Listening. Learning. Leading.®

Research Report
ETS RR-13-18

The Impact of Sampling Approach on Population Invariance in Automated Scoring of Essays

Mo Zhang

October 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.

The Impact of Sampling Approach on Population Invariance in Automated Scoring of Essays

Mo Zhang
Educational Testing Service, Princeton, New Jersey

October 2013

Action Editor: Shelby Haberman

Reviewers: Jiahe Qian and Yue Jia

Abstract

Many testing programs use automated scoring to grade essays. One issue in automated essay scoring that has not been examined adequately is population invariance and its causes. The primary purpose of this study was to investigate the impact of sampling in model calibration on population invariance of automated scores. This study analyzed scores produced by the *e-rater*® scoring engine using a *GRE*® assessment data set. Results suggested that equal allocation stratification by language sampling approach performed optimally in maximizing population invariance using either human/e-rater agreement or their correlation pattern differences with external variables as evaluation criteria. Guidelines were given to assist practitioners in choosing a sampling design for model calibration. Potential causes for lack of population invariance, study limitations, and future research are discussed.

Key words: automated essay scoring, sampling, population invariance in automated essay scoring

Acknowledgments

Thanks go to my colleagues at Educational Testing Service—Shelby Haberman, David Williamson, Isaac Bejar, Michael Kane, and Randy Bennett—and to my advisors at Washington State University—Professors Brian French, Michael Trevisan, and Jan Dasgupta—for their valuable advice on the study. I also thank the reviewers and the editors for their review and editorial support.

Table of Contents

Page

Method 3
Instrument 3
Participants 4
Procedure 6
Results 11
Population Invariance for Human/E-rater Agreement 11
Population Invariance for Human/E-rater Correlation Pattern With External Variables 15
Discussion 22
Practical Implications 26
Limitations 26
Recommendations for Additional Research 27
References 29
Notes 32

List of Tables

Page

Table 1 Demographics and Characteristics of the Test-Takers in Generic and Prompt-Specific Scoring Group 5
Table 2 Stratification Variables Used in Sampling Design 7
Table 3 Listing of Sampling Approaches and Their Abbreviations 7
Table 4 Sampling Comparison Based on Population Invariance for Human/E-rater Agreement by Group Classification 12
Table 5 Sampling Comparison Based on Population Invariance for Human/E-rater Agreement by Agreement Index 13
Table 6 Weighted Mean and Standard Deviation of Human/E-rater Agreement Indices 14
Table 7 Sampling Comparison Based on Population Invariance for Correlation Difference by Group Classification 16
Table 8 Sampling Comparison Based on Population Invariance for Correlation Difference by External Variable 17
Table 9 Weighted Mean and Standard Deviation of the Correlation Differences of Human and E-rater With External Variables 19
Table 10 Aggregated Results of Sampling Comparison Based on Population Invariance 21
Table 11 Human Scores, Feature Weights, and Outliers for Selected Countries/Territories 24

Several consequential testing programs incorporate an essay component, including the *TOEFL iBT*® test, the GMAT assessment, the *GRE*® revised General Test, and the Pearson Test of English. For efficiency as well as for measurement reasons, many programs use automated scoring systems. Among the better known systems are the *e-rater*® scoring engine developed at Educational Testing Service (2010); Intelligent Essay Assessor developed by Knowledge Analysis Technologies, now Pearson Knowledge Technologies (Pearson Education Inc., 2010); and Project Essay Grade initially developed by Page (1966). Many characteristics of the scores produced by these systems have been empirically examined, including agreement with human ratings (e.g., Lee, Gentile, & Kantor, 2008), agreement with external variables (e.g., Wang & Brown, 2008), factor structure (e.g., Attali, 2007), and reliability (e.g., Attali & Powers, 2009). One characteristic that has been less frequently examined is population invariance and its causes.

The notion of population invariance is central to test fairness (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education [AERA, APA, & NCME], 1999; Kane, 2013; Messick, 1998). The concept indicates that a test (or items in a test) should function similarly (i.e., scores should have the same meaning) from one population group to the next. The literature on population invariance is voluminous, covering various aspects of test (or item) functioning across groups categorized by race/ethnicity, age, gender, language, and disability status (e.g., Dorans & Holland, 2000; Dorans, Schmitt, & Bleistein, 1992; French & Mantzicopoulos, 2007; Mattern, Patterson, Shaw, Kobrin, & Barbuti, 2008; Ruth, 2000).

In the automated essay scoring context, tests are usually composed of only a few essay prompts, or, alternatively, tests may combine one or two prompts with a considerable number of multiple-choice items. In the latter case, even though the writing scores are usually reported along with scores on other sections, those writing scores are often reported separately as the only indictor of a test-taker's writing proficiency. When the total writing score is composed of only two prompts, as is the case for the GRE Analytical Writing (GRE AW) section or TOEFL iBT writing assessment, each individual prompt can have significant effects on the total writing score that is reported. As such, lack of invariance for an item can have a notable impact on total test score; thus, invariance has been studied at the item level.[1]

The few studies that have been conducted on this issue have yielded consistent findings; that is, a lack of population invariance occurred and was more severe for some population groups

than for others. Such lack of invariance has manifested itself in several ways. For example, Bridgeman, Trapani, and Attali (2009) found that the correlation between e-rater scores and human ratings for the GRE AW prompts was significantly lower for Japanese, Taiwanese, Nigerian, and particularly Chinese population groups than for other population groups. In a second paper, Bridgeman, Trapani, and Attali (2012) reported on the direction of the discrepancy between automated and human scores, specifically, that the average e-rater scores were higher than human ratings for Korean and Chinese population groups in both GRE writing tasks, while the reverse scenario arose for the Canadian, Turkish, English, and French population groups. Similar results were reported by Ramineni, Trapani, Williamson, Davey, and Bridgeman (2012a), who found that the average e-rater score on GRE AW issue prompts was lower than the average human rating by 0.18 standard deviations for African American test-takers, but was higher by 0.06 standard deviations for Chinese test-takers.

Similar phenomena have been found in TOEFL iBT. In an investigation of TOEFL iBT independent prompts, the e-rater scores on average were higher than human ratings for Chinese, Korean, and Japanese population groups, but lower than human ratings for Arabic, Turkish, and Spanish groups (Bridgeman et al., 2009). Similarly, Ramineni, Trapani, Williamson, Davey, and Bridgeman (2012b) reported that for TOEFL iBT integrated prompts, the e-rater mean score was higher than the human mean score by 0.21 standard deviations for Chinese-speaking examinees, but was lower than the human mean score by 0.19 standard deviations for Arabic-speaking examinees. Population groups with Hindi and Spanish language backgrounds also showed signs of lack of invariance relative to other groups (Ramineni et al., 2012b).

While these studies have suggested the presence of a lack of invariance in automated scores, specific causes are unknown. Many automated essay grading systems produce scores using regression (Drasgow, Luecht, & Bennett, 2006; Haberman, 2007). That is, these systems use a model that predicts, from computable features of essays, the scores human judges would be likely to assign. One characteristic of regression-based prediction is that the algorithms are conditioned on the calibration sample (i.e., the sample of examinees whose essays are used to train the automated scoring system). For example, Zhang (2012) found that the choice of sampling method directly affected the regression estimates and the quality of resulting automated scores in terms of the scores' agreement with human ratings on the same essay and with ratings on another writing task.

The sampling method that is used to calibrate the automated scoring model may also affect population invariance. Depending on the sampling method, the scoring model may not behave consistently for test-takers with different group membership. An inadequate calibration sample (e.g., with some groups over- or underrepresented) could, therefore, potentially lead to a scoring model that yields biased automated scores for certain populations. The purpose of this study was to investigate the impact of different sampling methods on the population invariance of automated scores, in particular, whether stratification-based approaches alleviate such departures.

Method

Instrument

The data were collected from the GRE General Test's Analytical Writing section.[2] In the GRE AW section, test-takers complete two types of writing task. One is called the *issue* task, in which the test-takers state their perspectives on a subject of general interest. The other task is *argument*, in which the test-takers critique and analyze claims provided in a prompt. Although the underlying construct assessed by the two tasks is the same (i.e., writing proficiency), the measurement focus of the issue task and argument task is slightly different. The issue task tends to assess how well the test-takers provide relevant reasons and examples to support their statements, and the argument task tends to assess the soundness of the logic and reasoning presented by the test-takers.

This study analyzed 20 of 139 issue prompts administered between July 15, 2010, and June 30, 2011. These prompts were selected because they represented the ones taken by the largest number of examinees.

All essays written in response to the 20 prompts were processed by e-rater.[3] The version of e-rater used extracts a large number of micro-linguistic features that are further combined into a smaller, fixed set of 11 primary features (Enright & Quinlan, 2010): grammar, mechanics, style, usage, collocation-preposition, organization, development, word choice, word length, and two content features. E-rater employs a multiple linear regression approach for predicting human ratings from the features.

