Listening. Learning. Leading.®

Research Report
ETS RR–13-07

Length of Textual Response as a Construct–Irrelevant Response Strategy: The Case of Shell Language

Isaac I. Bejar

Waverely VanWinkle

Nitin Madnani

William Lewis

Michael Steier

April 2013

ETS Research Report Series

Since its 1947 founding, ETS has conducted and disseminated scientific research to support its products and services, and to advance the measurement and education fields. In keeping with these goals, ETS is committed to making its research freely available to the professional community and to the general public. Published accounts of ETS research, including papers in the ETS Research Report series, undergo a formal peer-review process by ETS staff to ensure that they meet established scientific and professional standards. All such ETS-conducted peer reviews are in addition to any reviews that outside organizations may provide as part of their own publication processes. Peer review notwithstanding, the positions expressed in the ETS Research Report series and other published accounts of ETS research are those of the authors and not necessarily those of the Officers and Trustees of Educational Testing Service.

The Daniel Eignor Editorship is named in honor of Dr. Daniel R. Eignor, who from 2001 until 2011 served the Research and Development division as Editor for the ETS Research Report series. The Eignor Editorship has been created to recognize the pivotal leadership role that Dr. Eignor played in the research publication process at ETS.

Length of Textual Response as a Construct–Irrelevant Response Strategy: The Case of Shell Language

Isaac I. Bejar, Waverely VanWinkle, Nitin Madnani, William Lewis, and Michael Steier
ETS, Princeton, New Jersey

April 2013

Action Editor: Beata Beigman Klebanov

Reviewers: Brent Bridgeman and Michael Heilman

Abstract

The paper applies a natural language computational tool to study a potential construct-irrelevant response strategy, namely the use of *shell language*. Although the study is motivated by the impending increase in the volume of scoring of students responses from assessments to be developed in response to the Race to the Top initiative, the data for the study were obtained from the *GRE*® Analytical Writing measure. The functioning of the shell detection computational tool was first evaluated by applying it to a corpus of over 200,000 issue and argument essays and by means of a study to evaluate whether the shell language score agreed with the characterization of shell by two scoring experts. It was concluded that the computational tool worked well. The tool was then used to select essays for rescoring to determine whether the presence of shell language had had an effect on the operational scores they received. We found no evidence that such an effect was present. However, we did find a leniency effect in the operational scores. That is, the essays that were rescored as part of this project received a lower score compared to the operational score. The validity implications of these results are discussed.

Key words: writing assessment, scoring, GRE, shell language

Acknowledgments

We would like to thank the following staff members: Catherine Trapani for her assistance in obtaining the data and advice on initial study design, Gerry Kokolis for preparing data files and data sets, Florian Lorenz for his assistance with data analysis and presentation, Melissa Lopez for her help inputting the annotated text into the tool, and Daniel Blanchard for his help calculating shell overlap. We would like to extend our gratitude to the scoring leaders who participated in the studies. Finally, we would like to acknowledge the editorial and substantive contributions of Brent Bridgeman, Michael Heilman, and Beata Beigman Klevanov.

Table of Contents

List of Tables

Page

List of Figures

Page

Background

The demand for high-stakes testing is growing and, by most accounts, it will increasingly rely on constructed responses, including writing samples. For example, the volume of scoring from the assessments being produced in response to the Race to the Top initiative (U.S. Department of Education, 2009) will be unprecedented. The volume of responses to be scored, and related challenges, could pose a challenge to the validity of scores. In this report we evaluate a computational tool for identifying *shell language* and report on a small study evaluating the potential effect of shell language on human scoring. We adopt the following definition of shell language: Shell refers to sequences of words used in persuasive writing or speaking to provide an organizational framework for an argument. This shell language, often generalized in nature, does not necessarily include specific details related to the prompt.[1]

The use of shell language is potentially a construct-irrelevant response strategy (CIRS) that could influence human and automated scoring by adding *wordiness* (Powers, 2005b), especially when the shell language is generalized and not related to the prompt. In the case of human scoring, the presence of shell language could be seen by scorers[2] as evidence of argument analysis and production. In the case of automated scoring, the presence of shell language could inflate scores if the scoring engine looks for words related to argumentation as evidence of argumentation skill and also because it would tend to lengthen the response. Response length is a construct-relevant correlate of writing performance (Quinlan, Higgins, & Wolff, 2009) but can also unduly influence automated scores (Bennett, 2011).

The increases in the volume of constructed responses can be assumed from the plans by the two primary state consortia that have been formed as a result of the Race to the Top initiative. The two consortia, the Partnership for Assessment of Readiness for College and Careers (PARCC) and the Smarter Balance Assessment Consortium (Smarter Balanced), will implement somewhat different approaches to accountability testing. However, both consortia call for increased use of items requiring constructed responses. For example, the PARCC application noted that tests, both end-of-year and through-course, "will develop challenging *performance tasks and innovative, computer-enhanced items* that elicit complex demonstrations of learning and measure the full range of knowledge and skills necessary to succeed in college and 21st century careers" (Partnership for Assessment of Readiness for College and Careers, 2010, p. 7, emphasis added).

The inclusion of performance tasks will ensure that the assessments measure skills that are difficult to measure in on-demand assessments, and they will help model effective classroom instruction. Both consortia are expected to collaborate on the scoring of constructed responses by automated means as a means of dealing with the volume of scoring:

> …because the assessment system designs of both Consortia rely heavily on the use of artificial intelligence (AI) scoring engines to score complex items quickly and cost efficiently, the two Consortia will collaborate on the development of standardized AI scoring protocols. They also will explore a possible collaboration on the procurement of an AI engine. (K-12 Center at ETS, 2011, p. 2)

In short, these assessments will be administered in *every* state that is a member of the consortium to *all* students in *multiple* grades *year after year*. This will mean that an unprecedented volume of responses will need to be scored within short periods of time. Given the high stakes involved, the motivation for obtaining high scores will be strong. Therefore, ensuring the validity of the scores will require increased vigilance.[3] The concern is not new, of course. With respect to writing assessment, Hillocks's (2002) criticism of state writing programs for rewarding formulaic writing is well known. He argued that by emphasizing five-paragraph essays and quick scoring the scoring process rewarded structure rather than substance.

For convenience, this study was conducted with essays from the *Graduate Record Examinations*® (*GRE*®) Analytical Writing measure. However, the computational approach we use is relevant to any writing testing program that includes writing samples that call for argument analysis, including the assessments that will be produced by the PARCC and Smarter Balanced consortia. Although the focus here is on human scoring, automated scoring is equally vulnerable to CIRS, as illustrated in the context of automated scoring of GRE[4] essays (Bejar, Flor, Futagi, & Ramineni, 2012) scored by the *e-rater*® scoring engine (Attali & Burstein, 2006). Bejar et al. (2012) found that substituting words in an essay with less frequent synonyms tended to inflate the scores of lower-scoring students when scored with the e-rater engine. Such malleability of scores from construct-irrelevant response strategies potentially weakens a validity argument. Moreover, because automated scoring is typically based on human scores, any vulnerability of human scoring is also relevant to automated scoring (Bejar, 2012).

A specific strategy that has been detected among some GRE test takers is the use of formulaic or shell language. A possible effect of shell language is to lengthen a response in a

construct-irrelevant fashion. The role of length in scoring has been discussed by Powers (2005b). Length, of course, is also a construct-relevant aspect of textual responses (Quinlan et al., 2009). The challenge is to distinguish responses that attempt lengthening a response in a construct-irrelevant manner to obtain a higher score from responses that are appropriately longer due to a more detailed elaboration of an argument, for example.

Computationally, it is fairly straightforward to identify and quantify shell language in a written response based on previous exemplars of such language and a tool for such purposes has been developed (Madnani, Heilman, Tetreault, & Chodorow, 2012). From a validity perspective, it is important to document that responses that contain what appears to be shell language are not inadvertently rewarded or punished. That is, shell language, to the extent that it is present, should be treated neutrally, and scorers are instructed accordingly.

We first validate the functioning of an algorithm for identifying shell language in GRE essays. Next, two scoring leaders will rescore a set of essays that have been previously scored operationally. Differences between the scores of the two scoring leaders will be adjudicated by ETS Assessment Development staff. Discrepancies between the adjudicated scores, presumably a better measure of the writing skills reflected in the essays, and the operational scores[5] will be compared to evaluate whether discrepancies are possibly due to the presence of shell language.

The Computation of Shell Language

The tool designed for the automatic detection of shell language is a rule-based system consisting of approximately 25 complex, regular expression[6] rules that process each sentence in a given essay and mark up the matching spans as shell language.[7] The rules were written on the basis of shell annotations of existing essays provided by ETS Assessment Development staff in charge of the Analytical Writing measure and were done independently of this project.

The following sentences are an example of what is considered shell language: The argument rests on the assumption that A is analogous to B in all respects. This assumption is weak, since although there are points of comparison between A and B, there is much dissimilarity as well.

As noted, scorers are instructed to neither reward nor penalize shell language per se. In the example above, if such a lengthy shell is followed by an appropriate analysis of similarities and dissimilarities, the assigned scorer would take that into account in a positive manner. However, if the lengthy shell is followed by repetition of parts of the prompt instead, rather than

original analysis, the test taker is not engaging in the analysis called for and would not be credited for simply using shell language. It is in this sense that the presence of shell language is not rewarded or penalized.

Besides rules for detecting fragments such as the example above, the system also uses some heuristics to prevent over- and underannotation of shell language, such as merging two discontinuous portions of shell language if a sufficient number of the words between them appear to be shell-like. Single words are not marked as shell. In addition to marking up the shell, the tool also computes a weighted shell score by using prestipulated weights that are a function of the length of the portions of text that have been identified as shell. The weights are chosen so as to penalize longer shell spans more heavily because shorter shell spans are often necessary to construct a coherent argument on any topic. Although the tool is believed to have adequate coverage of different types of shell instances, it may generate false positives in order to achieve such coverage. The shell score used in this study roughly indicates what proportion of the essay is considered shell.

