encounters of the trash kind?” ], about its’ relevance to a movie called “Close Encounters of the Third Kind.” Two of the three native test-takers commented that this question would most likely be confusing for non-native test takers, since many of them might have not seen the movie. This does indeed become a relevant issue, but we felt that the question was still relevant to the article itself, and that the question could indeed be answered, even if the test taker did not see that movie.

If a test-taker did perhaps see that movie this reference would probably provide them with prior information that they could integrate into the short answer. However, seeing the movie, or not, would not ultimately prevent someone from answering the question effectively, in the end. Therefore, this issue did not prompt us to make any changes to the test question, but did raise some awareness about the different possibilities for answers between people who are familiar with the movie, and those who are not familiar with it. Still, this was not the only feedback we received from the pilot study.

When asked about the basis of the questions, and whether the test-taker thought the questions were fair and relevant to the text, one native speaker commented that Questions #3 and #4 seemed awfully similar to Questions #1 and #2 (see *Appendix B*). She indicated that the questions were asking for the same information, and they possibly needed to be changed. We took this into consideration, but realized that the questions might have appeared similar for that particular person, but the other two native speakers answered the way we thought they might. In the first two questions of the test, we were looking for the main arguments for each passage, while in the third and fourth questions, we were looking for direct effects of space programs. These two questions were looking for supporting information that was relevant to the arguments produced by each author.

Also, by using the word “effects” in both the third and fourth questions, we were more specifically looking for *multiple consequences* of space programs, not just one summarizing effect. We intentionally made “effects” plural, rather than singular, so that test-takers would provide several examples of how space programs have affected the universe, directly from each text. Each text lists several examples of how space programs have affected our universe, in one form or another. For some time we considered rewording the question to ask for “specific effects,” but opted not to, since two out of three native English speakers didn’t express the same concern, even when asked specifically about this concern.

Perhaps our greatest contributions came from the classroom discussion we had with other groups. This time enabled us to receive crucial feedback from others who were also developing a reading/writing test very similar to our own. From this session, we learned that we should have included an eighth question, as well as resolve the ambiguity in the seventh question. Originally, we considered leaving out an eighth question, because we felt that the test might then be too long for non-native English speakers. However, we learned from this discussion that the specifications called for at least eight questions, and we added a new question as a result. Therefore, the native speakers were only tested with seven short answer questions, while the non-native speakers were tested with all eight.

In regards to Question #7, the other discussion group indicated that this question might possibly be confusing to most non-natives, especially since it was confusing to them (both native speakers). The question was originally worded as [Imagine that you are the author of Passage A, reading Passage B. What arguments would you use to have

space debris issues solved?]. They both had the impression that this question was asking the reader to imagine that they are an author of one passage trying to support another authors' argument, rather than support their own. After some debate, we all agreed that the question needed to be re-worded a bit, and it eventually became the question that it is now. This information proved useful for us, and hopefully cleared up any ambiguity within the question.

Lastly, we asked each of the three, native English test-takers to rate the difficulty of the overall test, using a 1 to 10 scale (1 = extremely easy, 10 = extremely difficult). Two of the three native people rated the test as a 4, while the third person rated the test as a 5. From this information, we assessed that an average rating of 4.3 from native speakers would be around an average rating of 7-8 for non-native speakers. As a result, we felt confident that the test would not be too difficult for GSTEP test-takers, and it would be a fair assessment of writing skills.

### References

Hamp-Lyons, L., & Kroll, B. (1996). Issues in ESL writing assessment: An overview. *College ESL*, 6,1, 52-72

Weigle, S.C. (in press). Integrating reading and writing in a competency test for non-native speakers. *Assessing Writing*