Two types of e-rater scoring models were used: generic (G) and prompt-specific (PS). G models are built on a group of prompts within the same genre. As a result, all prompts within a group have an identical scoring algorithm. PS models are built on individual prompts. All the

above 11 features are used in prompt-specific scoring, while only nine features, excluding the two content features, are used in generic scoring.

Each essay was also graded by at least one randomly assigned human rater.[4] Qualified human raters were selected and trained to use the scoring rubric and grade GRE writing responses. The quality of the human ratings was examined by using two data sets, one historical and the other current.[5] In the current data set, when there existed a second human rater (N = 106,637), the correlation coefficient between the first and second human was 0.70, and the human ratings on the issue prompt modestly correlated with the human ratings on the other writing task (r = 0.60). The historical interrater agreement was on average 0.74 across prompts, indicating reasonable levels of agreement. These values are comparable to the values found for other postsecondary examinations that incorporate an essay writing component (e.g., interhuman correlation of 0.69 for TOEFL® independent writing tasks, Ramineni et al., 2012b; correlation coefficient of 0.59 between the two SAT® II writing tasks and 0.56 between the two persuasive writing tasks in the new SAT, Breland, Kubota, Nikerson, Trapani, & Walker, 2004).

Participants

Participants were GRE General Test examinees responding to the 20 prompts selected for analysis in this study. The country/territory in which the examinees took the test was automatically recorded. Test-takers' backgrounds, including their English-as-best-language status and academic major, were self-reported. Additionally, test-takers' scores on the GRE General Test's Verbal (GRE-V) section, Quantitative (GRE-Q) section, and on the GRE AW argument prompt (i.e., the other writing item in a test form) were retained.

For analysis purposes, the examinee sample was divided as follows:

- Generic group 1 (G1): examinees responding to 10 randomly selected prompts from the 20-prompt total
- Generic group 2 (G2): examinees responding to the remaining 10 prompts
- Prompt-Specific 1 (PS1): examinees responding to the prompt with the highest volume from the 20-prompt total
- Prompt-Specific 2 (PS2): examinees responding to the prompt with the second highest volume from the 20-prompt total

The sizes of the two generic groups were 120,705 and 138,155. The two generic groups were comparable in terms of the composition of the test-taker population. Each generic group had examinees from 37 individual countries/territories and one combined.[6] The largest examinee population group was the United States, which accounted for nearly 70% of the population, followed by Indian and (mainland) Chinese, each of which accounted for roughly 10% of the population. The majority of the test-takers identified English as their best language (which was not surprising, given that most test-takers took the test in North America). Countries/territories were also grouped geographically into eight regions (described below) for stratification purposes.

In contrast to generic groups, the sizes of the two prompt-specific populations were 22,017 and 19,733. The number of countries/territories for prompts was smaller than that of groups due to the smaller individual prompt population size. There were still at least 18 countries that were large enough to stand alone as independent strata. Small countries/territories (fewer than than 200 examinees) had to be combined to form a separate stratum.

Table 1 gives the sample sizes and test-taker characteristics associated with each of the generic and prompt-specific groups.

Table 1

Demographics and Characteristics of the Test-Takers in Generic and Prompt-Specific Scoring Group

Popula-tion	*N*	Demographics		Ability (average scores)		
		Largest country/territory	English-as-best-language	GRE-Verbal	GRE-Quantitative	GRE AW argument
Generic 1	120,705	US: 68.6% India: 11.0% China: 10.0%	Yes: 71.6% No: 14.7%	461.5 (SD = 122.2)	603.8 (SD = 145.4)	3.5 (SD = 1.0)
Generic 2	138,155	US: 68.0% India: 10.8% China: 10.9%	Yes: 71.3% No: 14.8%	460.8 (SD = 122.6)	604.6 (SD = 146.2)	3.5 (SD = 1.0)
Prompt-Specific 1	22,017	US: 68.2% India: 11.0% China: 11.0%	Yes: 70.9% No: 14.7%	445.9 (SD = 120.8)	587.5 (SD = 149.8)	3.3 (SD = 1.0)
Prompt-Specific 2	19,733	US: 68.2% India: 10.3% China: 9.7%	Yes: 72.8% No: 14.7%	466.9 (SD = 121.3)	608.9 (SD = 143.4)	3.5 (SD = 1.0)

Note. GRE AW = GRE Analytic Writing.

Procedure

Sampling approaches. Three general classes of sampling approaches were used. Those three general classes were (a) simple random sampling without replacement (SRS), (b) stratified random sampling with proportional allocation (STRS) based on the empirical prompt population, and (c) stratified random sampling with equal proportional allocation (EQ).[7]

For SRS, the probabilities of being selected are equal for all elements in the target population. The advantage of SRS is that it eliminates subjectivity in the selection process. However, one major disadvantage of SRS is that the selected sample may not adequately represent the population structure, thereby leading to inaccurate estimation of the population distribution in model calibration.

In stratified sampling, the population is partitioned into L mutually exclusive groups called *strata*. Stratified sampling might be applied in either of the two following ways. For STRS, sometimes called *stratified probability proportional to size*, the goal is (a) to select an adequately representative sample of the population to be used for model building from which (b) to produce more precise and accurate statistical estimators (e.g., the population mean) than under SRS. When applied to the empirical prompt population, stratification with population proportional allocation is carried out as follows: within each stratum, a random sample without replacement of n_h units (h = 1, 2, … L) is independently selected proportional to its population representation. Let N_h (h = 1, 2…, L) denote the number of cases (or units) in stratum h; thus, the total population size $N = \sum_{h=1}^{L} N_h$.

For EQ, the same number of elements for each stratum is selected, regardless of their proportions in the population. This method is a potentially more robust approach to selecting samples, compared to stratification with population proportional allocation, because EQ is unaffected by shifts in population composition.

Stratification variables. For proportional and equal stratification, several different variables were used. They were country/territory, region, and status of English-as-best-language. These variables were chosen because, as noted earlier, indications of population invariance have been found based on language group and on native country, including for native countries close to one another. The attributes of each of those three stratification variables are described in Table

2. In addition to those three distinct variables, country/territory×language was also implemented for stratification.

Table 2

Stratification Variables Used in Sampling Design

Variable	Level of measurement	Description
Country/territory	Categorical: 37 individual countries /territories + 1 combined	The country/territory where the test-takers took their tests
Region	Categorical: 8 levels (Africa, Australia/Oceania, Eastern Asia, Europe, Middle East, Northern America, Central/South/Latin America & Caribbean, Southern/Southeastern Asia)	Grouped geographically based on test country/territory
English-as-best-language (language)	Categorical: 3 levels (yes/no/not available)	Test-takers' self-reported status

Note. The variable country/territory was automatically recorded by the ETS test administration system. There was no missing value for test country/territory. The variable English-as-best-language was self-reported. Test-takers who chose not to report whether English is their best language were grouped into one stratum named *not available*.

This set of variables led to six stratification-based sampling approaches. These approaches are listed, along with simple random sampling, in Table 3.

Table 3

Listing of Sampling Approaches and Their Abbreviations

Abbreviation	Sampling approach
EQ·CNTY	Equal allocation stratification by country/territory
EQ·LANG	Equal allocation stratification by language
STRS·CNLN	Proportional stratification by country/territory × language
STRS·CNTY	Proportional stratification by country/territory
STRS·LANG	Proportional stratification by language
STRS·RGN	Proportional stratification by region
SRS	Simple random sampling

Model calibration sample sizes. Model calibration sample sizes of 2,000 and 5,000 were used, depending on sampling method and type of scoring model. These sizes were chosen based on findings from Zhang (2012), which documented the superiority of these sample sizes over smaller ones in model calibration and cross-validation.

Model calibration sample sizes of 2,000 and 5,000 were used for STRS and SRS methods under both scoring models (i.e., generic and prompt-specific). EQ methods were used only for sample sizes of 5,000 because, in previous research, those methods did not perform effectively with smaller sample sizes (Zhang, 2012). For generic scoring, both EQ·CNTY and EQ·LANG were applied, but for prompt-specific scoring, only EQ·LANG was applied. EQ·CNTY was omitted because the small prompt population size would sometimes lead to exclusion of more than half of the country groups.

Comparison of the functioning of sampling approaches. The above sampling approaches were used to calibrate scoring models with examinees described above. The effectiveness of the different sampling approaches was evaluated in cross-validation data that consisted of all essays not used for model calibration. As a consequence, the cross-validation data sets necessarily differed by sampling approach because, by definition, the different approaches should lead to different calibration samples, resulting in variation across validation data sets. This variation, however, appeared immaterial, in that examination of the key demographics and examinee ability distributions suggested little difference among the data sets. In addition, because the initial population sizes for both generic and prompt-specific scoring were quite large, all resulting cross-validation data sets were generally big enough to allow for reasonably precise estimation of cross-validation indices.[8]

The impact of sampling approach on population invariance was examined across groups categorized by test country/territory, by language, and by region. The comparison was done using two criteria. The first criterion was human/machine agreement, which was indicated by four commonly used indices (i.e., Pearson correlation coefficient, quadratic-weighted kappa, exact percentage agreement, and standardized mean score difference; see Zhang, Williamson, Breyer, & Trapani, 2012, for computation details).