Data

The investigation is based on essays from students taking the GRE between 2006 and 2010 and precedes the release of the GRE revised General Test in 2011. There were 105 issue prompts represented in that period, and 114 argument prompts. The prompts are considered to be equivalent in the sense that the ETS Assessment Development staff did not identify some prompts as being more worthy of study for this project. (The GRE program discloses all its prompts and they are available to test takers; Powers, 2005a). For each prompt, 1,000 essays were randomly selected such that approximately 500 were from test takers that had made appointments to take the test in the United States and 500 had made appointments to take the test outside of the United States. For convenience, we refer to these two test-taking populations as U.S. and non-U.S. This represents an oversampling of foreign test takers because the U.S. test-taking population is larger. This process identified a sample of 212,434 test takers.

Analysis of Shell Score Distribution Across Prompts and by Prompts

Computing the shell score for all the data revealed that the shell was somewhat variable by prompt. The large majority of essays showed relatively little shell. Also, the shell score seemed to top at around 25, which approximately corresponds to 25% of the essay consisting of

shell language. To describe the shell scores and to facilitate subsequent sampling, we defined shell cutscores as seen in Table 1. We computed the shell score for all 212,434 essays.

Table 1

Recoding of Shell Scores Into Nine Categories

Shell score	New shell level
0–3	1
3–6	2
6–9	3
9–12	4
12–15	5
15–18	6
18–21	7
21–24	8
25 and higher	9

Note. The shell score is the output of the computational tool and corresponds roughly with the percentage of the essay that is marked as shell.

Figures 1–4 show several histograms of shell frequency broken down by prompt type and population. Figures 1 and 2 show a histogram for the issue prompt type for U.S. and non-U.S. populations, respectively. As can be seen, the non-U.S. population essays contain more shell language. Figures 3 and 4 show corresponding histograms for the argument prompt type. The argument prompt type seems to elicit more shell than the issue prompt type, and, as before, the non-U.S. essays seem to contain more shell language. This is to be expected; the argument prompt type is concerned with critiquing an existing argument and lends itself more readily to the use of more elaborate shell language.

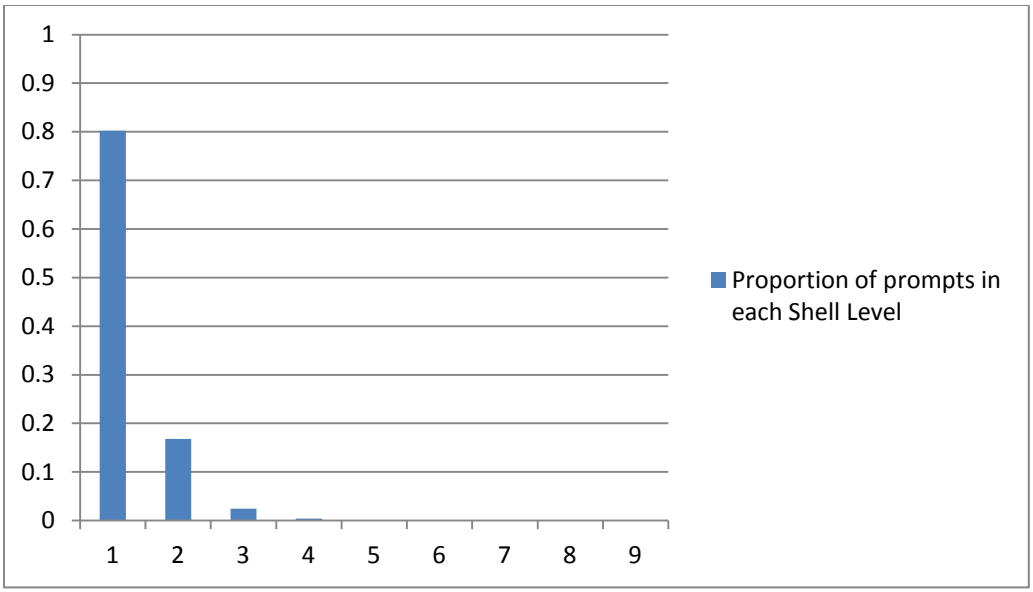


***Figure 1*. Frequency of shell level for issue prompts, U.S. (*N* = 56,788). The *x*-axis refers to the categories of shell score level, as defined in Table 1. The *y*-axis refers to the proportion of essays at each shell score level.**

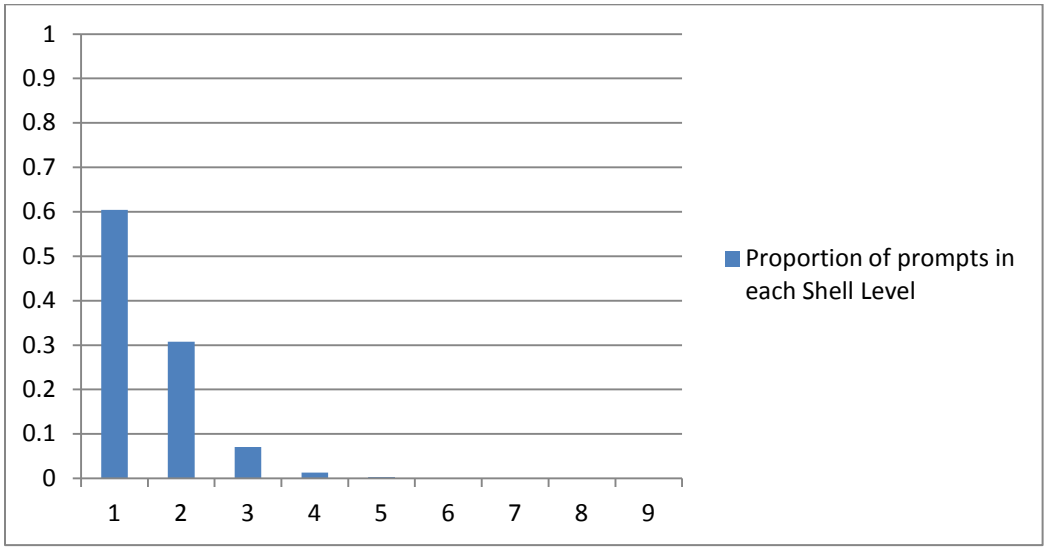


***Figure 2*. Frequency of shell level for issue prompts, non-U.S. (*N* = 43,974). The *x*-axis refers to the categories of shell score level, as defined in Table 1. The *y*-axis refers to the proportion of essays at each shell score level.**

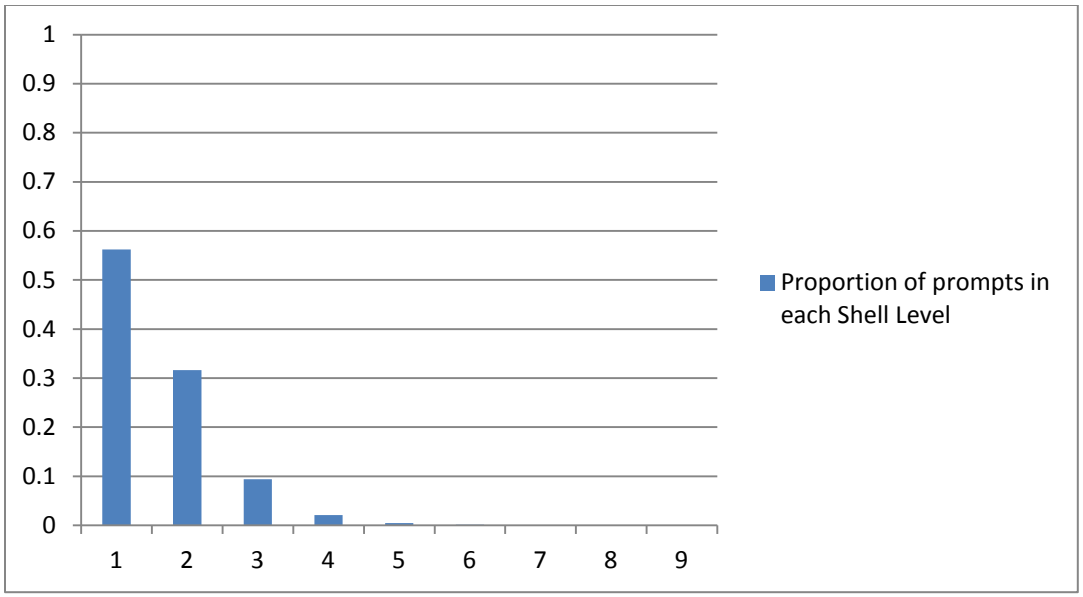


***Figure 3*. Frequency of shell level for argument prompts, U.S. (*N* = 61,549). The *x*-axis refers to the categories of shell score level, as defined in Table 1. The *y*-axis refers to the proportion of essays at each shell score level.**

