



GRE

Listening. Learning. Leading.

Graduate Record Examinations®

**The GRE® Analytical Writing Measure:
An Asset in Admissions Decisions**

August 2007

Copyright © 2007 by Educational Testing Service. All rights reserved.
EDUCATIONAL TESTING SERVICE, ETS, the ETS logos, GRADUATE RECORD EXAMINATIONS, GRE,
TOEFL, and TWE are registered trademarks of Educational Testing Service (ETS) in the United States of America
and other countries throughout the world. ScoreItNow! is a service mark of ETS. TEST OF ENGLISH AS A
FOREIGN LANGUAGE and LISTENING. LEARNING. LEADING. are trademarks of ETS.

The GRE General Test

Since October 1, 2002, the GRE General Test has been composed of Verbal, Quantitative, and Analytical Writing sections. The Verbal and Quantitative sections contain multiple-choice questions that test reasoning skills in the verbal and quantitative domains. The Analytical Writing section is a performance measure; examinees write two essays: a 45-minute “Present Your Perspective on an Issue” task and a 30-minute “Analyze an Argument” task. The “Issue” task states an opinion on an issue of general interest and asks test takers to address the issue from any perspective(s) they wish, as long as they provide relevant reasons and examples to explain and support their views. The “Argument” task presents a different challenge: it requires test takers to critique an argument by discussing how well reasoned they find it. Test takers are asked to consider the logical soundness of the argument rather than to agree or disagree with the position it presents. These two tasks are complementary in that the first requires the writer to construct an argument about an issue, and the second requires a critique of someone else's argument by assessing its claims.

The addition of the Analytical Writing measure to the General Test was made for several reasons:

- to respond to an expressed need from faculty for help in assessing higher level critical thinking and analytical writing skills of applicants to graduate programs, and
- to provide a performance assessment that measures a test taker's ability to make and critique arguments, skills that are central to the work done by graduate students in most fields.

This document provides detailed information about the nature of the Analytical Writing measure and the value that it adds above and beyond the Verbal and Quantitative measures. We will consider six topics:

- A. Value Added by the Analytical Writing Measure**
- B. Validity Evidence**
- C. Fairness for Examinee Subgroups**
- D. Comparability and Reliability**
- E. Test Scores**
- F. Test Preparation**

A. Value Added by the Analytical Writing Measure

For virtually all disciplines, the Writing measure adds value to the General Test because it provides unique information about test taker abilities, over and above skills measured in the Verbal and Quantitative measures.

Before the Writing measure became operational, a research study was conducted in April 2001 in which subjects reported their undergraduate grades and took the Verbal (V),

Quantitative (Q), and Analytical Writing (AW) measures, along with the old Analytical Reasoning (A) measure. With these data, various combinations of measures were investigated to see which was best at predicting undergraduate grade point average (UGPA). The results are shown in the Table 1 below, which lists the R^2 values for the predictor variables; this value represents the percent of the variance explained by the various combinations of tests.

Table 1

Group	N	Predictions of UGPA ¹		
		V, Q	V, Q, A	V, Q, AW
Total Group	5,733	.084	.091	.100
<i>Language Group:</i>				
US Citizens	4,038	.090	.098	.103
Non-US Citizens	1,616	.089	.096	.103
<i>Broad Undergraduate Major Field:</i>				
Business	141	.224	.218	.216
Education	185	.054	.053	.098
Engineering	794	.111	.109	.141
Humanities & Arts	793	.107	.107	.110
Natural Sciences	2,078	.081	.093	.094
Social Sciences	1,396	.076	.081	.090
Other	346	.039	.051	.086

These data indicate that when Analytical Writing is added to Verbal and Quantitative, the contribution to predicting UGPA is higher than that of the old Analytical Reasoning measure.² For the total group, there is a .016 increase over Verbal and Quantitative when Analytical Writing is added, and a .007 increase over Verbal and Quantitative when Analytical Reasoning is added. These results are found both for US citizens and non-US citizens. For the undergraduate major fields, the results for Business show a slight decrease in the contribution to predicting UGPA for Analytical Writing compared to Analytical Reasoning (.218 versus .216), but for all other fields Analytical Writing increases the degree to which undergraduate grades are predicted. This is especially true in Education and Engineering.