Because the population sizes differed by groups, for each index, the weighted mean and weighted standard deviation were computed as shown in Equation 1. (Of note is that, because the raw values of standardized mean score difference could be negative, absolute values were used to compute the weighted mean of this index.)

$$\overline{x}_w = \frac{\sum_{i=1}^{N} w_i x_i}{\sum_{i=1}^{N} w_i} \qquad (1)$$

where

$\bar{x}_w$ refers to the weighted mean of agreement index *w*,

w_i refers to the weight associated with observation/population group *i*,

x_i refers to the raw value of an agreement index for observation/population group *i*, and

N refers to the number of observations/population groups.

$$SD_w = \sqrt{\frac{N\sum_{i=1}^{N} w_i (x_i - \bar{x}_w)^2}{(N-1)\sum_{i=1}^{N} w_i}} \quad , \tag{2}$$

where

$\bar{x}_w$ refers to the weighted mean of agreement index *w*,

w_i refers to the weight associated with observation/population group *i*,

x_i refers to the raw value of an agreement index for observation/population group *i*, and

N refers to the number of observations/population groups.

The second criterion was based on the relationship of automated scores with external measures. Arguably, automated scores should correlate with external variables in a similar fashion to the correlations of human ratings with those same external variables. A substantial discrepancy in correlational pattern across population groups would suggest that the automated scoring system and humans were evaluating somewhat different aspects of writing.

For each population group, the difference of Pearson correlations, $r(e,x) - r(h,x)$, was computed, where *e* refers to e-rater scores, *h* refers to human ratings, and *x* refers to an external variable (i.e., GRE-V, GRE-Q, and GRE AW argument). Next, the weighted mean and weighted standard deviation of the correlation differences were computed and used to compare sampling approaches. A scoring model based on a sampling approach would be generally invariant if it produced a low weighted mean and a low weighted standard deviation of the correlation differences across population groups (i.e., country/territory, language, and region groups).

The computation procedures for the weighted mean and weighted standard deviation were the same as for the first criterion. Of note is that, for this second criterion, the weighted

mean was computed based on the absolute values, while the weighted standard deviation was computed on the raw values.

Because no conventionally accepted criteria for comparing the effectiveness of sampling approaches for automated scoring exist, a judgmental threshold was chosen to maximally differentiate among the sampling methods. The threshold was the percentage of times that a sampling approach yielded an index value above the mean across all draws for all sampling approaches. This method was similar to that used by Qian, von Davier, and Jiang (in press) in their comparison of sampling approaches for test equating.

Problematic group identification. For purposes of cross-validation, 17 of the 37 individual population country/territory groups were selected that were shown to be more problematic in either past automated scoring research or in the current data. The intention behind choosing a subset of groups was to maximize the opportunity for the various sampling methods to differentially reduce model fit problems that manifest as lack of invariance.

In the current data, problematic population groups were identified through the following indicators (using all 20 prompts; N = 258,860 test-takers): increase in mean squared error (MSE) of greater than 10% from group-specific to population-based scoring models or from group-specific to population-based models adjusted to match the mean of each group, absolute magnitude of the residuals of greater than 0.10 resulting from the population-based model for each group, a percentage of times greater than 45 that a second rater was needed to resolve a large discrepancy between the first human rating and operational automated score, and a correlation of less than 0.70 between human ratings and operational automated scores for a population group.[9] The selection threshold for each index was chosen based on either operational ETS practice (e.g., 0.70 as satisfactory correlation coefficient; Williamson, Xi, & Breyer, 2012) or a judgment as to what values might best identify bad model fit for a group.

From the current data, 16 country/territory groups that exceeded the thresholds on three or more indicators were selected for inclusion in the cross-validation data set. Those 16 countries/territories were India, China, Japan, Singapore, Taiwan, Bangladesh, Nepal, Thailand, Malaysia, Saudi Arabia, Puerto Rico, Philippines, Hong Kong, Colombia, Italy, and Indonesia. In addition, Korea was chosen due to model fit issues reported in previous research (e.g., Bridgeman et al., 2009, 2012). Finally, the U.S. population was also included as a baseline because its test population was the largest of all countries/territories.

With respect to region and language groups, results from the current data set showed that two language groups (i.e., those not indicating English as their best language and those who did not report on this variable) and several region groups (i.e., Middle East, Eastern Asia, and Southern/Southeastern Asia) had more violations on the aforementioned model fit indices than other language and region groups. Because the total number of language and region groups was small, and existing knowledge on population invariance by region was limited, all three language groups and eight region groups were included in the analyses.

Results

Population Invariance for Human/E-rater Agreement

Tables 4 and 5 each present two indices computed using cross-validation data. One index was the agreement between human and automated scores and the other was the variability of that agreement, a measure of population invariance. Table 4 gives results for sampling methods by three population-group classifications (i.e., 18 country/territory groups, three language groups, and eight region groups), and Table 5 presents the results by agreement index.

In both tables, sampling approaches are denoted by their abbreviations (see Table 4), to which is affixed the relevant model calibration sample size. Agreement is shown in the mean columns and was computed as follows. First, for each sampling method, the number of cases was identified for which a given agreement index produced a population-weighted mean (taken across population groups) that was above the average for all sampling methods on that index. This operation was conducted separately on each of the four indices: correlation coefficient; quadratic-weighted kappa; exact percentage agreement, for which higher values are desirable; and absolute standardized mean score difference between e-rater and human scores, for which lower values are preferred. Second, for Table 4, the number of above identified cases was combined across indices, whereas for Table 5, the cases were combined across classification groups. As such, in Table 4, the figures are the percentage of times, combined across indices, that a sampling approach produced an index value that was beyond the average for that index. In Table 5, the figures are the percentage of times, taken across classification groups, for which a sampling method produced an above average result for a particular index. (These data are further disaggregated in Table 6 and are presented in raw-value form.)

Table 4

Sampling Comparison Based on Population Invariance for Human/E-rater Agreement by Group Classification

Sampling method	*N* (mean/ SD/sum)	Percentage of occurrence with best weighted mean agreement (above average) and lowest weighted standard deviation (below average) collapsed across correlation coefficient, weighted kappa, standardized difference, and percentage agreement								
		Across country/territory			Across language			Across region		
		Mean	SD	Sum	Mean	SD	Sum	Mean	SD	Sum
EQ·CNTY·5,000	8/8/16	**63** [a]	**75** [a]	**69** [a]	25	**75** [a]	**50** [a]	38	**88** [a]	**63** [a]
EQ·LANG·5,000	16/16/32	38	**100** [a]	**69** [a]	44	**88** [a]	**66** [a]	38	**81** [a]	**59** [a]
STRS·CNLN·2,000	16/16/32	56	**56** [a]	56	**63** [a]	44	53	**63** [a]	**56** [a]	**59** [a]
STRS·CNLN·5,000	16/16/32	50	25	38	**69** [a]	**56** [a]	**63** [a]	**63** [a]	50	56
STRS·CNTY·2,000	16/16/32	**63** [a]	31	47	44	31	38	56	13	34
STRS·CNTY·5,000	16/16/32	56	38	47	**63** [a]	31	47	56	38	47
STRS·LANG·2,000	16/16/32	**69** [a]	50	**59** [a]	56	38	47	**63** [a]	38	50
STRS·LANG·5,000	16/16/32	38	19	28	**63** [a]	38	50	50	13	31
STRS·RGN·2,000	16/16/32	44	19	31	44	44	44	44	13	28
STRS·RGN·5,000	16/16/32	50	19	34	44	38	41	**63** [a]	19	41
SRS·2,000	16/16/32	44	**56** [a]	50	31	19	25	31	38	34
SRS·5,000	16/16/32	50	44	47	**69** [a]	31	50	**69** [a]	38	53

Note. Best was indicated by high values for correlation coefficient, quadratic-weighted kappa, and percentage agreement and by low values for standardized mean score difference. Mean = weighted mean, SD = weighted standard deviation, sum = numerically the average of the mean and SD columns and provides an indication of overall performance for a sampling approach.

[a] Indicates the three highest values within each column (also in boldface).

The variability of agreement is shown in the SD columns and was computed in a manner similar to the two-step process described above. Variability for a sampling method is represented by the weighted standard deviation of an agreement index taken across population groups. For each sampling method, the cell values indicate the percentage (taken across the four agreement indices for Table 4 or taken across the three population group classifications for Table 5) of all weighted standard deviation values that were below average.