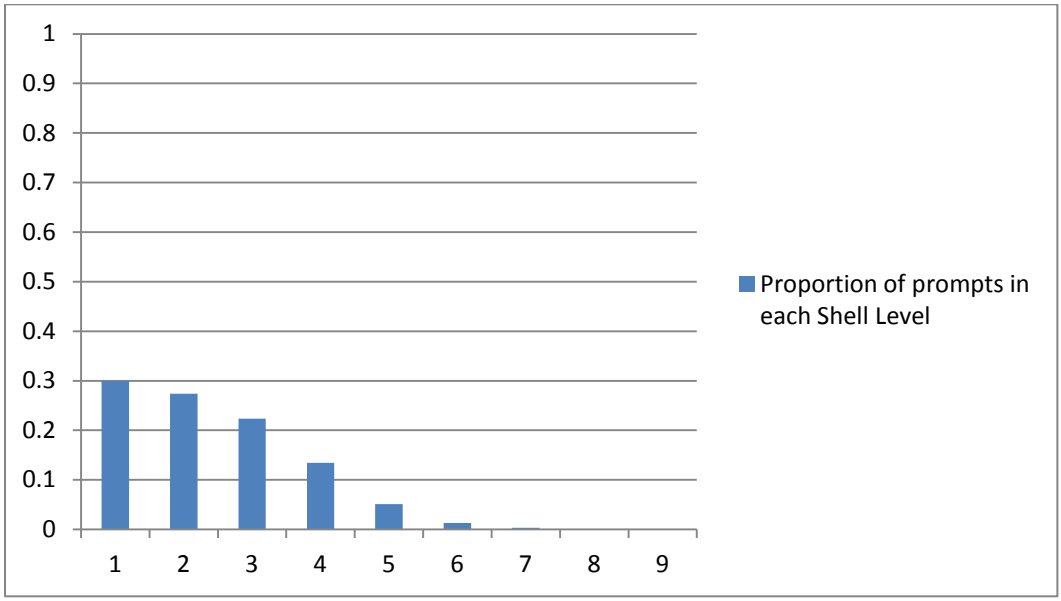


***Figure 4.* Frequency of shell level for argument prompts, non-U.S. (*N* = 50,213). The *x*-axis refers to the categories of shell score level, as defined in Table 1. The *y*-axis refers to the proportion of essays at each shell score level.**

We then examined the distribution of shell scores by prompt separately for both prompt types and both test-taking populations. Figures 5–8 summarize the results by plotting the relative frequencies associated with each interval defined in Table 1 for each prompt for issue and argument by region. The *x*-axis is the prompt. For ease of illustration, the prompts are ordered with respect to the relative frequency at Interval 1 defined in Table 1. Each figure displays the frequency distribution of shell scores for all the prompts in a very economical fashion and show very clear patterns. Note that each of the plots is color coded but the color is not essential to reading the chart. The highest leftmost curve corresponds to Interval 1, the second highest to Interval 2, and so forth.

Perhaps the most noticeable effect is that distribution of shell is not constant across prompts. On the surface, such a trend could be due to different prompts eliciting more or less shell language, perhaps because they have been in circulation for a longer period, or it could be a seasonality effect, because the prompts are rotated and test takers of different skill levels may predominate at different times. However, all the prompts are of essentially the same vintage. Moreover, the prompt effect is consistent across regions. To evaluate if the same ordering was present in the U.S. and non-U.S. population test-taking populations, we plotted the proportion of Level 1 shell for both test-taking populations. The results appear in Figure 9 and 10 for argument and issue prompts, respectively. Each point in Figure 9 represents a prompt. For example, the rightmost prompt shows a little over 45% of the essays for non-U.S. have a shell at Level 1, whereas approximately 80% of the U.S. test takers do. The strong linear trend suggests that the effect of the prompt is similar in both test-taking populations. Figure 10 shows a similar pattern for issue prompts. In short, prompts order themselves with respect to shell roughly in the same manner across two test-taking populations and the reliance on shell is far more prevalent in the non-U.S. population.

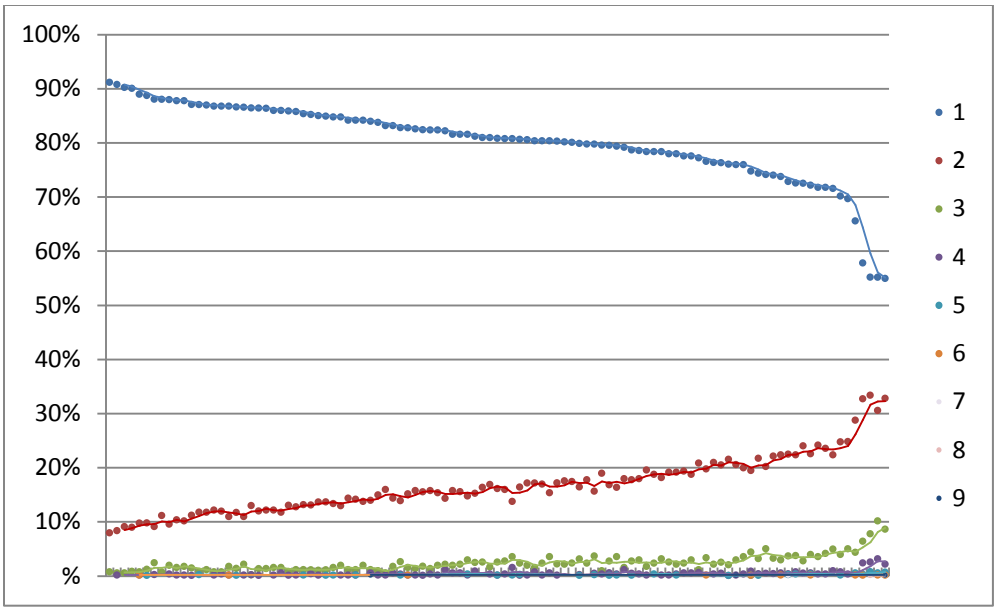


***Figure 5*. Relative frequency of shell language by prompt for issue (U.S.). Each mark along the *x*-axis refers to a different prompt.**

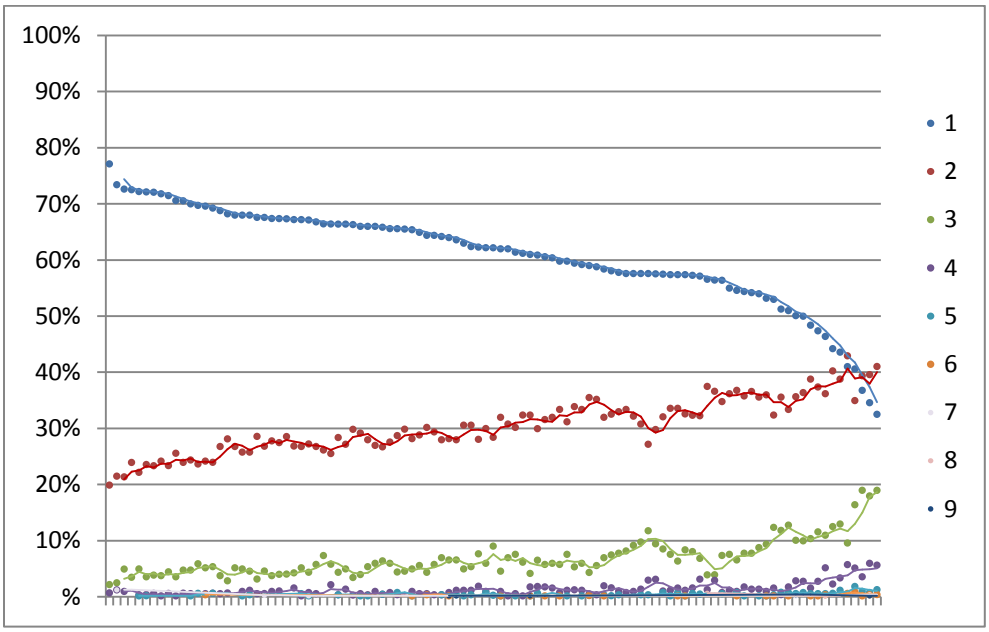


***Figure 6*. Relative frequency of shell level by prompt for issue (non-U.S.). Each mark along the *x*-axis refers to a different prompt.**

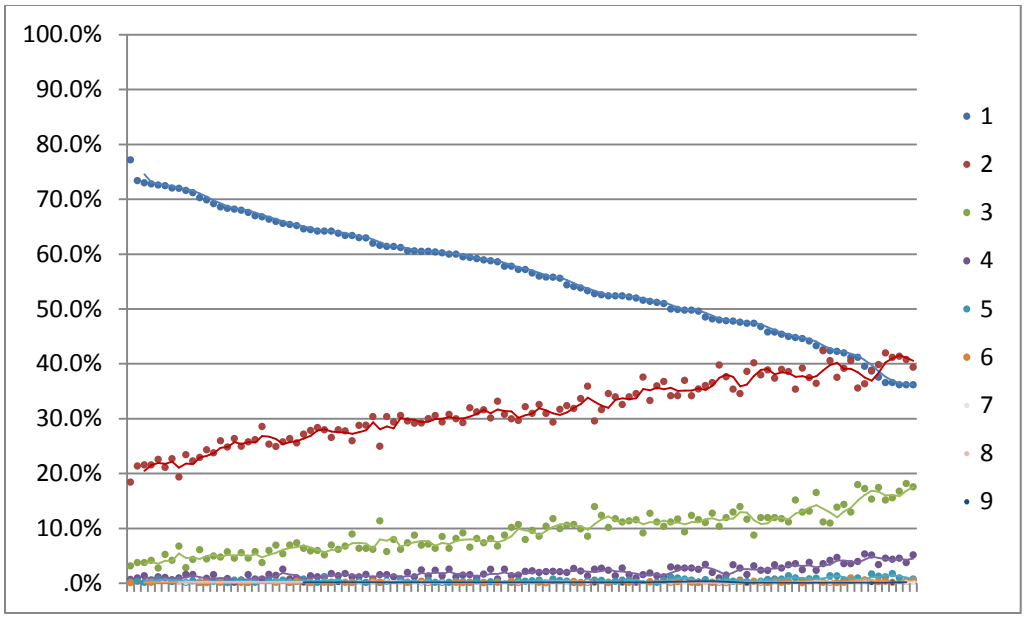


Figure 7. **Relative frequency of shell language by prompt for argument (U.S.). Each mark along the *x*-axis refers to a different prompt.**