Of course, the purpose of the GRE General Test is not to predict undergraduate grades. It will require several years of operational use before we are able to determine the relationship between Analytical Writing scores and performance in graduate school. But the importance of making and critiquing arguments in graduate school, and the relative

¹ These R^2 values have been adjusted for the number of predictor variables.

² It should be noted that the Analytical Writing measure might overlap with UGPA to a greater extent than the Analytical Reasoning measure which would lessen the incremental validity of the measure for predicting graduate grades.

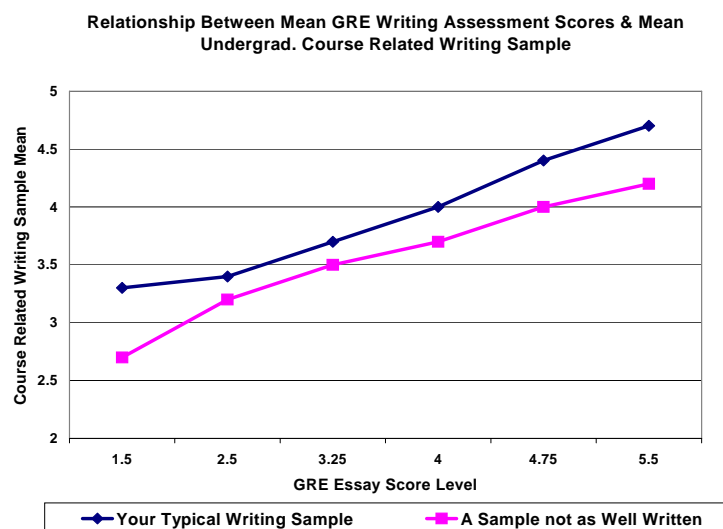
independence of Analytical Writing performance from performance on the Verbal and Quantitative measures, suggest that Analytical Writing will be an important part of overall GRE predictive validity. In coming years we will be conducting validity studies to confirm that this is the case.

B. Validity Evidence

The validity foundations established for the Analytical Writing measure are quite extensive. (A full listing of the studies is included in Appendix A.) The validity of the measure can be demonstrated by: 1) linking the content of the test to necessary skills identified by potential users in the graduate community, 2) establishing the construct that is intended to be measured, and 3) conducting special validity studies.

- 1) Interviews with graduate faculty have consistently identified critical thinking and writing skills as important for success in graduate school. The two tasks that comprise the Analytical Writing measure are both considered essential in many fields of graduate study. These two tasks are complementary in that the first requires the writer to construct his/her own argument about an issue, and the second requires a critique of someone else's argument by assessing its claims. The structure of the test can thus be shown to provide content-related evidence of validity because the test assesses skills identified by the graduate community as essential for success in many fields of graduate work.
- 2) One way to demonstrate the construct validity of the Analytical Writing measure is to produce evidence of both convergent and discriminant validity of the measure. Convergent validity focuses on the correlation of the measure with other measures that it is intended to resemble; discriminant validity examines the extent to which the measure does not correlate with other measures that should be very different from it (Campbell & Fiske, 1959). Because of the extensive research done in the development phase, it is known that the Analytical Writing measure correlates with other measures of academic writing produced by examinees—a result that should be expected (Power, Fowles, & Welsh, 1999) and that is shown in Figure 1 below.

Figure 1



Evidence of discriminant validity comes from several sources. Although students submit a personal statement in essay form with many of their applications, the Analytical Writing measure is more closely correlated with other indicators of writing skill than the personal statement in almost all cases (Powers and Fowles, 1997c). Furthermore, data collected in an April 2001 study indicate that the Analytical Writing measure has a very low correlation (.02) with the Quantitative measure and a moderate correlation (.55) with the Verbal measure—a finding that is consistent with the structure and intent of those various measures.³

- 3) Three large validity studies were conducted during the development of the Analytical Writing measure that contain evidence of its psychometric quality (Powers, Fowles, & Boyles, 1996; Powers, Fowles, & Welsh, 1999; Schaeffer, Briel, & Fowles, 2001). Other studies have focused on particular aspects of validity. For example, Powers & Fowles (1997a) showed that graduate deans and faculty evaluate the same features as GRE essay readers when they judge essay quality. In another example, examinees were given a greater amount of time to complete the essays. Results showed that while the essay scores improved, the rank ordering of the test takers remained nearly the same (Powers & Fowles, 1996). This demonstrates that the time limits on the Analytical Writing essays do not distort our assessment of critical thinking abilities.