Lastly, the cells in the sum columns numerically are the average of the cell values in the mean and SD columns.

Table 5

Sampling Comparison Based on Population Invariance for Human/E-rater Agreement by Agreement Index

Sampling method	*N* (mean/ SD/sum)	Percentage of occurrence with best weighted mean (above average) and lowest weighted standard deviation (below average) collapsed across country/territory, language, region groups											
		Correlation coefficient			Weighted kappa			Standardized difference			Percentage agreement		
		Mean	SD	Sum	Mean	SD	Sum	Mean	SD	Sum	Mean	SD	Sum
EQ·CNTY·5,000	6/6/12	33	**83** [a]	**58** [a]	17	**83** [a]	**50** [a]	17	50	33	**100** [a]	**100** [a]	**100** [a]
EQ·LANG·5,000	12/12/24	17	**100** [a]	**58** [a]	8	**92** [a]	**50** [a]	33	**67** [a]	50	**100** [a]	**100** [a]	**100** [a]
STRS·CNLN·2,000	12/12/24	**67** [a]	**58** [a]	**63** [a]	50	**50** [a]	**50** [a]	**75**	42	58	50	**58** [a]	**54** [a]
STRS·CNLN·5,000	12/12/24	**92** [a]	42	**67** [a]	**100** [a]	42	**71** [a]	42	**58** [a]	50	8	33	21
STRS·CNTY·2,000	12/12/24	58	17	38	50	25	38	58	33	46	50	25	38
STRS·CNTY·5,000	12/12/24	25	50	38	50	17	33	**83** [a]	50	**67** [a]	**75**	25	50
STRS·LANG·2,000	12/12/24	**67** [a]	17	42	**83** [a]	42	**63** [a]	50	**75** [a]	**63** [a]	50	33	42
STRS·LANG·5,000	12/12/24	**67** [a]	0	33	**75** [a]	17	46	58	33	46	0	42	21
STRS·RGN·2,000	12/12/24	**75** [a]	0	38	**75** [a]	17	46	25	50	38	0	33	17
STRS·RGN·5,000	12/12/24	50	17	33	42	25	33	**67** [a]	**58** [a]	**63** [a]	50	0	25
SRS·2,000	12/12/24	50	50	50	58	33	46	25	42	33	8	25	17
SRS·5,000	12/12/24	58	25	42	**75** [a]	25	**50** [a]	**67** [a]	**58** [a]	**63** [a]	50	42	46

Note. Best was indicated by high values for correlation coefficient, quadratic-weighted kappa, and percentage agreement and by low values for standardized mean score difference. Mean = weighted mean, SD = weighted standard deviation, sum is numerically the average of the mean and SD columns, which intends to provide an overall performance of a sampling approach.

[a] Indicates the three highest values within each column (also in boldface).

Table 6

Weighted Mean and Standard Deviation of Human/E-rater Agreement Indices

	Median value of the weighted mean and weighted standard deviation for an agreement index by population group classification							
	Correlation coefficient		Weighted kappa		Standardized difference		Percentage agreement	
Sampling method	Mean	SD	Mean	SD	Mean	SD	Mean	SD
For country/territory groups								
EQ·CNTY·5,000	0.717	0.065	0.645	0.088	0.088	0.157	62.667	4.486
EQ·LANG·5,000	0.712	**0.055** [a]	0.649	**0.076** [a]	0.095	**0.141** [a]	63.316	**2.626** [a]
STRS·CNLN·2,000	0.715	0.073	0.662	0.088	0.092	0.172	62.162	4.178
STRS·CNLN·5,000	0.717	0.072	0.664	0.089	0.097	0.179	62.227	4.323
STRS·CNTY·2,000	0.719	0.070	0.663	0.090	0.084	0.200	62.382	4.062
STRS·CNTY·5,000	0.717	0.073	0.664	0.092	0.084	0.164	62.663	4.011
STRS·LANG·2,000	0.720	0.072	0.663	0.088	0.089	0.164	62.536	3.965
STRS·LANG·5,000	0.719	0.071	0.663	0.090	0.093	0.182	62.192	4.241
STRS·RGN·2,000	0.719	0.073	0.663	0.090	0.099	0.176	61.941	4.218
STRS·RGN·5,000	0.716	0.071	0.661	0.090	0.092	0.164	62.140	4.386
SRS·2,000	0.715	0.069	0.658	0.087	0.088	0.177	62.125	4.357
SRS·5,000	0.715	0.071	0.659	0.089	0.079	0.171	62.309	4.211
For language groups								
EQ·CNTY·5,000	0.739	0.037	0.683	0.038	0.074	0.082	62.586	3.373
EQ·LANG·5,000	0.747	**0.027** [a]	0.706	**0.027** [a]	0.040	0.029	63.245	**1.900** [a]
STRS·CNLN·2,000	0.749	0.040	0.704	0.042	0.017	0.043	62.145	3.267
STRS·CNLN·5,000	0.751	0.038	0.705	0.043	0.030	0.031	62.144	3.534
STRS·CNTY·2,000	0.749	0.041	0.702	0.047	0.032	0.034	61.262	2.995
STRS·CNTY·5,000	0.749	0.039	0.701	0.043	0.026	0.030	62.609	3.141
STRS·LANG·2,000	0.750	0.043	0.704	0.042	0.031	0.032	62.511	3.066
STRS·LANG·5,000	0.741	0.040	0.705	0.043	0.027	0.031	62.171	3.175
STRS·RGN·2,000	0.750	0.041	0.705	0.044	0.032	0.028	61.780	3.114
STRS·RGN·5,000	0.745	0.040	0.705	0.041	0.031	**0.027** [a]	61.326	3.260
SRS·2,000	0.749	0.041	0.704	0.045	0.036	0.037	61.981	3.498
SRS·5,000	0.751	0.040	0.707	0.045	0.023	0.028	61.973	3.294
For region groups								
EQ·CNTY·5,000	0.719	0.061	0.652	0.063	0.105	0.096	62.564	4.370
EQ·LANG·5,000	0.717	**0.047** [a]	0.662	**0.059** [a]	0.099	**0.093** [a]	63.200	**2.417** [a]
STRS·CNLN·2,000	0.723	0.061	0.669	0.073	0.082	0.118	62.147	4.041
STRS·CNLN·5,000	0.726	0.062	0.672	0.076	0.088	0.127	62.173	4.230
STRS·CNTY·2,000	0.723	0.063	0.669	0.077	0.081	0.119	62.340	3.939
STRS·CNTY·5,000	0.722	0.062	0.669	0.074	0.078	0.121	62.615	3.855
STRS·LANG·2,000	0.725	0.063	0.671	0.073	0.080	0.115	62.506	3.816
STRS·LANG·5,000	0.724	0.063	0.671	0.077	0.085	0.129	62.171	4.094
STRS·RGN·2,000	0.724	0.064	0.670	0.077	0.087	0.123	61.810	4.050
STRS·RGN·5,000	0.724	0.062	0.669	0.076	0.083	0.116	62.112	4.135
SRS·2,000	0.725	0.061	0.670	0.074	0.081	0.119	62.011	4.226
SRS·5,000	0.724	0.064	0.671	0.077	0.066	0.124	62.249	4.074

Note. Mean = weighted mean, SD = weighted standard deviation.

[a] Indicates the lowest median weighted standard deviation among all sampling approaches (also in boldface).

The percentages in the mean, SD, and sum columns together provide different perspectives on population invariance. The SD is the most direct indicator of the extent to which a sampling method reduces lack of invariance; the mean column suggests whether that reduction comes at the price of lower agreement; and the sum column shows how a sampling method balances, in a compensatory manner, the two factors.

The previous tables show how often a sampling approach produced high means and low standard deviations in human and e-rater agreement indices across population groups, but give no idea of the magnitude of the differences among the approaches. To offer a sense of the sizes of the differences in weighted mean and standard deviation among sampling approaches, Table 6 provides the median value of the weighted mean and weighted standard deviation of an index associated with each sampling approach. The results are presented for each of the four agreement indices separately for each of the three population group classifications.

Similar to the findings presented in Tables 4 and 5, results in Table 6 indicate that sampling approach EQ·LANG·5,000 yielded the lowest degree of variability, manifested by the smallest weighted standard deviation across population groups for all agreement indices except for one index in one population group classification (i.e., the standardized mean score difference across language groups). Of note is that the reduction in variability due to EQ·LANG·5,000 was quite significant in some cases. For example, compared with SRS·5,000 (i.e., the approach most similar to current e-rater operational practice), the weighted standard deviation was reduced by 15% to 42%. Moreover, there was minimal agreement loss, if any, with EQ·LANG·5,000.

In summary, EQ·CNTY·5,000 and EQ·LANG·5,000 appeared to be the best in reducing the lack of population invariance for human/e-rater correlation without agreement loss, regardless of whether the results are organized by agreement index or by population group classification.