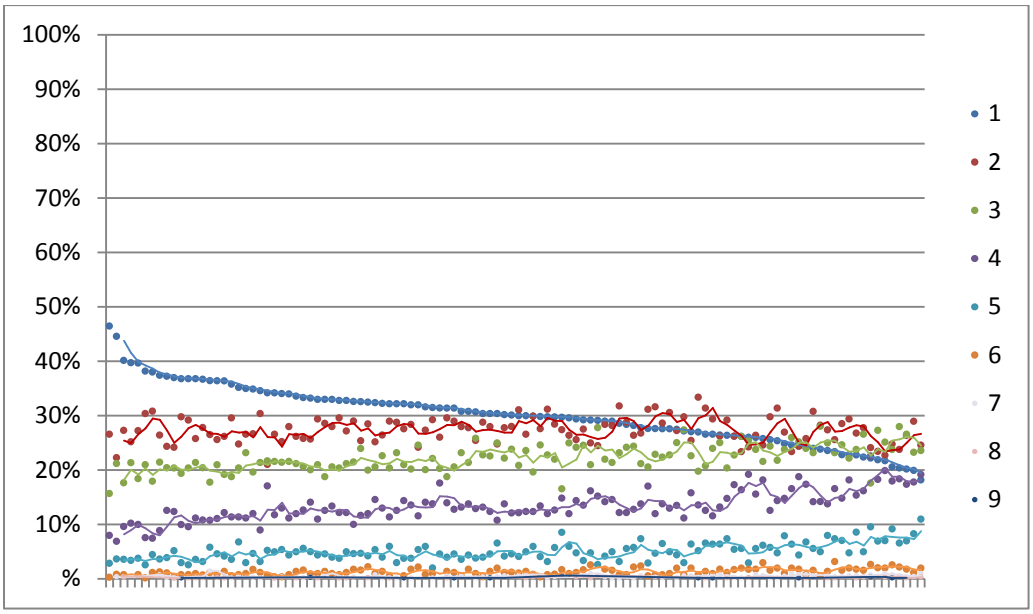


Figure 8. **Relative frequency of shell level by prompt for argument (non-U.S.). Each mark along the *x*-axis refers to a different prompt.**

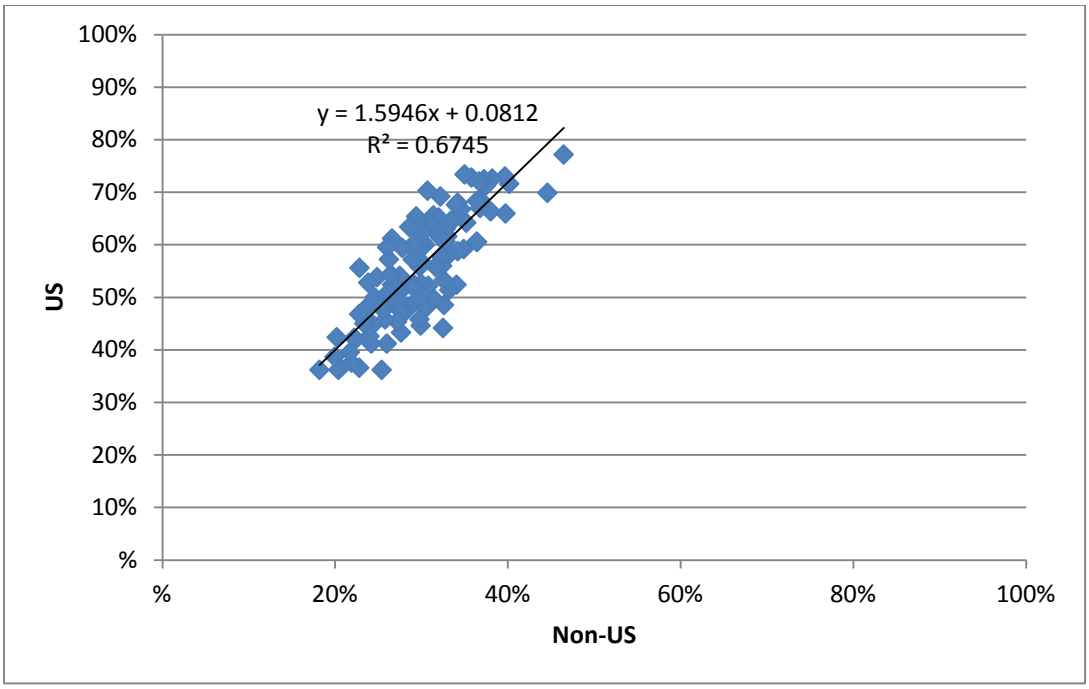


Figure 9. **Scatter plot of Level 1 shell for each argument prompt in U.S. and non-U.S. test-taking populations.**

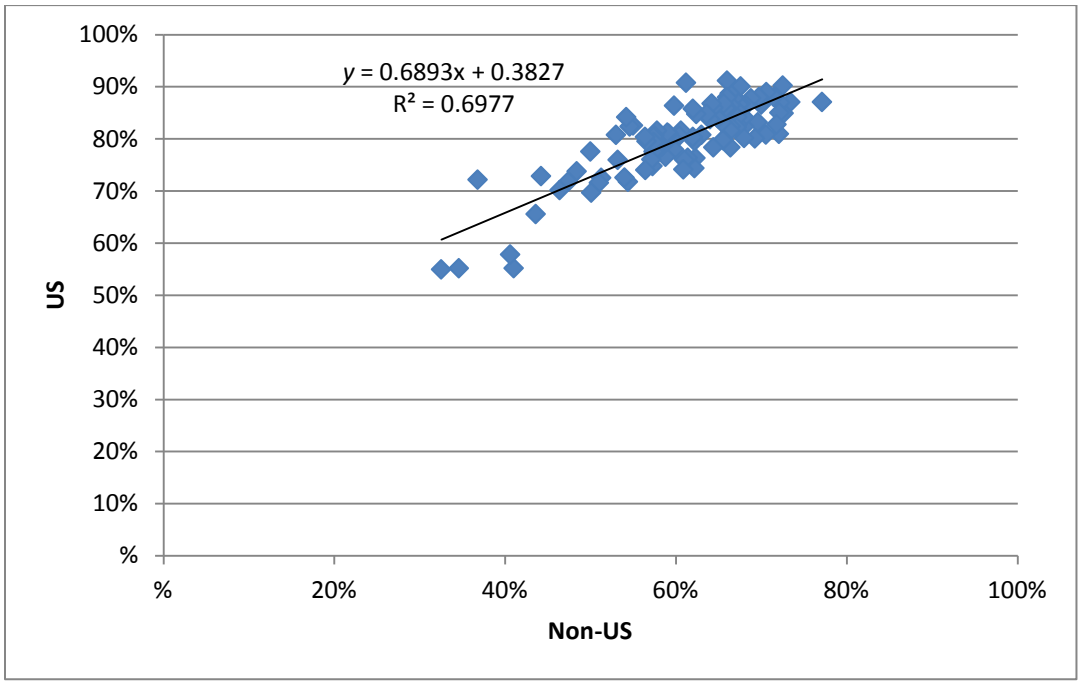


Figure 10. **Scatter plot of Level 1 shell language for each issue prompt in U.S. and non-U.S. test-taking populations.**

Conclusions

The foregoing results support the following conclusions. The presence of shell is higher for argument compared to issue prompts. The non-U.S. test-taking population shell levels are higher than those of the U.S. test-taking population. Finally, the shell scores vary systematically by prompts in the same manner across the U.S. and non-U.S. test-taking populations. These findings add support to the functionality of the shell detection software. The first two findings are as expected. The fact that the presence of shell language varies systematically across prompts was unexpected and only became evident in trying to summarize the distribution of shell across all prompts in a succinct manner.[8]

Validating the Shell Score

Because the shell detection software was still under development when it was used for this study, we conducted a study to evaluate the functionality of the software for our purposes. Specifically, is the computed shell score correlated with the expert scoring leaders' perception of the amount of shell? For that matter, do raters agree on the ranking of essays by amount of shell? Do they agree on what text constitutes shell? That is, there can be agreement on the ranking of essays by the amount of shell even when the two raters or the rater and the computer are not counting as shell the *same text*. For example, the expert could designate a portion of the text as shell while the machine identifies the same amount shell but in a different portion of the essay. Therefore, to reach firm conclusions, it is important that both conditions hold, namely that the expert and the machine agree on the amount of text *and* the text that is to be considered shell. Specifically, the goal of this study was to evaluate whether highly experienced raters agree on the relative amount of computed shell language and what text constituted shell. In addition, we evaluate the same question with respect to the shell software.

Subjects

Two experienced GRE scoring leaders[9] were used in this first study. Scoring leaders within driving distance of the ETS headquarters were invited to participate. Four candidates were identified, but only the two that were available sooner were used in the study. Their task was to rank the essays from four prompts according to the presence of shell language. They were instructed to make a global judgment of the presence of shell language and order the essays accordingly. Importantly, they also highlighted the text they thought was shell. The instructions

given to the raters for ranking the essays are found in Appendix A. In effect, they were asked to categorize essays into eight levels of shell through a sorting method. They were given an overview of the project and asked to read the instructions and raise any questions. They then proceeded to analyze four prompts (two issue and two argument prompts).

Selection of Prompts and Essays for Study 1

Two prompts each for issue and argument were chosen for the study. When deciding which prompts to use for Study 1, we found that some of the prompts did not have any or enough essays at the shell level of 6 and higher and that issue prompts had, comparatively speaking, much less shell language. Thus, when deciding which prompts to use, we limited our pool to only those prompts that had essays with a shell level of six and higher.

In total, there were 114 argument prompts and 105 issue prompts to choose from. The prompts were sorted on the percentage of shell at Level 1, from highest to lowest. We initially selected four issue and four argument prompts for a total of eight prompts. However, the scope of the investigation had to be reduced and we chose four prompts. Because the issue prompt type elicits less shell that is detectable by the tool, we essentially chose two issue prompts that had the highest shell scores as determined by the shell detection software. For the argument prompts, we chose prompts from the middle of the distribution.