All research studies for the Analytical Writing measure are available on the GRE website (www.ets.org/gre/research.html).

C. Fairness for Examinee Subgroups

The GRE Board has long been concerned that examinee groups not be disadvantaged by changes in the General Test. In response to that concern, extensive analyses of group differences in Analytical Writing were performed before it became operational as a standalone test in October 1999. These analyses have since been supplemented by data from those who have taken the standalone test operationally,⁴ and by data from a special research study conducted in April 2001. The findings from each of these data sources indicate that there is less difference in the scores of men and women on the Analytical Writing measure than on the multiple-choice measures. The differences between African American and White examinees and between most groups of Hispanic examinees and White examinees are also smaller on Analytical Writing than on the multiple-choice measures. The difference between Asian American and White examinees is about the same as the difference on the Verbal measure.⁵ Group performance results from the April 2001 study are shown in Table 2 below.

³ The correlation between the Verbal and Quantitative measures is .23.

⁴ It should be kept in mind that many of the examinees who have taken the standalone Writing measure are a self-selected group who might be presumed to have better writing skills than the general GRE population.

⁵ Asian American examinees outscore White examinees on the quantitative measure.

Table 2

Examinee Group ⁶	N	Standardized Difference ⁷				
		Operational General Test			April Analytical	April Writing
		V	Q	A		
<i>Gender Groups:</i>						
Males	1,192	+.29	+.67	+.24	+.26	-.01
Females	2,489		-			
<i>Ethnic Groups:</i>						
American Indian, Alaska Native	20	+.47	+.33	+.29	+.45	+.67
Black or African American	270	+.92	+1.13	+1.19	+1.17	+.79
Mexican or Chicano	72	+.52	+.53	+.87	+.71	+.48
Asian Amer. or Pacific Islander	210	+.16	-.34	+.21	+.42	+.26
Puerto Rican	38	+1.09	+.58	+.82	+.83	+1.04
Other Hispanic, Latin American	69	+.52	+.39	+.76	+.68	+.50
Other	137	-.24	+.04	+.06	+.10	-.05
White	2,874					
<i>Language Groups:</i> ⁸						
US Citizen	3,699	+.61	-.86	+.03	+.21	+.88
Non-US Citizen	1,493					
<i>Broad Undergraduate Major Field:</i>						
Business	119	+.20	+.12	+.28	+.37	+.27
Education	162	+.24	+.64	+.42	+.38	+.01
Engineering	641	+.48	-1.12	-.28	-.13	+.69
Humanities & Arts	711	-.67	+.42	+.02	-.07	-.55
Natural Sciences	1,854	+.12	-.32	-.14	-.14	+.14
Social Sciences	1,259	-.08	+.37	+.06	+.06	-.26
Other	313	+.13	+.59	+.34	+.33	-.07

⁶ Examinees in this table participated in the April 2001 research study and also had an operational test score. For the gender and minority analyses, results are based only on US citizens.

⁷ The standardized difference is the difference between the mean scores for two groups divided by the pooled standard deviation. In the gender analysis, the female mean score is subtracted from the male mean score (meaning that a negative value shows women scoring higher). In the ethnic group analyses, the minority means are subtracted from the White mean. For foreign examinees, the non-US citizen mean is subtracted from the US citizen mean. In the major field analyses, the mean of the target group is subtracted from the mean for everyone else.

⁸ Analyses are based on US citizenship because this variable has proved a better indicator of foreign students than the English language question. (For example, Asian foreign students will often answer English is their best language when it appears that it is not.)

This table shows that mean scores among women are about the same as those of men. For African Americans, there is a decrease in the score difference on Analytical Writing compared to the former Analytical measure (+.79 versus +1.19). A decrease is also obvious for Mexican and Other Hispanic groups. Substituting Analytical Writing scores for those of the Analytical measure might lead to the admission of more women and more of these minority students into graduate school, depending on how graduate departments use the writing information.