Population Invariance for Human/E-rater Correlation Pattern With External Variables

Tables 7 and 8 show the magnitude of the correlational differences, that is, the quantity $r(e,x) - r(h–x)$, and variability of such differences across population groups resulting from different sampling methods from two perspectives. One perspective (shown in Table 7) gives the

results by the three population-group classifications, and the other perspective (shown in Table 8) is by three external variables (i.e., GRE-Q, GRE-V, and GRE AW argument). In the tables, the magnitude is quantified by the sample-size weighted mean of the absolute correlation differences across population groups, and variability is quantified by the sample-size weighted standard deviation of the differences in correlation across population groups.

Table 7

Sampling Comparison Based on Population Invariance for Correlation Difference by Group Classification

		Percentage of occurrence with lowest weighted mean absolute differences for $r(e,x) - r(h,x)$ and lowest weighted standard deviation of the differences (below average) collapsed across quantitative, verbal, and AW argument								
		Across country/territory			Across language			Across region		
Sampling method	*N* (mean/ SD/sum)	Mean	SD	Sum	Mean	SD	Sum	Mean	SD	Sum
EQ·CNTY·5,000	6/6/12	0	50	25	0	**67** [a]	33	0	33	17
EQ·LANG·5,000	12/12/24	**83** [a]	**75** [a]	**79** [a]	**83** [a]	**58** [a]	**71** [a]	**83** [a]	42	**63** [a]
STRS·CNLN·2,000	12/12/24	50	42	46	33	17	25	33	33	33
STRS·CNLN·5,000	12/12/24	25	42	33	8	**58** [a]	33	8	**67** [a]	38
STRS·CNTY·2,000	12/12/24	42	**75** [a]	**58** [a]	25	50	38	42	50	46
STRS·CNTY·5,000	12/12/24	50	42	46	50	33	42	67	50	**58** [a]
STRS·LANG·2,000	12/12/24	58	42	50	50	**58** [a]	54	50	42	46
STRS·LANG·5,000	12/12/24	58	50	54	50	50	50	50	**58** [a]	54
STRS·RGN·2,000	12/12/24	**83** [a]	58	**71** [a]	**75** [a]	**75** [a]	**75** [a]	**83** [a]	**75** [a]	**79** [a]
STRS·RGN·5,000	12/12/24	**83** [a]	25	54	**75** [a]	50	**63** [a]	**83** [a]	33	**58** [a]
SRS·2,000	12/12/24	50	**67** [a]	**58** [a]	33	33	33	58	42	50
SRS·5,000	12/12/24	50	58	54	42	42	42	42	42	42

Note. Mean = weighted mean, SD = weighted standard deviation, sum = numerically the average of the mean and SD columns, which provides an overall index for the performance of a sampling approach. In $r(e,x) - r(h,x)$, *r* refers to correlation coefficient, *e* refers to e-rater scores, *h* refers to human ratings, and *x* refers to an external variable (i.e., GRE-Verbal, GRE-Quantitative, and GRE Analytic Writing argument).

[a] Indicates the three highest values within each column (also in boldface).

Table 8

Sampling Comparison Based on Population Invariance for Correlation Difference by External Variable

Sampling method	*N* (mean/ SD/sum)	Percentage of occurrence with lowest weighted mean absolute differences for $r(e,x) - r(h,x)$ and lowest weighted standard deviation of the differences (below average) collapsed across country/territory, language, and region groups								
		Quantitative			Verbal			Argument		
		Mean	SD	Sum	Mean	SD	Sum	Mean	SD	Sum
EQ·CNTY·5,000	6/6/12	0	**67** [a]	33	0	50	25	0	33	17
EQ·LANG·5,000	12/12/24	**100** [a]	42	**71** [a]	50	**75** [a]	**63** [a]	**100** [a]	58	**79** [a]
STRS·CNLN·2,000	12/12/24	33	17	25	25	**58** [a]	42	58	17	38
STRS·CNLN·5,000	12/12/24	0	**58** [a]	29	25	42	33	17	**67** [a]	42
STRS·CNTY·2,000	12/12/24	**58** [a]	**67** [a]	**63** [a]	25	**58** [a]	42	25	50	38
STRS·CNTY·5,000	12/12/24	**58** [a]	17	38	50	50	50	58	58	58
STRS·LANG·2,000	12/12/24	42	33	38	**75** [a]	50	**63** [a]	42	58	50
STRS·LANG·5,000	12/12/24	50	50	50	67	42	54	42	**67** [a]	54
STRS·RGN·2,000	12/12/24	50	50	50	**92** [a]	**58** [a]	**75** [a]	**100** [a]	**100** [a]	**100** [a]
STRS·RGN·5,000	12/12/24	**67** [a]	50	**58** [a]	**75** [a]	8	42	**92** [a]	58	**75** [a]
SRS·2,000	12/12/24	**58** [a]	50	54	33	25	29	50	**67** [a]	58
SRS·5,000	12/12/24	50	50	50	25	50	38	58	42	50

Note. Mean = weighted mean, SD = weighted standard deviation, sum = numerically the average of the mean and SD columns, which provides an overall index for the performance of a sampling approach. In $r(e,x) - r(h,x)$, *r* refers to correlation coefficient, *e* refers to e-rater scores, *h* refers to human ratings, and *x* refers to an external variable (i.e., GRE-Verbal, GRE-Quantitative, and GRE Analytic Writing argument).

[a] Indicates the three highest values within each column (also in boldface).

Both tables can be read in a similar way to Tables 4 and 5. The magnitude of the differences is provided in the mean columns, which present the percentage of absolute differences for which a sampling method produced a value that was below average across draws for all sampling approaches with all three external variables. High percentages indicate small differences between e-rater and humans' correlations with the same external variables, suggesting similarity in score meaning between the two scoring methods. Values in the SD columns indicate the percentage (taken across external variables, for Table 7, or across population classifications, for Table 8) of all weighted standard deviation values that were below average. High percentages are desirable because they indicate greater invariance (of the correlational difference) across groups, which offers further evidence for similarity in score meaning. Last, the sum columns give the average values in the mean and SD columns, which represent an overall outcome.

Results in Tables 7 and 8 suggest that EQ·LANG·5,000; STRS·RGN·2,000; and STRS·RGN·5,000 were the best with regard to minimizing the magnitude and variance of correlational differences with external measures between human ratings and e-rater scores across population groups.

Although Tables 7 and 8 show how frequently a sampling approach reduced the magnitude and variance of the correlational pattern difference, they offer no indication of the absolute size of the differences among the approaches. Table 9 presents the median value of the weighted mean and standard deviation of the correlational differences between human and e-rater with external variables across population groups associated with each sampling approach. Of note is that differences among methods with respect to the magnitude of the correlational differences were generally small, especially in the case of GRE AW argument, the external measure most closely related to writing skill. Here, the largest difference among sampling approaches was only 0.007 within country/territory group classification, 0.008 within language group classification, and 0.005 within region group classification.

Table 9

Weighted Mean and Standard Deviation of the Correlation Differences of Human and E-rater With External Variables