Choosing Essays

The next step was to select essays from the selected prompts to use in the study. For each prompt, we created a grid to select essays in order to identify the number of essays available from each prompt at each human score level (1–6) as well as at each shell level (1–6), where the sixth category is a collapsing of Shell Levels 6–9. This resulted in a total of 3,456 essays in our pool. We selected manually the 24 essays needed per prompt. We selected four essays per shell level (1 through 6) while trying to obtain different human scores (from 1 to 6) at each shell level. We tried to select essays from both testing regions; however, this was not always accomplished. In the end, 24 essays for each of the four prompts were obtained, for a total of 96 essays.

Results Study: Agreement on Ranking

Do scoring leaders agree on a ranking of the essays by shell? Tables 2 and 3 show the cross-tabulation of the rankings for the two scoring leaders for the 48 issue essays and the 48 argument essays.

Table 2

Cross-Tabulation of the Rankings of Essays by Shell Language for Argument Essays

	Rater 2 ranking recoding						
Rater 1 ranking recoding	1	2	3	4	5	6–9	Total
1	3	3	0	0	0	0	6
2	3	1	0	2	0	0	6
3	0	1	2	1	2	0	6
4	0	1	2	0	1	2	6
5	0	0	0	2	7	3	12
6–9	0	0	1	1	1	9	12
Total	6	6	5	6	11	14	48

Table 3

Cross-Tabulation of the Rankings of Essays by Shell Language for Issue Essays

	Rater 2 ranking recoding				
Rater 1 ranking recoding	1	2	3	4	Total
1	5	0	0	0	5
2	3	2	1	0	6
3	0	3	6	4	13
4	0	1	5	18	24
Total	8	6	12	22	48

Tables 2 and 3 show that raters agree, to some extent, on the ranking of essays by level of shell. For both issue and argument essays, the kappa measure of agreement was significantly different from 0. For argument, kappa was .33 and for issue, kappa was .47. For both issue and argument essays, the weighted kappa measure of agreement was significantly different from 0. For argument, the weighted kappa was .80 and for issue, weighted kappa was .82.

Agreement on What Is Shell: Overlap

As noted, agreement on the amount of shell is potentially ambiguous because what is being called shell by two raters could be entirely different text. Therefore, we also computed an overlap measure between any two sources or annotators. Here, we are interested in the overlap

between the two raters and the raters and the computer. Essays are marked up during the computation of the shell score to designate where shell begins and ends as described previously. A similar markup was done for the highlighting completed by the human raters.

For purposes of computing overlap, an essay was represented as a 0/1 vector of length equal to the number of words in the essay, such that the word occurring in a portion of the text marked up as shell would be coded a 1 and 0 otherwise. With the text represented in this fashion, overlap was computed as the dot product of two vectors representing the same essay marked up by two sources.

Let o_r designate overlap, where r designates one source either a human rater or the machine, and r' a different source. Let also t_r be the 0/1 vector of length *l* representing the text such that

t_r = if the *i*th word is marked up as shell;

t_r = 0 otherwise.

Overlap is the number of words in common designated as shell is defined as the dot product of two vectors,

$$o_{r,r'} = t_r \,.\, t_{r'}$$

$$= \sum_{i=1}^{l} t_r t_{r'}.$$

In this form, overlap refers to the *number* of words that have been designated in common as shell by *r* and *r′*. We express that number relative to the total number of words that were marked as shell by either source,

$$s_{r,r'} = \left(\sum_{i=1}^{l} t_r + \sum_{i=1}^{l} t_{r'}\right) - o_{r,r'}.$$

Finally, *overlap* is defined as the proportion of words designated as shell by either source,

$$Overlap_{r,r'} = o_{r,r'}/s_{r,r'}.$$

Figure 11 reports that the relationship of the overlap of each rater with the shell software for both argument and issue prompts goes as high as 50%. Nevertheless, the level of overlap

varies and can be 0 because there may not have been any shell indicated, in which case the overlap is 0. It is reassuring, however, that the overlap of the shell scores produced by the software with two different raters is positively related with an *R* squared of .53.

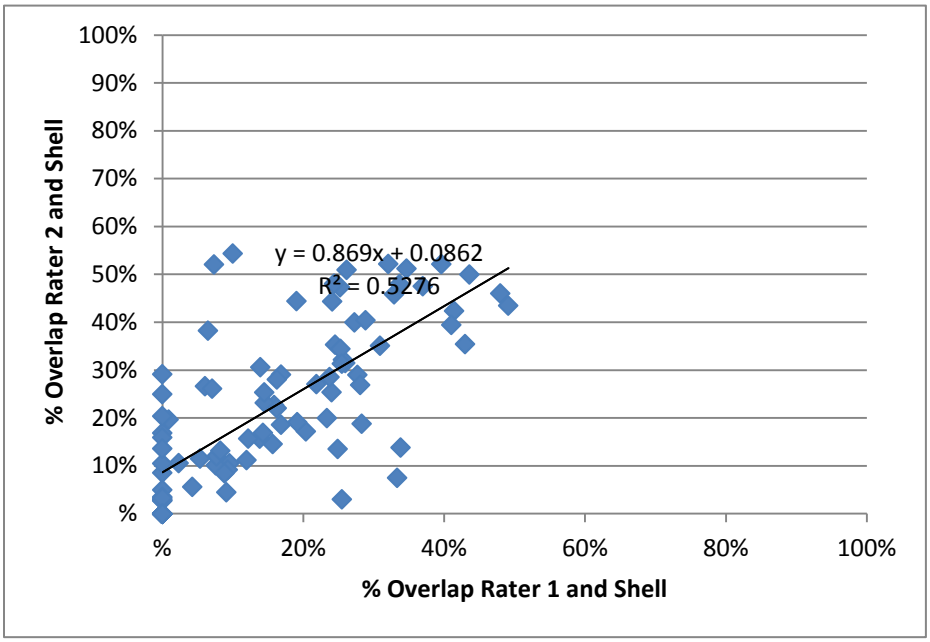


Figure 11. **Relationship of the overlap of shell language between machine and rater for two raters.**

Conclusions

We conclude that the two scoring leaders agreed substantially on the ranking of essays with respect to the presence of shell. The weighted kappa agreement index was significantly different from 0 and very high. Moreover, the overlap in the designation of shell language by the software and what the raters marked up as shell in a positive fashion across the two raters. Therefore, the software tool can be relied on to select essays that vary in the amount of shell they contain.

Main Study

The foregoing analysis was necessary to rely on the shell identification software for the main part of the study. The primary question in this study is whether raters appropriately treat shell language neutrally when scoring under operational conditions. Specifically, are operational

scores higher as a function of the amount of shell language? One possibility is that under operational conditions scorers are overly influenced by the presence of shell language and therefore assign inappropriately higher scores than under the more leisurely conditions of the present study. To answer that question, the 96 essays selected for this study were rescored by experienced raters different than those involved in the first study. We examine the score discrepancies between operational scores and the rescoring from this study.

Raters

Two experienced GRE scoring leaders, different from the two used in the previous study, were used in this study. Scoring leaders within driving distance of the ETS headquarters were invited to participate. Four candidates were identified but only the two that were available sooner were used in the study.[10] After the scoring leaders completed their work, they were debriefed. Appendix B contains the scoring directions. Because the two raters were very experienced, the scoring proceeded very smoothly. Both raters were able to complete the rating assignment in approximately 5 hours. After the scoring was completed, the raters were debriefed. Appendices D and E contains transcripts of the debriefing interviews.

Choosing Prompts and Essays

The same prompts used in Study 1 were used in this study. However, a different set of essays was used. To choose essays, the 1–2 and 5–6 writing score levels were collapsed. This left the essays sorted into four score levels: 1–2, 3, 4, and 5–6. The shell score categories in Table 1 were collapsed as well into three levels: 1–2, 3–4, and 5–9. Tables 4 and 5 show the number of essays at each of the writing and shell score levels. The goal was to sample randomly two essays from each cell for a total of 24 essays per prompt. As can be seen, in cases where there were fewer than two essays in a cell, an essay from an adjacent cell was borrowed. This was only necessary in cases with a shell score of 5–9. In effect, this approach over-represents essays with high levels of shell language but not to an extent that would make the set of essays atypical from the point of view of the raters. The resulting sets of 24 essays from each prompt were used in both studies described below.

Results

The essays were rescored as part of this project under more leisurely conditions to reduce the possibility that shell language could influence the scores as a result of raters rushing thorough the evaluation of the essays. Although the interrater agreement was high at a weighted kappa of .82, there were also systematic differences among the two scoring leaders. Specifically, there were 11 discrepancies at the 3–4 score range out of 31 discrepancies at all score levels. Such discrepancies are, in a sense, more consequential because there is a qualitative highly construct-relevant distinction between the upper half of the score range (Scores 4–6) and the lower half (Scores 1–3). In an effort to obtain scores as free of error as possible, we had these differences adjudicated and annotated by the ETS Assessment Development department staff.

Table 4

Number of Argument Essays at Each Writing and Shell Score

	Shell score					
	Prompt 1			Prompt 2		
Writing scores	1–2	3–4	5–9	1–2	3–4	5–9
1–2	52	22	2	37	23	8
3	95	85	34	85	89	30
4	69	27	2	67	42	4
5–6	28	15	1	37	9	1

Table 5

Number of Issue Essays at Each Writing and Shell Score

	Shell score					
	Prompt 1			Prompt 2		
Writing scores	1–2	3–4	5–9	1–2	3–4	5–9
1–2	38	49	10	48	16	5
3	98	91	4	100	83	7
4	77	28	1	87	38	3
5–6	33	3	0	39	6	0

For 10 of the 11 discrepancies in the 3–4 score range, the ETS Assessment Development staff sided with one of the scoring leaders, which suggests that the lower-/upper-half distinction is not equally understood by all scoring leaders. In addition, for nine of the 11 discrepancies in

the 3–4 score range, the assessment developers agreed with the lower score of 3. Again, this suggests the two scoring leaders do not have the same conception of the attributes that discriminate between essays in the 3–4 score range.