For Asian American examinees, there is a slightly increased difference on Analytical Writing, compared to the operational Analytical Reasoning measure: .26 versus .21. However, this difference is not substantively significant.⁹

These results also indicate that ESL test takers, as indicated by the non-US citizen category, find the Analytical Writing measure more challenging, on average, than do native speakers of English. Steps have already been taken to ensure that these performance differences are not due to differences in the cross-cultural accessibility of the prompts. Special fairness reviews occur for all prompts to ensure that the content and tasks are clear and accessible for all groups of test takers, including international students. Before grading essays, scorers are trained to accept a wide variety of reasoning styles and strategies for structuring the essays because there is not a single right way to address the essay topics. In addition, scorers are trained to focus on the analytical logic of the essays more than on spelling, grammar, or syntax. The mechanics of writing are weighed in their ratings only to the extent that these impede clarity of meaning. Since the Analytical Writing measure is tapping into different skills than the multiple-choice measures, it may not be surprising that the performance of ESL examinees differs on this measure. Given that graduate faculty have indicated that analytical writing is an important component of work in most graduate fields, they may find the addition of an Analytical Writing measure to the General Test will provide a more valid indicator of applicants' ability.

It is clear that the net effect of substituting the Analytical Writing measure for the Analytical Reasoning measure is to emphasize skills on which ESL examinees have generally done less well. Such differences in performance might be more troubling were it not related to the construct that is intended to be measured by Analytical Writing. Given that the Analytical Writing measure taps into different skills than the Analytical Reasoning measure, it should not be surprising that it results in a different rank ordering of test takers. Of course, English-language writing skills do count, and so it may be important for graduate departments to place a greater reliance on Test of English as a Foreign Language™ (TOEFL®) scores for ESL examinees to help determine whether a low score on the Analytical Writing measure is due to lack of familiarity with English or to a lack of ability to produce and analyze logical arguments.

Research on Analytical Writing also shows that Humanities and Arts majors (fields that are often writing intensive), Social Science majors, and Education majors all fare better on Analytical Writing than they do on Analytical Reasoning measure. While these cross-field

⁹For this group, the difference in performance on the operational Analytical Reasoning test (.21) compared with the April Analytical Reasoning test (.42) is much more noticeable.

results may be interesting, they are not critical for most faculty. Since admissions decisions are made among applicants within a discipline, the expected gain in predictive validity is more important than any differences in mean scores across fields. For example, although Engineering examinees do less well on Analytical Writing than on the Analytical Reasoning measure compared with other graduate students, the use of Analytical Writing for Engineering students is likely to have greater predictive validity than the use of the Analytical Reasoning measure (see Table 1).

D. Comparability and Reliability

Comparability of Writing Prompts

The major concern in prompt comparability is whether all writing prompts are equally difficult. The issue of prompt comparability takes on special significance if it is also associated with group membership, that is, if a prompt is differentially more difficult for members of some groups. Because of this concern, researchers investigated the comparability of prompts and subgroup performance in the Psychometric Evaluation of the New GRE Writing Assessment (Schaeffer, Briel & Fowles, 2001), a project originally begun during the test development phase in 1996. This investigation, based on selected subgroups of English-best-language examinees who were US citizens, concluded that most of the writing prompts were comparable in difficulty and that no important subgroup interactions with prompt classifications were observed.

Efforts to ensure comparability for examinees include pretesting all prompts in the GRE Analytical Writing measure before they are used operationally in order to obtain score distribution and difficulty information. In the current test (consisting of an Issue task and an Argument task), steps are also taken to ensure that the overall test difficulty is similar across examinees. In addition, prompt difficulty is monitored in an on-going basis to determine whether there are changes in performance.

Inter-Rater Reliability

Each essay response receives two independent ratings from trained raters, using a 6-point holistic score scale.¹⁰ The scores on each essay task are averaged and the final essay score is the average of the two task scores. The final essay scores range from 0 to 6 in half-point increments.

