ED 422 724	FL 025 411
AUTHOR	Liu, Angie H. C.
TITLE	Constructing and Validating Parallel Forms of Performance-Based Writing Tasks in Academic Settings.
PUB DATE	1997-00-00
NOTE	32p.; Paper presented at the Annual Meeting of the Language Testing Research Colloquium (19th, Orlando, FL, 1997).
PUB TYPE	Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	College Students; Comparative Analysis; *English (Second Language); Higher Education; *Language Tests; Statistical Analysis; *Test Construction; Test Reliability; Test Validity; *Writing Evaluation
IDENTIFIERS	*Placement Tests

ABSTRACT

Due to concern for test security and fairness, three new performance-based English placement test writing prompts were developed in a large midwestern university for incoming students of English as a second language, and the degree of prompt variability was investigated from multiple perspectives, including "fit-to-specification," decision reproducibility, skill profile consistency, and prompt information similarity. Additionally, potential differential prompt functioning was examined for subgroups, including ESL examinees' gender, academic status, and field of study. The study of idiosyncratic response patterns was also conducted using outlier analysis. Multi-faceted Rasch models were selected to analyze examinees' essay performance because of their capacity to untangle the complex interaction between observed ratings and the unobserved latent trait. Findings are reported in four sections: prompt comparability in the target population; potential differential prompt functioning in various student groups; idiosyncratic response patterns of examinees and raters; and lecture-audience interaction impact of prompt comparability. Implications for future development of parallel forms of performance-based writing tasks in academic settings and the use of multi-faceted Rasch models in validating prompt comparability are discussed. (Contains 15 references.) (Author/MSE)

Constructing and Validating Parallel Forms of Performance-based Writing Tasks in Academic Settings

Angie H.C. Liu
Texas Education Agency

Abstract: *Due to the concern of test security and fairness, three new performance-based English Placement Test writing prompts were developed in a large university in the mid-west for its incoming ESL students and the degree of prompt comparability was investigated from multiple perspectives, including "fit-to-specification", decision reproducibility, skill profile consistency, and prompt information similarity. Additionally, potential differential prompt functioning was examined at the subgroup level in terms of ESL examinees' gender, academic status, and field of study. The study of idiosyncratic response patterns was also conducted using outlier analysis. Multi-faceted Rasch models were selected to analyze examinees' essay performances because its capability to untangle the complex interaction between observed ratings and the unobserved latent trait. Findings of this study were presented in four sections: a) prompt comparability in the target population, b) potential differential prompt functioning in various student groups, c) idiosyncratic response patterns of examinees and raters, and d) "lecture-audience interaction" impact on prompt comparability. Moreover, implications were discussed in terms of future development of parallel forms of performance-based writing tasks in academic settings, as well as the use of multi-faceted Rasch models in validating prompt comparability.*

Performance-based Second Language Writing Assessment in Academic Settings

The assessment of second language (L2) writing ability has been receiving increased attention in the past decade because important decisions, such as those regarding admission, placement, diagnosis, license, and achievement, are often made on the basis of the results of a writing assessment (Brand, 1991; DeMauro, 1992; Hamp-Lyons, 1990; Bode, 1994). In more and more testing situations, examinees are being asked to demonstrate their writing ability by producing writing samples to be evaluated. Because the ratings of examinees' writing samples are the joint product of measurement instruments (i.e., writing prompts, scoring rubrics and procedures, as well as essay raters) and examinees' writing ability, the reliability and validity of the use of performance-based essay ratings in making crucial decisions about examinees has always been a principal concern for test users and L2 writing assessment researchers. Moreover, due to the crucial role of the testing context in defining the criteria of "good / successful" writing and in determining content and construct validity of writing assessment, the validity and reliability of performance-based writing tests in any academic setting is subject to empirical evaluation. In other words, the results of writing performance under one academic setting cannot be considered automatically transferable to a different academic setting since, as suggested by Messick (1995), the comparability and generalizability of scores obtained across tasks goes to the very heart of score meaning. The extent to which score meaning holds across tasks, settings or contexts is an important empirical question.