Table 6 summarizes the relationship of the discrepancies between operational and adjudicated shell scores as computed by the shell tool. We have reason to believe, based on the earlier study, that these discrepancies are largely due to raters' conception of what shell language is. Based on these results, the most reasonable conclusion is that the discrepancies between operational scores and the adjudicated scores are *not* due to the presence of shell, because the shell scores for cases with no discrepancy is somewhat higher. There appears to be, however, a leniency effect in the operational scores. Table 7 summarizes the operational scores, the scoring leaders' scores, and the adjudicated scores. As can be seen, for issue essays, the mean operational score is 3.33 whereas the adjudicated score is 3.17. Similarly, for argument essays, the mean operational score is 3.46, whereas the adjudicated score is 3.20.

Table 6

Mean Shell by Discrepancy Between Adjudicated and Operational Scores

	Discrepancy											
	-1			0			1			2		
Prompt	*N*	Mean	SD	*N*	Mean	SD	*N*	Mean	SD	*N*	Mean	SD
Issue	5	2.60	1.52	32	3.63	1.79	9	3.33	1.32	2	3.5	0.71
Argument	5	3.40	0.89	27	3.56	1.70	15	3.47	1.51			

Note. Shell score ranges from 1 to 9. Discrepancy was computed as H1 minus the adjudicated score, where H1 is the first operational scoring.

Table 7

Descriptive Statistics for Scoring Leaders Scores (SC1 and SC2), Operational Score, and Adjudicated Scores for Both Argument and Issue Prompts

	Operational score			SC 1			SC 2			Adjudicated scoring leader score		
Prompt	*N*	Mean	SD	*N*	Mean	SD	*N*	Mean	SD	*N*	Mean	SD
Issue	48	3.33	1.17	48	3.29	1.01	48	3.48	1.11	48	3.17	0.93
Argument	48	3.46	1.20	47[a]	3.17	0.87	48	3.35	0.96	48	3.20	0.94

[a]Due to a missing essay in one of the packets, only one scoring leader scored the essay.

Summary and Conclusions

We relied on a computational tool under development (Madnani et al., 2012) to quantify shell language. We applied the tool to a large data set of GRE essays and conducted a study to validate the results from the tool. Results were presented that suggest that the computational tool worked well as evidenced by the agreement of the tool with raters as to the amount of shell and what text constitutes shell. The distribution of shell computed by the tool follows the expectation that the argument prompts elicit more shell language than the issue prompts. Moreover, consistent with anecdotal reports, we found that the use of shell language is more common in test-taking populations outside of the United States.

To evaluate the possibility that shell language could influence operational scores, we had essays rescored and adjudicated under nonoperational conditions. The discrepancies between operational and adjudicated scores were not a function of shell level. We conclude from that result that GRE scorers have been adequately trained to handle shell. We reach that conclusion by assuming that the adjudicated scores are as close as we can get to the true scores the essays deserve and that the discrepancies between operational and adjudicated scores obtained as part of this study were not related to the amount of shell.

However, the study did find that adjudicated scores were lower than the operational scores for both issue and argument prompts. Because the data was from test takers from 2006 to 2010, it is ambiguous whether such a difference is a leniency effect on the part of raters or a change in the scoring standards or some other form of drift. However, the annotation of the discrepancies at the 3–4 range revealed that the upper-/lower-half distinction may not be equally understood by all scoring leaders. Specifically, one scoring leader consistently over-rated papers that merited only a score of 3 based on the adjudicated scores. Because the 3–4 score range is the most frequent, such consistent discrepancies could account for the leniency effect we observed if, in fact, not all raters have the same understanding of the distinctions and the raters who are lenient predominate. A validity argument for writing scores would be weakened accordingly under those conditions. However, evidence that suggests that scorers are immune to the presence of shell language would contribute positively to a validity argument.

Although the study was based on GRE essays, the study was motivated by the impending increased use of performance measures and the use of automated scoring in school-based testing. In that context, the use of construct-irrelevant response strategies is also a potential threat to

validity. The results presented here cannot be generalized to a school-based context but the study does serve to illustrate the type of validation analysis that would be useful in that context, given that automated scoring is expected to help in handling the volume of constructed responses.

References

Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater® V.2. *Journal of Technology, Learning, and Assessment*, *4*(3). Retrieved from http://ejournals.bc.edu/ojs/index.php/jtla/article/view/1650/1492

Bejar, I. I. (2012). Rater cognition: Implications for validity. *Educational Measurement: Issues and Practice, 31*(3), 2–9.

Bejar, I. I., Flor, M., Futagi, Y., & Ramineni, C. (2012). *Effect of a construct-irrelevant response strategy (CIRS) on automated scoring of writing.* Manuscript submitted for publication.

Bennett, R. E. (2011). *Automated scoring of constructed-response literacy and mathematics items.* Washington, DC: Arabella Philanthropic Advisors. Retrieved from http://www.ets.org/s/k12/pdf/k12_commonassess_automated_scoring_math.pdf

Hillocks, G., Jr. (2011). *The testing trap*. New York, NY: Teachers College Press.

Jurafsky, D., & Martin, J. H. (2000). *Speech and language processing*. Upper Saddle River, NJ: Prentice Hall.

K-12 Center at ETS. (2011). *The SBAC supplemental application.* Princeton, NJ: ETS. Retrieved from http://www.k12center.org/rsc/pdf/sbac_supplemental_summary.pdf

Madnani, N., Heilman, M., Tetreault, J., & Chodorow, M. (2012, June). *Identifying high-level organizational elements in argumentative discourse*. Paper presented at the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Montreal, Canada. Retrieved from http://www.aclweb.org/anthology-new/N/N12/N12-1003.pdf

Partnership for Assessment of Readiness for College and Careers. (2010). *The Partnership for Assessment of Readiness for College and Careers (PARCC) application for the Race to the Top Comprehensive Assessment Systems competition*. Retrieved from http://www.fldoe.org/parcc/pdf/apprtcasc.pdf

Powers, D. E. (2005a). *Effects of preexamination disclosure of essay prompts for the GRE Analytical Writing Assessment* (Research Report No. RR-05-01). Princeton, NJ: Educational Testing Service.

Powers, D. E. (2005b). *"Wordiness": A selective review of its influence, and suggestions for investigating its relevance in tests requiring extended written responses* (Research Memorandum No. RM-04-08). Princeton, NJ: Educational Testing Service.

Quinlan, T., Higgins, D., & Wolff, S. (2009). *Evaluating the construct-coverage of e-rater® scoring engine* (Research Report No. RR-09-01). Princeton, NJ: Educational Testing Service.

U.S. Department of Education. (2009). *Race to the Top program executive summary*. Washington, DC: Author. Retrieved from http://www2.ed.gov/programs/racetothetop/executive-summary.pdf

Notes

[1]There is no shareable document describing guidelines for shell annotation. However, we used the same guidelines that were developed by Bejar, Flor, Futagi, and Ramineni (2012) for the software used in this project.

[2] Scorers and raters are used interchangeably in this report.

[3]Certain measures are already in place in admissions testing, such as essay similarity detection in the case of essays responses (http://www.ets.org/gre/institutions/scores/how/).

[4]A revised Analytical Writing measure was included with the GRE revised General Test in 2011. Unless otherwise noted, *GRE essays* refers to essays from the *previous* Analytical Writing measure.

[5]*Operational scores* here refers to two human scores.

[6]A regular expression is a language for concisely expressing a text pattern. Those patterns are then the basis for searching text (Jurafsky & Martin, 2000). Complex regular expressions can be formed by using several operators.

[7]The set of rules used in this study was a precursor to a more elaborate system (Madnani et al., 2012) that goes beyond matching rules to designate text as shell.

[8]A potential explanation is related to operational details on how the test was administered outside of the United States.

[9]A scoring leader is the most experienced rater and supervises other raters.

[10]The data collection took place in August 2011, immediately after the introduction of the GRE revised General Test. Because the writing component of the GRE revised test is substantially different, we were interested in completing the study as soon as possible to prevent potential contamination of the scoring process. (See the transcription of the debriefing of the two raters in Appendix D and E. *Both raters indicated that they experienced no confusion*.)