The inter-rater reliability, also referred to as scoring reliability, is an important index measuring the consistency of writing scores. The inter-rater reliability indicates the extent to which individual test takers would receive the same essay scores that they would get if their essay responses were scored by all possible essay raters. Using the statistical procedure recommended by Livingston (Livingston, 2004) and the operational writing score data from October 2002 to September 2003,¹¹ results show that the inter-rater reliability of the Analytical Writing measure is 0.93 (standard error of scoring = 0.24).

¹⁰ If the essay score assigned by two readers differs by more than one point, the essay is referred to a third experienced reader for adjudication.

¹¹ Based on examinees who indicated that English was their best language and that they were US citizens.

Training of Readers for Analytical Writing

The reliability of the operational test scores is dependent on the pool of essay readers. Many different strategies are used to ensure that all readers use the same scoring standard. In order to qualify as a GRE reader, a college or university faculty member who has been trained on the scoring criteria must pass a certification test for each task type. The certification test contains a total of 50 essays (10 essays on each of 5 different topics). In order to qualify as a GRE scoring leader, who monitors readers and serves to resolve discrepant scores, a GRE reader must demonstrate a high degree of scoring accuracy and an ability to mentor readers.

Once a reader has been certified, monitoring of reader quality continues. At the beginning of each scoring session, or when switching from one task type to the other, the readers must score a calibration set of 10 pre-scored essays with 90% accuracy.¹² To familiarize themselves with each new topic, readers review topic notes, read prescored benchmark essays and commentary, and then practice scoring rangefinder essays before beginning operational scoring on that topic. Prescored but unidentified papers are included in most reader assignments to monitor continued reader accuracy. Scoring leaders also monitor readers' performance throughout the scoring session by reading already-scored essays, reviewing scores on monitor and calibration papers, reviewing score distributions, and monitoring readers' activities. Readers who deviate from the acceptable level of accuracy are retrained or dismissed. In the current operational test, 97 percent of scores are within one point of agreement with each other, as indicated in Table 3. Appendix B provides further information about the recruitment and training of essay readers.

Table 3

Reader Agreement on Scores	Issue	Argument
Exact	57%	60%
Within one point	97%	97%

Although these results for the total group of examinees are impressive, 17% of the time readers will agree exactly by chance and 44% of the time they will agree within one point by chance. In order to correct for chance-expected agreement, a Cohen's kappa statistic was calculated for exact scores on the operational test, and these values are .40 for the Issue task and .46 for the Argument task

¹² Exact and adjacent scores are considered accurate. A discrepancy occurs when the assigned score differs by more than 1-point from the pre-assigned score. A reader must have no more than one discrepant score and at least five exact scores on a calibration set in order to proceed to operational scoring.

Reliability of Analytical Writing Total Score

The reliability of Analytical Writing total score is 0.72.¹³ This reliability coefficient was computed using the data from a study in which some examinees responded to two Issue prompts and some responded to two Argument prompts (Schaeffer, Briel, & Fowles, 2001). The reliability was computed from the covariances of scores on prompts of the same type. This reliability is in the expected range of reliability for a two-essay test.

E. Test Scores

Score Interpretation

For the Analytical Writing section, each essay receives a score from two trained readers, using a 6-point holistic scale. In holistic scoring, readers are trained to assign scores on the basis of the overall quality of an essay in response to the assigned task. If the two assigned scores differ by more than one point on the 6-point scale, the discrepancy is adjudicated by a third, very experienced reader. Otherwise, the scores from the two readings of each essay are averaged. The scores on the two tasks are then averaged and a single score (rounded up to ½-point intervals) is reported for the Analytical Writing measure.

The primary emphasis in scoring the Analytical Writing measure is on examinees' critical thinking and analytical writing skills rather than on grammar and mechanics. Interpreting the score is, thus, tied to the score level descriptions (available in the *GRE Guide to the Use of Scores* and on the GRE website at www.ets.org/gre/edupubs.html). The skills that are evaluated include the ability to:

- articulate complex ideas clearly and effectively
- examine claims and accompanying evidence
- support ideas with relevant reasons and examples
- sustain a well-focused, coherent discussion
- control the elements of standard written English (a factor that plays a role only to the extent that poor writing skills impede readers' understanding of the argument).