	Median value of the weighted mean and weighted standard deviation of the (absolute) differences between human and e-rater correlationwith external variables					
	With quantitative		With verbal		With argument	
Sampling method	Mean	SD	Mean	SD	Mean	SD
For country/territory groups						
EQ·CNTY·5,000	0.050	**0.032** [a]	0.057	0.047	0.080	0.029
EQ·LANG·5,000	0.038	0.034	0.055	**0.044** [a]	**0.073** [a]	**0.018** [a]
STRS·CNLN·2,000	0.042	0.033	0.057	0.047	0.078	0.025
STRS·CNLN·5,000	0.045	**0.032** [a]	0.059	0.047	0.080	0.021
STRS·CNTY·2,000	0.038	0.033	0.058	0.048	0.078	0.023
STRS·CNTY·5,000	0.041	0.034	0.057	0.048	0.079	0.022
STRS·LANG·2,000	**0.037** [a]	0.035	0.059	0.049	0.078	0.023
STRS·LANG·5,000	0.041	0.034	0.052	0.047	0.077	0.026
STRS·RGN·2,000	0.039	0.033	0.055	0.048	0.076	0.022
STRS·RGN·5,000	0.041	0.035	**0.049** [a]	0.048	0.074	0.028
SRS·2,000	0.041	0.033	0.059	0.048	0.075	0.022
SRS·5,000	0.042	0.034	0.061	0.052	0.077	0.026
For language groups						
EQ·CNTY·5,000	0.076	0.039	0.051	0.023	**0.068** [a]	**0.012** [a]
EQ·LANG·5,000	**0.033** [a]	0.038	0.054	**0.022** [a]	0.069	0.016
STRS·CNLN·2,000	0.055	0.041	0.054	0.026	0.072	0.015
STRS·CNLN·5,000	0.060	0.039	0.055	0.024	0.074	0.013
STRS·CNTY·2,000	0.052	0.038	0.055	0.025	0.071	**0.012** [a]
STRS·CNTY·5,000	0.051	0.040	0.053	0.024	0.070	0.014
STRS·LANG·2,000	0.049	0.038	0.049	0.026	0.071	0.013
STRS·LANG·5,000	0.054	0.041	0.049	0.025	0.076	0.013
STRS·RGN·2,000	0.055	0.039	0.050	0.025	0.071	**0.012** [a]
STRS·RGN·5,000	0.054	**0.037** [a]	**0.046** [a]	0.025	**0.068** [a]	0.013
SRS·2,000	0.053	0.040	0.055	0.025	0.072	**0.012** [a]
SRS·5,000	0.054	0.040	0.057	0.023	0.072	0.014
For region groups						
EQ·CNTY·5,000	0.051	**0.030** [a]	0.048	0.013	0.079	0.021
EQ·LANG·5,000	**0.039** [a]	0.032	0.050	0.012	**0.074** [a]	**0.015** [a]
STRS·CNLN·2,000	0.044	0.033	0.050	0.012	0.078	0.018
STRS·CNLN·5,000	0.046	0.032	0.051	0.012	0.079	0.016
STRS·CNTY·2,000	0.041	0.033	0.051	0.012	0.077	0.017
STRS·CNTY·5,000	0.043	0.035	0.049	0.012	0.077	**0.015** [a]
STRS·LANG·2,000	0.039	0.035	0.048	0.013	0.079	0.016
STRS·LANG·5,000	0.043	0.034	0.045	0.012	0.077	0.017
STRS·RGN·2,000	0.042	0.033	0.047	**0.011** [a]	0.077	**0.015** [a]
STRS·RGN·5,000	0.043	0.035	**0.042** [a]	0.013	**0.074** [a]	0.017
SRS·2,000	0.043	0.034	0.051	0.012	0.077	0.016
SRS·5,000	0.042	0.033	0.054	0.014	0.079	0.017

Note. Mean = weighted mean.

[a] Indicates the lowest median weighted mean and standard deviation among all sampling approaches (also in boldface).

With respect to the variance of the pattern differences, it is worth contrasting EQ·LANG, the best approach for reducing population variance in both e-rater/human agreement and correlation pattern with external measures, with three other approaches: STRS·CNLN and STRS·CNTY, which performed most optimally in model calibration (as reported in Zhang, 2012) and SRS because of its common use in operational e-rater scoring. In the case of EQ·LANG·5,000, the reduction in variance (of the pattern differences) was (a) 17% compared with STRS·CNLN·5,000; (b) 22% compared with STRS·CNTY·5,000; and (c) 44% compared with SRS·5,000. These reductions came with virtually no loss in the magnitude of the correlation between automated scores and GRE AW argument. The loss in using EQ·LANG·5,000 was (a) 0.007 (i.e., 10%) compared with STRS·CNTY·5,000; (b) 0.006 (i.e., 8%) compared with STRS·CNTY·5,000; and (c) 0.004 (i.e., 5%) compared with SRS·5,000.

Also of note is that the reduction in variance was smaller than that produced for human/ e-rater agreement. For example, compared with SRS·5,000 (i.e., the approach most similar to current operational practice), implementing EQ·LANG·5,000 reduced the variance of the differences in the external correlation for human and for e-rater by 10% on average. In contrast, EQ·LANG·5,000 reduced the variance in human/e-rater agreement by more than 29% on average from the SRS·5,000 sampling approach.

Finally, Table 10 provides an overall summary of the effectiveness of each sampling method in maximizing population invariance. The left side of the table aggregates the findings for human/e-rater agreement given in Table 4 (across population group classification) and Table 5 (across agreement indices). The right side aggregates the results regarding the e-rater/human correlation pattern differences with external variables given in Tables 7 and 8.

Table 10 documents that, as mentioned, EQ·LANG is the most desirable approach across both invariance criteria (i.e., human and external variables), given that it yielded the highest overall percentages in maximizing invariance (see sum columns). Although it appears that there is a loss in human/e-rater agreement for EQ·LANG, the absolute loss in agreement index values was negligible, as previously discussed (see Table 6).

Additionally, EQ·CNTY also produced low variance in human/e-rater agreement across population groups and was the second best approach overall when using human ratings as an invariance criterion. Even though STRS·CNLN and STRS·LANG sampling approaches yielded the highest human/e-rater agreement, those two approaches were not preferable in maintaining

the invariance of the e-rater score meaning across population groups, regardless of the criterion. Finally, the STRS·RGN sampling approach was comparable to EQ·LANG in reducing the correlation differences between human and e-rater with external measures.

Table 10

Aggregated Results of Sampling Comparison Based on Population Invariance

Sampling approach	Percentage of occurrences with best weighted mean agreement (above average) and lowest weighted standard deviation (below average)				Percentage of occurrence with lowest weighted mean absolute differences for $r(e,x) - r(h,x)$ and lowest weighted standard deviation of the differences (below average)			
	N (mean/ SD/sum)	Percentage			*N* (mean/ SD/sum)	Percentage		
		Mean	SD	Sum		Mean	SD	Sum
EQ·CNTY	24/24/48	42	**79** [a]	**60** [a]	18/18/36	0	50	25
EQ·LANG	48/48/96	40	**90** [a]	**65** [a]	36/36/72	**83** [a]	**58** [a]	**71** [a]
STRS·CNLN	96/96/192	**60** [a]	48	54	72/72/144	27	44	35
STRS·CNTY	96/96/192	56	30	44	72/72/144	46	50	48
STRS·LANG	96/96/192	**57** [a]	33	44	72/72/144	53	50	52
STRS·RGN	96/96/192	48	25	37	72/72/144	**81** [a]	**53** [a]	**67** [a]
SRS	96/96/192	49	38	43	72/72/144	46	47	47

Note. The left side of the table (columns 2 through 5) aggregates Tables 4 and 5. The right side of the table (columns 6 through 9) aggregates Tables 7 and 8. The results for sample sizes of 2,000 and 5,000 for each sampling approach were combined by taking the average. External variables include GRE-Quantitative, GRE-Verbal, and GRE Analytic Writing argument essay writing. Mean = weighted mean, sum= numerically the average of the mean and SD columns, which provides an overall index of the performance of a sampling approach. In $r(e,x) - r(h,x)$, *r* refers to correlation coefficient, *e* refers to e-rater scores, *h* refers to human ratings, and *x* refers to an external variable (i.e., GRE-Verbal, GRE-Quantitative, and GRE Analytic Writing argument).

[a] Indicates the most desirable two values within each column (also in boldface).

Discussion

This study attempted to identify sampling approaches that produced the greatest similarity of score meaning between human and automated methods across population groups. Similarity of score meaning (i.e., invariance of construct validity) across population groups is one common conception of test fairness (AERA et al., 1999; Messick, 1998). In this study, similarity was investigated at the item level because assessments like the GRE AW section typically consist of very few items. Similarity was operationally defined in terms of (a) the (high) magnitude and (high) invariance of the agreement between automated scores and human ratings and (b) the (low) magnitude and (high) invariance of the correlational pattern differences of human and automated scores with external variables (across population groups). It is important to note that these two criteria are related: maximizing agreement between human and automated scores should logically lead to minimizing the differences in their correlations with external variables.

As noted earlier, in cases where automated scores correlate with external variables higher than do human ratings, attempts to maximize the similarity of score meaning across population groups may have the effect of reducing in the overall population (or in selected subgroups) the relationship of automated scores with those external variables, including with other measures of writing. Such a reduction is of concern only if the correlations are viewed solely as evidence of the external-relations aspect of validity and not also as construct-validity invariance criteria. If and when a reduction in external relations occurs, how to balance invariance with external relations is a value judgment that must be resolved, based on the claims testing programs wish to make about the meaning of test scores.

In this study, results showed that with certain sampling designs, departures from population invariance were substantially reduced. The traditional simple random sampling approach did not perform as well as several other approaches. Instead, equal proportional allocation by language with a large sample size of 5,000 (EQ·LANG·5,000) outperformed all other approaches by yielding the least variation in (a) human/machine agreement and (b) correlational pattern difference between human and machine with external measures, with no or negligible loss in agreement strength across population groups or in the correlation of automated scores with the argument essay.

Additionally, depending on the aspects of invariance that one intends to optimize, there are other sampling approaches that one may consider. When the correlation with human ratings was the primary invariance criterion, equal stratification by country/territory (i.e., EQ·CNTY) was effective in yielding low variation across population groups with minimal agreement loss. However, this approach appeared to be one of the least favorable in reducing the discrepancies between human and e-rater correlation patterns with external variables within, and across, groups. For the latter type of invariance, STRS·RGN was an effective approach.

Why were equal allocation approaches effective in maximizing population invariance for human/machine agreement? E-rater models are built by regressing human scores on feature values. To the extent that the regression is different across population groups, the sampling approach used in model calibration could exacerbate or dampen the impact of those (regression) differences on the scoring model.