List of Appendices

Appendix A

Instructions for Shell Validation Study

August 9, 2011

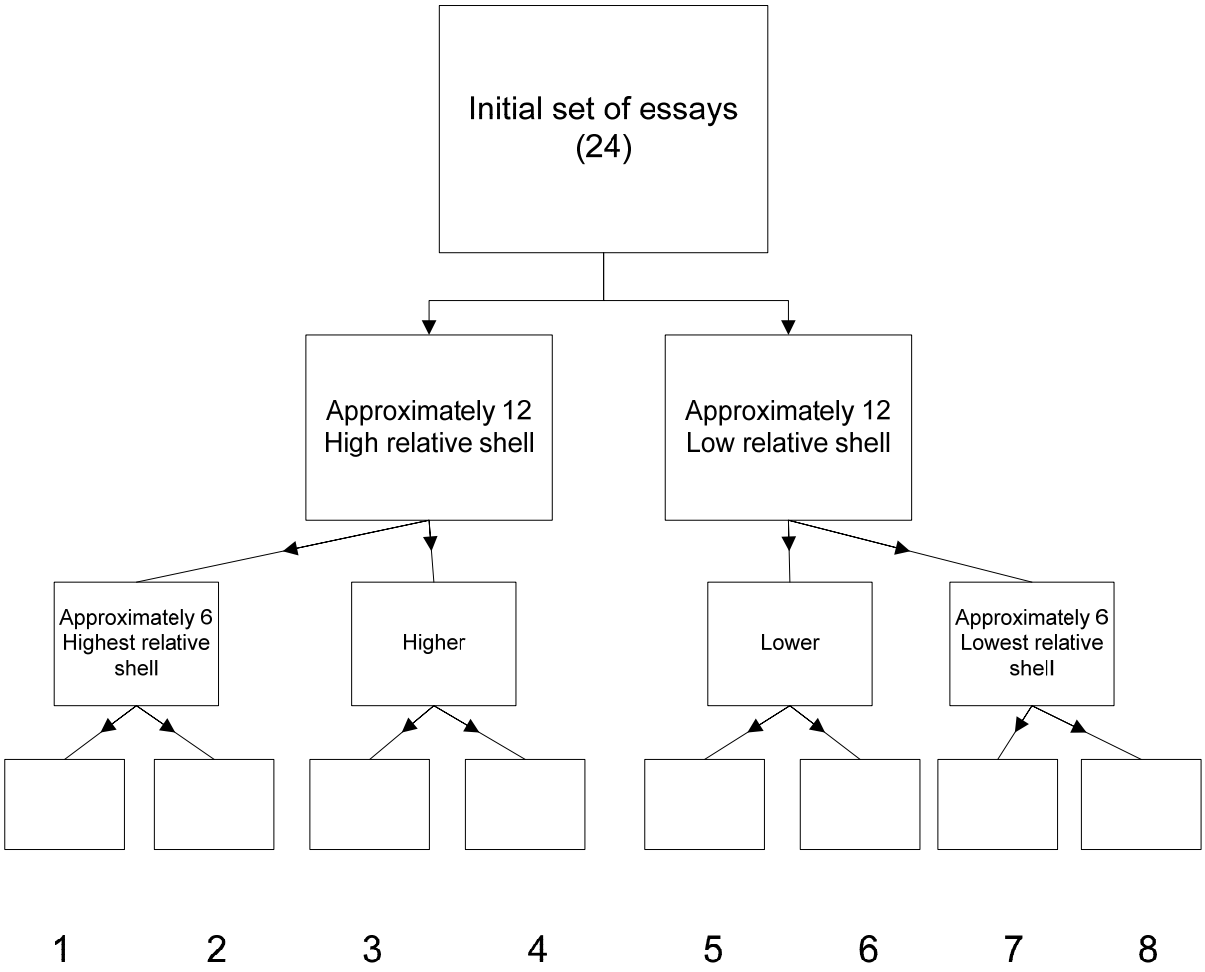
We are doing a special study of eight (4 issue/4 argument) prompts to determine the prevalence of shell language. Your role in this study is to make a holistic determination of the relative amount of shell language by ranking the essays from each of the prompts separately**. *For purposes of this study, you are not scoring the essays and you should ignore the quality of the response.***

You will rank the essays from each prompt according to what you perceive to be the *relative proportion* of shell language across the essays. That is, you should take into account the length of the essay in your judgment. For example, given two essays with the same amount of shell language, you should rank the shorter one higher than the longer one because it *proportionately* contains more shell language.

Initially, the essays are not ordered in any way. Please follow this procedure. As your read each essay, first put the essays into "high" and "low" piles of approximately the same number of essays each such that all the "high essays contain more shell than the "low essays." Then, repeat the process with "high and "low" piles. At this point you should have four piles. Then repeat the process one more time by taking each of the piles and splitting them into a high and low. After the third round you should have eight piles. The figure below describes the process graphically assuming 24 essays per prompt.

As a final step, go over each pile and, if you feel it is appropriate, fine tune the assignment of essays to piles.

When you have completed the process for a prompt, please enter your ranking for each essay on the sheet provided to you. On the top left-hand corner of each essay there is a Person ID (e.g., 9995000000758340). On the top right-hand corner is the Prompt ID. Please make sure to use the Ranking sheet that has the correct prompt ID (at the top of the page). On the Ranking sheet, please write your name at the top of the page, write down the Person ID in the Person ID column, and write your ranking for each essay.



By the end of the third round you should have 8 piles with approximately 3 essays each. Please assign the ranking 1–8 to each of the 24 essays. You will be given a ranking sheet with the essay IDs. Simply write down the corresponding ranking (1–8) for each of the 24 essays.

***Figure A1*. Essay ranking.**

Appendix B

Scoring Instructions

The purpose of this study is to evaluate the accuracy of operational GRE writing scores. The study is based on the "old GRE" and, therefore, you should score these essays based on the old rubric as you normally would have scored the essays. Please use the old rubric for issue and argument (included in your materials) as you score the essays today.

To accommodate the fact that different readers reading at different rates, we have 24 essays from two issue and two argument prompts (96 essays altogether) for you to score. Please pace yourself accordingly to complete this during the course of the day. In this study we are focusing on the accuracy of scoring over productivity.

After you have completed the scoring, I will return to ask you a few questions about your participation in the study.

Your materials include:

- Scoring rubric
- The prompts
- The essays
- Scoring sheet

Do you have any questions?

Appendix C

Debriefing Process

Thank you for your participation.

Since we will continue data collection for a few more weeks, please refrain from discussing the study with your colleagues.

The study is concerned with the different factors that affect the accuracy of scoring, including the use of "shell." Some of the essays you scored were selected to better understand those factors. The scores that you and your colleagues will provide will be considered as potentially more accurate than scores obtained under normal circumstances.

Do you have any questions?

To complete the study I'd like to ask you a few questions and record your answers, if you don't mind. Your comments will be transcribed and not attributed to you personally.

1. Were you able to score these essays without interference from the scoring instructions for the GRE revised test?
2. Compared to the OSN scoring of the "old" GRE, was this process similar, different? How so?
3. How typical or atypical were the folders you scored compared to old GRE folders?
4. What is your understanding of shell language in general, and with respect to issue and argument prompts?
5. Did the essays you scored contain shell?
6. Roughly speaking, what proportion of the issue and argument essays you scored do you think contained a significant amount of shell?
7. Can you describe the strategy that you used to score essays that contained shell?
8. What role, if any, does the length of the essay you are about to score play in your strategy for scoring issue and argument essays?

Appendix D

Debriefing of Subject A

This is a transcript of the researcher (R) debriefing Scoring Leader 1 (SL1).

R: This is [researcher name], debriefing [Scoring Leader 1]. It is 1:18 [pm]. Thank you for your participation. Since we will continue the data collection for a few more weeks, please refrain from discussing the study with your colleagues.

SL1: Absolutely.

R: The study is concerned with different factors that affect accuracy of scoring, including the use of shell. Some of the essays you scored were selected to better understand those factors. The scores that you and your colleagues will provide will be considered as potentially more accurate than scores obtained under normal circumstances. Do you have any questions?

SL1: No.

R: To complete this study, I would like to ask you a few questions and record your answers. If you don't mind, your comments will be transcribed but not attributed to you personally.

SL1: OK.

R: So, the first question. Were you able to score these essays without interference from the scoring instructions for the revised GRE?

SL1: Yes.

R: There was no interference?

SL1: No.

R: Compared to the OSN scoring of the old GRE, was this process similar, different, and how so?

SL1: Compared to—sorry, say that again.

R: The old GRE, in other words, referring to the old GRE as opposed to the revised GRE. . . was the process similar or different and how so?

SL1: I think [in] the old process of GRE we were looking at how the student responded to the prompt and to those task directions, and with the new GRE we have to keep in mind that the task—the new task direction, the new variants—have more consideration in how you determine the score. In other words, the essay could possibly be well written and even well analyzed on some level but not following the task directions, which would affect the score. So I think, in keeping that in mind, the two things are a little different.

R: Right, comparing the old GRE to the revised GRE, the process you went through, which obviously is paper and pencil—compared to the old GRE, how is it similar?

SL1: It is the same process. We use the same principle when we are doing paper/pencil for GRE under the old GRE, and when we are on OSN, it is still the same process—you read the essay, you determine whether the essay meets upper half- or lower-half criteria, and then you determine, "Well, if it is lower half, where does it fit in—1, 2, or 3?" If it is upper half—4, 5, or 6—so you go through the same process.

R: So this process you describe as upper and lower half—is that sort of an official component of scoring or is that your own strategy?

SL1: Well, we have through the time period of doing this. There have been different test development strategies given to us as scoring leaders and that is usually how they approach a score—that it is pretty easy to determine from reading the essay usually whether it is a lower-half or an upper-half paper and then deciding where it fits in there.

R: So it is a big decision to sort of commit to upper half or lower half—once you commit to that decision you stick to it, or are there instances where you actually retract that decision?

SL1: Yes, when you, for me, anyway, making a decision whether a paper could be upper half or lower half is largely part of language skill. So if you're reading a response that has persuasive errors, a number of errors in it, you—it is not going to be an obviously upper-third paper, which needs to have facility of language and skill of language. It may not even be a 4—adequate—because if it affects the clarity, then it is not going to even be a 4, so that the first thing you notice [is]…that there are continual problems with language control. Then you probably are not looking at an upper-half paper so then you determine

other aspects of the scoring guide to see where it falls in that score and on the 3, 4 line. You can sometimes hesitate to make a decision—upper-half or lower-half based—but that is not usually [based on] language control. Then it is usually a matter [of] how is the analysis—how well developed is the analysis—than it is more of a content question, usually. How much elaboration, how much development is considered adequate as opposed to perhaps insufficient or limited? So in that case, it may be [a] factor.

R: How typical or atypical were the folders you scored, compared to the old GRE folders?

SL1: Well, lately there seem to be a lot of papers that have persistent language problems.

R: Lately, meaning?

SL1: Lately, meaning, I would say, in the last 6–7 months or I would say even in the last year, there just seems to be a greater population perhaps of test takers around the world, and so we get a higher population sometimes of people that don't quite have mastery over basic grammar and things like that. There also seems to be a lot [of] influence with places that coach students on how to respond to the essays. So we do see a greater number lately, I would say, of frameworks that are given to students on how to approach the GRE, and so you see a lot of that kind of language. That organizational framework comes back, so that [would have] been, I would say, in more recent [years] than perhaps than when we first started doing this in the late 1990s.