Score Interpretation for ESL Examinees

The issue of score interpretation is somewhat more complicated for ESL students. As a performance assessment, the GRE Analytical Writing measure provides a snapshot of examinees' analytical writing ability before entry into graduate school. If ESL examinees do not understand the task being posed to them, their performance will be affected. Test users should consider a variety of pieces of information about ESL applicants, including TOEFL and TWE (Test of Written English) scores, to determine whether to admit these students.

¹³ This estimate takes into account both the sampling of prompts and the sampling of scorers as possible sources of unreliability.

Looking at both the TOEFL and GRE writing measures can be helpful because these two assessments are very different. Unlike Analytical Writing, the TWE® that is part of TOEFL is not designed to measure higher levels of thinking and analytical writing, but centers instead on command of English vocabulary, grammar, spelling, and syntax. Because the TOEFL test emphasizes fundamental writing skills, the TOEFL TWE score can supplement an Analytical Writing score by helping faculty determine whether a low score on the GRE Analytical Writing measure is due to lack of familiarity with English or to lack of ability to produce and analyze logical arguments.

Analytical Writing Score Distribution Information

A full score distribution with percentiles is contained in the *GRE Guide to the Use of Scores* and on the GRE website at www.ets.org/gre/edupubs.html.

Score Reporting Schedule

The General Test is given year-round on the computer. For the Analytical Writing section of the test, essay tasks are delivered on the computer, and test takers type their responses. Unlike the unofficial Verbal and Quantitative scores that examinees can view at the test center, Analytical Writing scores are not available on the testing date.

Verbal, Quantitative, and Analytical Writing scores will be sent to institutions and test takers within 15 days of the test administration. Score reporting timeframes are published to allow test takers to plan accordingly.

Use of Scores

The GRE program has prepared a guide to help departments determine how they can integrate Analytical Writing information into their admissions decisions. This guide, *How to Interpret and Use GRE Analytical Writing Scores*, can be downloaded from the GRE website at www.ets.org/gre/edupubs.html. It is important to note that scores from the Analytical Writing measure should not be combined with the Verbal and Quantitative scores. Each measure should be considered separately because each provides insight into a different aspect of the applicant's abilities.

F. Test Preparation

Individuals may view information about the Analytical Writing measure on the GRE website at www.ets.org/gre/edupubs.html. The materials available include information about the nature of the test, directions for the two essay tasks, the entire pool of topics, scoring criteria, samples of scored essays, and the word processing tutorial. Individuals who register for the GRE General Test are automatically sent a CD-ROM containing *GRE POWERPREP Software—Test Preparation for the GRE General Test*. This test preparation software includes advice on how to write effective essays for the Issue and Argument tasks. It also lets users practice writing essays under simulated GRE testing conditions, with the

same GRE word processing and testing tools that appear on the test. A downloadable version of *POWERPREP* software is also available for free to anyone who visits the GRE website at www.ets.org/gre/greprep.html.

Individuals can also access *ScoreItNow!*SM *Online Writing Practice* for additional preparation for the Analytical Writing measure. Users submit essays and receive immediate scores on their responses to GRE Analytical Writing topics as well as brief diagnostic feedback on grammar, usage, mechanics, style, organization, and development.

Appendix A

GRE Research on Writing and the Writing Assessment

Breland, H. (1999) Exploration of an automated editing task as a GRE writing measure. GRE Research Report 96-01. Princeton, NJ: ETS.

Breland, H., Bridgeman, B. & Fowles, M.E. (1999) Writing assessment in admission to higher education: Review and framework. GRE Research Report 96-12. Princeton, NJ: ETS.

Cooper, P.L. (1984) The assessment of writing ability: A review of research. GRE Research Report 82-15. Princeton, NJ: ETS.

Kaplan, R.M., Wolff, S.E., Burstein, J.C., Lu, C., Rock, D.A. & Kaplan, B.A. (1998) Scoring essays automatically using surface features. GRE Research Report 94-21. Princeton, NJ: ETS.

Livingston, S.A. (in press) An interesting problem in the estimation of scoring reliability. Journal of Educational and Behavioral Statistics.