Table 11 shows how feature weights vary when models are created within individual country/territory. Included are four English-speaking examinee populations and a subset of the problematic (non–English-speaking) countries/territories (listed in the Method section). This subset had both low correlations between human and e-rater, as well as large human/e-rater standardized mean differences. These problematic countries/territories are further divided into ones with higher human than e-rater scores and others with the opposite pattern. The table shows noticeably higher weights for the English-speaking countries than for the two groups of problematic countries on organization (mean weights = 29% vs. 23% and 15%, respectively) and development (29% vs. 23% and 19%, respectively). The opposite occurs with respect to grammar (6% vs. 10% and 10%, respectively) and mechanics (6% vs. 23% and 19%, respectively).

Further, as Table 11 shows, the average human scores, an indication of overall proficiency in writing, also varied from one population group to another. For example, English-speaking countries/territories received higher human ratings than non–English-speaking countries/territories. In addition, as shown in the far right columns of the table, the writing feature profiles varied noticeably. For example, some population groups had well-rounded, better English writing proficiency as measured by e-rater (e.g., Canada), while some population groups exhibited irregular profiles and lower proficiency (e.g., Nepal).

Table 11

Human Scores, Feature Weights, and Outliers for Selected Countries/Territories

			Percentage feature weight in group-specific generic models									Count of outliers	
Country /territory	Sample size	Average human score	Grammar	Mechanics	Style	Usage	Word choice	Word length	Develop-ment	Organiza-tion	Colloca-tion preposi-tion	+1 SD	-1 SD
Nonproblematic groups and English-speaking countries/territories													
US	176,715	3.8	3.40	6.10	1.40	7.10	6.30	6.90	32.40	34.30	1.90	5	0
CA	2,285	4.0	5.00	5.20	1.60	11.90	10.10	4.20	29.40	29.20	3.40	8	0
UK	1,114	3.9	6.30	9.30	-1.20	12.30	10.70	1.40	27.70	29.40	4.00	9	0
AU	277	3.8	8.90	4.20	1.00	21.50	12.70	0.80	26.30	22.00	2.50	7	0
Mean	-	3.9	5.90	6.20	0.70	13.20	9.95	3.33	28.95	28.73	2.95	-	-
Problematic groups (e.g., low agreement between e-rater and human; standardized mean score difference of e - h < 0)													
TW	884	2.7	11.40	12.80	-0.50	7.00	5.80	9.30	27.10	23.50	3.50	0	4
NP	749	2.7	11.90	19.20	4.50	7.80	1.80	8.60	19.50	18.90	7.80	1	6
TH	705	2.9	15.50	10.40	-2.00	12.80	2.80	7.20	25.00	23.10	5.10	0	2
SA	572	2.7	6.20	12.30	1.80	12.10	4.50	7.90	20.50	24.00	10.60	0	8
IN	27,752	2.9	9.30	14.90	4.80	10.60	3.20	5.80	20.40	22.50	8.50	0	3
BD	779	2.8	5.60	14.60	1.60	6.30	6.30	4.20	24.00	24.50	12.90	0	5
ID	261	3.0	7.10	12.60	5.10	8.50	3.20	5.60	23.50	24.10	10.30	0	2
Mean	-	2.8	9.57	13.83	2.19	9.30	3.94	6.94	22.86	22.94	8.39	-	-
Problematic groups (e.g., low agreement between e-rater and human; standardized mean score difference of e - h > 0)													
CN	27,133	3.1	12.10	16.10	2.20	8.50	6.90	7.20	21.60	18.20	7.10	2	0
SG	1,509	3.3	7.60	16.60	2.20	17.2	15.80	2.40	17.90	15.00	5.30	1	0
MY	692	3.0	9.80	13.70	0.00	13.10	8.00	11.50	19.70	16.90	7.30	0	0
HK	394	3.2	9.70	8.20	2.50	23.80	13.10	2.00	20.50	10.70	9.50	1	0
Mean	-	3.1	9.80	13.65	1.73	15.13	10.95	5.78	19.93	15.20	7.30	-	-

These patterns imply that the regressions are considerably different across population groups and that sampling methods that tend to overweight large groups (e.g., U.S. test-takers) may contribute to a lack of population invariance. For example, because STRS (and potentially SRS) approaches differentially value each population group, the regression of human ratings on e-rater features will be disproportionally determined by the dominant population groups. Lack of population invariance is likely to be exacerbated when within-group regressions for smaller groups differ considerably from the dominate groups' regressions.

In contrast, EQ·LANG and EQ·CNTY should balance regressions across population groups for variables directly associated with GRE writing performance, hence maximizing the population invariance for the human/machine correlation. Because equal allocation includes in the calibration sample the same number of examinees from each group, the influence of large groups is dampened; that is, the regression of human ratings on e-rater features for the population is conceptually similar to an average across groups rather than being dominated by the large groups. Further, equal allocation approaches by language and by country/territory function to balance the distribution of the automated scores across these population groups, instead of pulling the distribution toward dominant groups. This balance reduces the incidence of human/machine distributionwise discrepancies, thereby enhancing that type of population invariance.

Why did STRS·RGN (in addition to EQ·LANG) work in minimizing departures from invariance regarding external variables? Regions differ from one another in the distribution of undergraduate major, which is, in turn, differentially associated with the three invariance criteria (i.e., GRE-V, GRE-Q, GRE AW argument). For example, examinees from Eastern Asia tend to be from more quantitatively oriented academic majors than examinees from North America, producing an association between region and GRE-Q score. STRS·RGN takes into account this relationship with the population invariance criteria in selecting the model calibration sample. Other sampling strategies may not work as well because they do not capitalize as effectively on this relationship. The only other strategy that functioned like STRS·RGN on this criterion, EQ·LANG, may have done so because it represents each group equally on a variable that directly relates to the external measures.

While sampling approaches appear to have an impact on population invariance, two other factors may directly contribute to this phenomenon. One potential cause is a lack of population invariance in human ratings (e.g., differences in interrater reliability), which might, in turn, cause

the automated scores to correlate differentially with human scores across groups.[10] This cause occurs when humans are the target of prediction in automated scoring modeling. (See Zhang, 2012, for further discussion.) One other potential cause is the linear modeling approach itself. Linear modeling, particularly linear regression, tends to shrink the scale by pulling the predicted values toward the grand mean of the population. Consequently, an individual whose human rating is lower than the grand mean tends to receive an automated score that is higher than the human score, and vice versa. In practice, the predicted values are usually rescaled (i.e., enlarged) to deal with the above mentioned shrinkage by matching to the human distribution. Such rescaling possibly contributes to the directional distributionwise differences between e-rater and human scores for a population group.

Practical Implications

The findings from this study imply that, depending on the criterion chosen, one can use any one of three approaches to optimize invariance within, and across, population groups. To enhance the population invariance of automated scores simultaneously across the two criteria considered in this study, equal allocation by language is recommended for model calibration. This strategy is also robust in dealing with changes in examinee population composition. To maximize a specific aspect of population invariance, equal allocation by country/territory and proportional stratification by region can also be considered. EQ·CNTY is preferable in maximizing population invariance regarding human/machine agreement, while STRS·RGN is favorable in minimizing departures from invariance regarding correlational pattern differences.

Limitations

This research is subject to the following limitations. First, the conclusions are most properly limited to regression-based automated essay scoring systems that function similarly to e-rater. Findings may not be generalizable beyond the expository, academic writing genre that is measured by the issue prompt in the GRE General Test, stratification variables, and the cross-validation criteria and variables used in this study.

Second, the quality of the human ratings was not directly evaluated due to the lack of a second randomly assigned human rater. Instead, other relevant data were examined to confirm the rating quality. Due to this limitation, direct comparisons between machine/human and human/human agreement were not possible.

Third, results are best generalized to the process for creating cross-validation data sets applied here. Although the data sets for the different sampling approaches were not the same, they were generally large in size and highly comparable to one another in terms of the demographics and test-taker ability levels.

Last, guidelines were not available to establish thresholds to distinguish better and worse sampling approaches. Additionally, there was no statistical mechanism to aggregate quantitatively the results from the various indices used in comparing sampling approaches. Consequently, judgment was used to establish thresholds and to come to overall conclusions about sampling methods. Other thresholds or approaches to aggregation might have produced somewhat different results.

Recommendations for Additional Research

This research is best viewed as a starting point to investigate sample selection as an underexamined component in automated scoring. The following research topics are suggested.

One, researchers are encouraged to examine whether the gain in population invariance with certain sampling designs can be extended to other test countries/territories and other population group classifications, such as by gender and by ethnicity.

Two, researchers are encouraged to compare the impact of different cross-validation sampling methods, including the commonly used approach of conducting the cross-validation on the responses remaining from model calibration. Different approaches may cause variations in cross-validation samples (which may affect the evaluation results). Adopting different target populations may also cause differences in the cross-validation samples.