R: So are you able to recognize those frameworks that you refer to?

SL1: Yes. I mean I am not saying I know all of them, but, yes, they are pretty noticeable. I mean it is pretty obvious, and some people can use them and use them quite well. I mean it is not that the framework itself [that] determines the score of the paper, because if the student is able to use that framework effectively, it can still turn out to be a well constructed paper, but they are more noticeable in argument than in issue.

R: What is your understanding of shell language in general and with respect to issue and argument prompts that you just mentioned?

SL1: Ok. Well, as English teachers, we all teach our students to outline and how to use an organizational framework to present their ideas, so it is not that different. (Are we running out of time? No? Ok.) It is not that different from what the coaching schools are

telling their students, but it is kind of obvious to see that they are telling them to take, rephrase the prompt and then say the prompt is wrong or that there [are] flaws in the prompt or somehow that this is not effective and then come up with some reasons why it is flawed. But if their own support for those ideas isn't constructed as well as the framework is, then you can tell that the person has memorized that framework but does not have the skills to do anything with it.

R: And you are able to determine that?

SL1: I think, after a certain amount of experience with them—I think probably newer raters who are unaccustomed to seeing that kind of framework would be fooled by it sometimes because the paper appears to be organized and appears to be following the task. But I think that an experienced rater would certainly see that, usually, I would say.

R: Did the essays that you scored contain any shell?

SL1: Yes, they did—especially the argument ones were pretty noticeable.

R: Roughly speaking, what proportion of the issue and argument essays you scored do you think contained a significant amount of shell?

SL1: I noticed that almost all of them in the argument did contain shell framework, so I would say it was very noticeable on the argument essays.

R: Can you describe the strategy that you use to score essays that contain shell?

SL1: The strategy I use, well again we, I, first of all, read the essay, OK? And then when I am trying to make the determination on the score, I will look to see how much of the [writing supports the framework], how was that done well, was it reasonable, was it valid even or was it just clutching at straws to try to come up with vague and general support? Then, if the writing in spite of the shell is still valid, relevant to the prompt, and is still written clearly, then you know the score. That's fine, I mean the shell itself doesn't influence the score but to the extent that the writer used the shell as a crutch even as a smoke screen, as opposed to just utilizing it effectively, that makes the determination.

R: What role, if any, does the length of the essay that you are about to score play in your strategy for scoring issue and argument essays?

SL1: OK, well, we are told time and time again not to let length be a determining factor but the facts of the matter are that it is impossible to get a 6 when you are writing three or four sentences. So to say that length is not a quality or characteristic is not true, OK, because you know 5s and 6s are, by necessity, almost are going to be longer than, say, typical 1s and 2s. However, we all know—and we have seen ample, ample evidence—that the fact [is] that you can have a two-page 1. I mean, you can have a long, long 1. You can have a long , long 2. Any of those score point values can be very, very long, so, in that sense, length does not make the determination. In terms of getting raters to see the quality of the development and the quality of the elaboration and the level of detail that—comes [with] experience, I think, and it is a little tricky because very, very precise details and very, very cogent examples or reasons—a person who can skillfully use words isn't penalized either for coming up with a very, very cogent response and deserves a 5 or 6. That's a good thing, so more words don't necessarily mean better words so the level of vocabulary, variety, all of that is a factor as well and even using examples, if they come up with some typical examples like everybody uses—Bill Gates, Martin Luther King—things like that. But if that example isn't connected in any way to the prompt or to the support that the person should be using for their argument, then it is not a relevant example, so the amount of time they spent on it doesn't necessarily mean it was good development or a good elaboration….[It] comes, too, with experience and with looking at lots and lots of different papers. And the benchmarks and the range finders sometimes help raters see some of those things to watch for, so all that is [under] consideration.

R: Can you say more about range finders and benchmarks: Are those given at one point in time? Are they always available to you?

SL1: With the old GRE, we [had] a shortage of sample papers…but in the old GRE, we had a set of benchmarks, which were typical middle-of-the-road, common ways that writers might approach an example for each of the score points. So you have Benchmarks 6, 5, 4, 3, 2, 1, in theory. They weren't always there for every prompt and then you would have a series of range-finder papers because each score point has a continuum; there are good 5s that are not quite 6s but are very, very good 5s. And then there are 5s that are just a little bit better than 4s, and they meet the criteria for 5s so there is always a range and the

range finder would kind of help to see where some of those boundaries might be drawn and exemplify those different characteristics.

R: How do the range finders and rubrics come together?

SL1: Well, that's kind of how the range finder and benchmarks are determined in the first place. If you can't, if you can't articulate why that particular paper is an exemplar of the scoring guide then it probably would…not have been vetted as an example of that paper.

R: In practice, do you rely more on benchmarks or the scoring rubric?

SL1: When I'm articulating the rationale or trying to get a rater to articulate the rationale for assigning a score, we use the language of the scoring guide. If we are trying to use an example of how a writer might have approached a particular prompt, if it is similar to the paper that the person is looking at, then sometimes the benchmarks and the range finders can help the rater. [A benchmark or the scoring rubric] can serve as an illustration of how that approach might also work, so they are both useful but perhaps in different contexts and at different times.

R: Good. Thank you very much.

SL1: That's it?

R: The official Q & A [is] over. We have been talking for 16 minutes.

Appendix E

Debriefing of Subject B

This is a transcript of the researcher (R) debriefing Scoring Leader 2 (SL2).

R: Ok, this is [researcher name] . . . debriefing with [name of Scoring Leader 2]. Thank you for your participation. Since we will continue data collection for a few more weeks, please refrain from discussing this study with your colleagues.

SL2: Okay.

R: The study is concerned with different factors that affect the accuracy of scoring, including the use of shell [language]; some of the essays you score were selected to better understand those factors. The scores that you and your colleagues will provide will be considered as potentially more accurate than scores obtained under normal circumstances. Do you have any questions?

SL2: No.

R: To complete this study, I would like to ask you a few questions and record your answers, if you don't mind.

SL2: No problem.

R: Your comments will be transcribed and not attributed to you personally.

SL2: Ok.

R: Were you able to score these essays without interference from the scoring instructions of the revised GRE?

SL2: Yes.

R: Compared to the OSN scoring of the old GRE, was this process similar, different…how so?

SL2: Very similar—I mean, it was very similar to what we would normally follow in the paper reading ‘cause it was the previous GRE test—so we applied the previous scoring guide. So—no problems at all.

R: Ok.

R: How typical or atypical were the folders you scored compared to the old GRE folders?

SL2: These were fairly typical. There seemed to be—I don't have a hard count but I would say that, just glancing at my score sheets, it seems most of the responses were in the 4, 3, 2 range for both the issue and the argument.

R: What is your understanding of a shell language in general and with respect to issue and argument prompts?

SL2: Well, shell language seems to be one of two types. Either it is a particular phrase or series of sentences that could be memorized by an examinee (and oftentimes these phrases, these sentences occur either near the introduction of a response or at the conclusion of a response) [or] sometimes they are generic statements. For example, in the argument, I think an example of a shell statement might be something like, "This argument could be strengthened if there was more research" and, of course, you could apply that to any argument or topic, so there we go.

R: I do that all the time.

SL2: And the other of type (and it didn't occur very much in these papers) is the similarity of examples, depending, of course, on the topic. Thomas Edison may show up, Abraham Lincoln, or…I think this one had two responses that cited Helen Keller, but again, they were dissimilar enough that they could have been legitimate examples that the examinee identified on his or her own, so that's the type of shell language. And often another indicator is that depending again on the facility of the examinee, the shell language is often error free and may include vocabulary that seems to be beyond what the writer uses—the examinee uses—in the revision of the response.

R: Interesting.

SL2: So, another distinguishing characteristic to show the difference between original text created by the examinee and some type of shell construction that was imported.

R: In other words, in respect to issue and argument, are there differences with respect to shell? Another distinguishing characteristic to show the difference between original text created by the examinee and some type of shell construction that was imported?

SL2: I think [of] the shell language in terms of phrases and sentences, like the example I gave about calling for more research or more evidence. Those type[s] of shell construction seem to occur more in the argument. The similarity of examples seems to occur more in the issue responses.

R: Ok. Did the answers you scored contain shells?

SL2: Yes, some did.

R: Roughly speaking, what proportion of the issue and argument essays you scored do you think contained a significant amount of shell?

SL2: I would think somewhere probably between 10–15%.

R: Can you describe the strategy that you use to score essays that contain shell?

SL2: Well, the guideline that we have developed and worked with as raters and scoring leaders, with the assistance of test development, is to disregard any type of shell construction and evaluate the essay, the responses should say, on the remaining text the examinee has provided.

R: Those are the guidelines they provide but when you score, as a scoring leader, what strategies do you use to, sort of, disregard [the shell construction]?

SL2: Well, I simply read the entire response, obviously, and then if I recognize what I think is shell construction or shell language then, in a second read, I will simply concentrate more on the original text created by the examinee and base the score on that.

R: What role if any does the length of the essay that you are about to score play in your strategy for scoring issue and argument essays?

SL2: Length doesn't necessarily outweigh any other characteristics. Length is often evaluated, or usually evaluated, not often based, on its relationship to the development of the response. And, of course, if it is an issue response, then the length should in some ways be reflective of the examinee's engagement with the prompt, whether agreeing or disagreeing, and the amount of detail and the examples that the examinees has used to illustrate his or her point. In argument, of course, in order to receive an upper-half score in the previous test, the examinees must identify the flawed prompt and then analyze why

the prompt is flawed; the upper-third responses and sometimes a response in the four category will also provide ways to strengthen the argument in order to correct the flaw.

R: Good, well, those are all the questions we have as far as the study—do you have any other questions?

SL2: No, it was interesting, I am glad to be invited to participate—it was a pleasure working with you, [researcher name].

R: Thank you.