O'Neill, K.A., & Rizavi, S. (2002) Performance of examinee groups on a measure of analytical writing. Presentation at NCME national conference, New Orleans.

Powers, D.E., Burstein, J.C., Chodorow, M.S., Fowles, M.E. & Kukich, K. (2000) Comparing the validity of automated and human essay scoring. (GRE Research Report 98-08a. Princeton, NJ: ETS.

Powers, D.E., Burstein, J.C., Chodorow, M.S., Fowles, M.E. & Kukich, K (2001) Stumping E-Rater™: challenging the validity of automated essay scoring. GRE Research Report 98-08b. Princeton, NJ: ETS.

Powers, D.E. & Fowles, M.E. (1996) Effects of applying different time limits to a proposed GRE writing test. JEM, 33, 433-452. (Also available as GRE Research Report 93-26c. Princeton, NJ: ETS.)

Powers, D.E. & Fowles, M.E. (1997a) Correlates of satisfaction with graduate school applicants' performance on the GRE writing measure. GRE Research Report 93-18. Princeton, NJ: ETS.

Powers, D.E. & Fowles, M.E. (1997b) Effects of disclosing essay topics for a new GRE writing test. GRE Research Report 93-26a. Princeton, NJ: ETS.

Powers, D.E., & Fowles, M.E. (1997c). The personal statement as an indicator of writing skill: A cautionary note. Educational Assessment, 4(1), 75-87. (Also available as GRE Research Report 93-26d. Princeton, NJ: ETS.)

Powers, D.E. & Fowles, M.E. (1998a) Effects of preexamination disclosure of essay topics. AME, 11, 139-157

Powers, D.E. & Fowles, M.E. (1998b) Test takers' judgments about GRE writing test prompts. GRE Research Report 94-13. Princeton, NJ: ETS.

Powers, D.E. & Fowles, M.E. (2000) Likely impact of the GRE writing assessment on graduate admissions decisions. GRE Research Report 97-06. Princeton, NJ: ETS.

Powers, D.E., Fowles, M.E. & Boyles, K. (1996) Validating a writing test for graduate admissions. GRE Research Report 93-26b. Princeton, NJ: ETS.

Powers, D.E., Fowles, M.E. & Welsh, C.K. (1999) Further validation of a writing assessment for graduate admissions. GRE Research Report 96-13. Princeton, NJ: ETS.

Powers, D.E., Fowles, M.E. & Willard A (1994). Direct assessment, direct validation? An example from the assessment of writing. Educational Assessment, 2, 89-100.

Schaeffer, G.A., Briel, J.B. & Fowles, M.E. (2001) Psychometric Evaluation of the New GRE Writing Assessment. GRE Research Report 96-11. Princeton, NJ: ETS.

Appendix B

Recruitment and Training of Essay Readers

How are readers recruited?

Readers are recruited from a variety of academic disciplines throughout the United States. The criteria for readers are (a) reside in the continental U.S. or Hawaii; (b) be a U.S. citizen, a resident alien, or authorized to work for remuneration in the U.S.; (c) hold, at minimum, a completed master's degree or equivalent academic credential(s); and (d) currently teach or have recently taught a college- or university-level course in any field or discipline where writing and/or critical thinking skills are important.

Anyone who meets these criteria and is interested in becoming a reader for the GRE Analytical Writing measure is advised to contact the ETS Online Scoring Network office at GREScore@ets.org or go to the ETS website (www.ets.org).

Where are the readers located?

Readers are located throughout the United States. Most scoring of the Analytical Writing measure is conducted through the ETS Online Scoring Network (OSN). OSN is an Internet-based scoring system through which readers login score essays online, either from home via the web or at an OSN center.

How are readers trained?

To qualify as official scorers, readers must take and pass a certification test demonstrating that they can apply the scoring standards to the same consistent standard as other scorers. At the beginning of each scoring session, or when changing to a different task type, trained readers must score a calibration set of 10 prescored essays with 90 percent accuracy. To familiarize themselves with each new topic, readers review topic notes, read prescored benchmark essays and commentary, and then practice scoring rangefinder essays before beginning operational scoring.