Three, the use of confidence intervals might be explored to compare sampling approaches. For example, a confidence interval for each sampling approach could be built around the weighted standard deviation of the agreement between human and automated scores across groups. Using such intervals, two sampling approaches might then be compared to determine the significance of the differences.

Four, simulation-based approaches might be useful. In such a controlled environment, the composition of a population or of population groups can be systematically designed. This characteristic allows the impact of sampling on population invariance to be disaggregated from extraneous factors that can also be designed into the simulation data.

Finally, more research is needed to investigate potential nonsampling causes for the lack of population invariance in automated scores, and, more importantly, explore solutions. For instance, alternative automated scoring features that are less sensitive to essay length might be investigated, since length-related features (i.e., organization and development) appear to be associated with the lack of population invariance. Modeling approaches that do not heavily rely on a single calibration sample (e.g., neural network) may also help reduce the lack-of-invariance problem.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *The standards for educational and psychological testing.* Washington, DC: American Psychological Association.

Attali, Y. (2007). *Construct validity of e-rater in scoring TOEFL essays* (Research Report No. RR-07-21). Princeton, NJ: Educational Testing Service.

Attali, Y., & Powers, D. (2009). Validity of scores for a developmental writing scale based on automated scoring. *Educational and psychological measurement*, *69*(6), 978–993.

Breland, H., Kubota, M., Nikerson, K., Trapani, C., & Walker, M., (2004). *New SAT writing prompt study: Analyses of group impact and reliability* (Research Report No. RR-04-03). Princeton, NJ: Educational Testing Service.

Bridgeman, B., Trapani, C., & Attali, Y. (2009, May). *Considering fairness and validity in evaluating automated scoring*. Paper presented at the annual meeting of the National Council of Measurement in Education, San Diego, CA.

Bridgeman, B., Trapani, C., & Attali, Y. (2012). Comparison of human and machine scoring of essays: Differences by gender, ethnicity, and country. *Applied Measurement in Education*, *25*(1), 27–40.

Deane, P. (2011). *Writing assessment and cognition* (Research Report No. RR-11-14). Princeton, NJ: Educational Testing Service.

Dorans, N. J., & Holland, P. W. (2000). Population invariance and the equitability of tests: Basic theory and the linear case. *Journal of Educational Measurement*, *37*(4), 281–306.

Dorans, N. J., Schmitt, A. P., & Bleistein, C. A. (1992). *The standardization approach to assessing differential speededness* (Research Report No. RR-88-31). Princeton, NJ: Educational Testing Service.

Drasgow, F., Luecht, R. M., & Bennett, R. E. (2006). Technology and testing. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 471–515). Westport, CT: Praeger Publishers.

Educational Testing Service. (2010). *ETS automated scoring technologies*. Retrieved from http://www.ets.org/s/commonassessments/pdf/AutomatedScoring.pdf

Enright, M. K., & Quinlan, T. (2010). Complementing human judgment of essays written by English language learners with e-rater scoring. *Language Testing*, *27*(3), 317–334.

French, B. F., & Mantzicopoulos, P. (2007). An examination of the first/second-grade form of the pictorial scale of perceived competence and social acceptance: Factor structure and stability by grade and gender across groups of economically disadvantaged children. *Journal of School Psychology*, *45*, 311–331.

Haberman, S. J. (2007). Electronic essay grading. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics: Vol. 26. Psychometrics* (pp. 205–233). Amsterdam, The Netherlands: Elsevier.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, *50*(1), 1–73

Lee, Y.-W., Gentile, C., & Kantor, R. (2008). *Analytic scoring of TOEFL CBT essays: Scores from humans and e-rater engine* (TOEFL Research Report No. RR-81). Princeton, NJ: Educational Testing Service.

Mattern, K. D., Patterson, B. F., Shaw, E. J., Kobrin, J. L., & Barbuti, S. M. (2008). *Differential validity and prediction of SAT* (College Board Research Report No. 2008-4). New York, NY: College Board.

Messick, S. (1998). *Consequences of test interpretation and use: The fusion of validity and values in psychological assessment* (Research Report No. RR-98-48). Princeton, NJ: Educational Testing Service.

Page, E. B. (1966). The imminence of grading essays by computer. *Phi Delda Kappa*, *47*, 238–243.

Pearson Education Inc. (2010). *Intelligent essay assessor (IEA) fact sheet*. Retrieved from http://kt.pearsonassessments.com/download/IEA-FactSheet-20100401.pdf

Qian, J., von Davier, A., & Jiang, Y. (in press). Achieving a stable scale for an assessment with multiple forms—Weighting test samples in IRT linking and equating. In *Proceedings of the 2012 Annual International Meeting of the Psychometric Society*.

Ramineni, C., Trapani, C., Williamson, D., Davey, T., & Bridgeman, B. (2012a). *Evaluation of e-rater for the GRE issue and argument prompts* (Research Report No. RR-12-02). Princeton, NJ: Educational Testing Service.

Ramineni, C., Trapani, C., Williamson, D., Davey, T., & Bridgeman, B. (2012b). *Evaluation of e-rater for the TOEFL independent and integrated prompts* (Research Report No. RR-12 -06). Princeton, NJ: Educational Testing Service.

Ruth, R. (2000). Using structural equation modeling to test for differential reliability and validity: An empirical demonstration. *Structural Equation Modeling*, *7*(1), 124–141.

Wang, J., & Brown, M. S. (2008). Automated essay scoring versus human scoring: A correlation study. *Contemporary Issues in Technology and Teacher Education*, *8*(4), 310–325.

Williamson, D. M., Xi, X., & Breyer, F. J. (2012). A framework for evaluation and use of automated scoring. *Educational Measurement: Issues and Practice*, *31*(1), 2–13.

Zhang, M. (2012). *Sampling issues and error control in automated scoring of essays* (Unpublished doctoral dissertation). Washington State University, Pullman, WA.

Zhang, M., Breyer, F. J., & Lorenz, F. (in press). *Investigating the suitability of implementing the e-rater automated essay scoring system in a large-scale English language testing program* (Research Report). Princeton, NJ: Educational Testing Service.

Zhang, M., Williamson, D. M., Breyer, F. J., & Trapani, C. (2012). Comparison of e-rater scoring model calibration methods based on distributional targets. *International Journal of Testing*, *12*, 345–364.

Notes

[1] In some writing assessments, one or two essay prompts may be combined with other forms of writing evaluation, as in the case of *CBAL*™ writing assessment (Deane, 2011) and *TOEIC*® writing assessment (Zhang, Breyer, & Lorenz, in press). A composite writing score may be reported. However, because the number of essay prompts usually is fairly small (e.g., two for CBAL and one for TOEIC), the performance of an individual item can still have a large impact on the total writing assessment.

[2] Data were provided by the *Graduate Record Examinations*® program for use in a larger study reported in Zhang (2012). While this research was being conducted, the *GRE*® program launched the GRE revised General Test in August 2011. The former GRE General Test was studied here.

[3] Essays were processed by e-rater Engine-10.

[4] Because e-rater made unnecessary the routine use of the second human rater, the current data set only has one human rating for most responses.

[5] The current data set can offer an estimate of human agreement for that nonrandom subset of essays that must be adjudicated by a second human, due to a large discrepancy between the first human and e-rater. In addition, for this data set, the correlation between the human scores on two prompts (issue and argument) was examined. Finally, historical data (October 2006–September 2007) with two randomly selected human raters grading the same prompts as used in this study can offer an unbiased, but older, estimate of the interhuman agreement.

[6] There were originally more than 80 test countries/territories, of which many countries had quite small test-taker populations, accounting for less than 0.001% of the population. For example, there was only one examinee who took the test in Tunisia and three examinees who took the test in Yemen. Therefore, countries/territories that had fewer than 200 test-takers were combined into one group, or stratum, for stratification sampling purpose.

[7] For practical purposes, I considered the test-taker population on either a single (for prompt-specific scoring) or a group of prompts (for generic scoring) as the empirical prompt population.

[8] The use of different cross-validation samples seems defensible for two additional reasons. One, any change in operational practice from one sampling approach to another would seem to bring with it a difference in the cross-validation data set. Therefore, it is logical to include this associated variation in the comparison of sampling approaches. Two, adding, or alternatively holding out, a single fixed cross-validation data set would introduce the complication of what sampling approach to use in selecting that validation data set. Seemingly, a fair comparison would require a range of sampling approaches for cross-validation that, in combination with the range of approaches examined for model calibration, would greatly increase the complexity of the study's design and the interpretation of results.

[9] Operational automated scores were generated by the GRE program.

[10] A point to stress is that the lack of invariance in human ratings does not make a similar lack of invariance in automated scores acceptable; in both cases, the meaning of scores from one group to the next is not constant.