When a writing test is administered more than once, multiple forms are often necessary due to practical considerations such as test security and test fairness (Petersen, Kolen, & Hoover, 1993). To ensure the interchangeability of different test forms and to maintain consistent and accurate judgments regarding ESL learners' writing ability across different tasks, it is generally considered essential that ESL test respondents be given prompts of comparable difficulty (i.e., parallel test forms). Linn (1995) contends that the issue of comparability poses major challenges for the study of performance-based assessment because of the complexity involved. However, most studies examining the generalizability of performance across tasks have not focused on the construction of comparable performance measures, which are intended to be used interchangeably. In other words, on the one hand, equivalent forms of performance-based writing tests need to be developed to maintain test security and fairness. On the other hand, when ESL writers respond to different test prompts, the influence of the fluctuations caused by different writing prompts on ESL writers' performance needs to be taken into account and appropriately adjusted to ensure the validity of score comparability across different writing prompts.

The meaningfulness of L2 writing assessment depends heavily on the degree to which L2 writers' compositions are representative of, or generalizable to, a stable trait, namely ESL examinee's writing ability (Lehmann, 1993); therefore, clear specification of the writing construct as well as the task domain is crucial in making valid generalizations or comparisons between examinees' writing performances on one writing prompt and those on other prompts. Consequently, to ensure that the writing performance elicited by prompts of "parallel" forms come from the same ability domain, a criterion-referenced framework of writing task specifications is needed to serve both as an analytic guide in constructing new "parallel" forms of writing prompts and but also an evaluating guide in validating the comparability of different forms of the writing prompts.

The English Placement Test at the University of Illinois at Urbana-Champaign

Based on the university policy, all international students admitted to UIUC are required to take the English Placement Test (EPT) unless their TOEFL score is above both the university's minimum (the current campus-wide minimum is 607) and the departmental minimum. The EPT is a performance-based test, locally developed at UIUC to determine whether students who are non-native speakers of English (NNS) need additional ESL instruction and, if so, to provide a basis for placing students into the appropriate level of ESL courses. The current EPT consists of two components, a written test and an oral interview.

The written component of the current EPT has two sections: a video-and-reading-based essay (used as a primary criterion for final placement decisions), and a multi-item discrete point English structure and usage exam (used as a measure of knowledge of basic grammatical conventions). The prompt for the essay writing was developed based on the *EPT Specification for Video-Reading Based Academic Essays*. The essay is scored holistically on a 1 to 4 scale by experienced teaching assistants of the ESL service courses. The structure test items are scored dichotomously and then converted to a standardized score. Unless exempted from any additional ESL instruction, examinees are placed into different levels of ESL service courses based on the composite results of the two sections (i.e., use essay rating as the basis for placement decision when the converted structure score is higher than 35; when the converted structure score is lower than 35, the assigned essay rating is lowered one level and that is used as the basis for the final placement decision).

The video-and-reading-based essay is designed to measure ESL examinees' abilities in successfully accomplishing academic writing tasks, where integrative skills -- understanding a professor's lecture, comprehending an academic reading text, and presenting both sources of information in a written format that is acceptable in academic settings -- are involved. Specifically, ESL examinees first watch a ten- minute-videotape in which a lecture corresponding to the theme of the later writing task is shown. Next, they are provided with a reading passage which addresses the same theme as the video-lecture, but presents opposing viewpoints. The ability to incorporate different viewpoints across sources and present them under the same thematic topic is one of the key components stated in the rating rubrics of the EPT test specifications. Finally, they are instructed to develop a main idea about the theme, support that idea with information from the videotape and the article, and present it in a general writing format (i.e. introduction, body, and conclusion). Examinees are given 50 minutes to read the article and to write the essay after the video is stopped. In addition, the criteria used in grading the examinee's essay are explicitly stated in the instruction booklet.

Purpose and Rationale of This Project

Currently, there is only one video-and-reading-based prompt developed for the EPT and it has been used for several years now. A large number of international students have been exposed to the current prompt and it is reasonable to suspect some examinees may already have some knowledge about the prompt when they take the EPT. Out of concerns of test security and fairness, the construction of a prompt bank with multiple forms of writing prompts is needed. To ensure the comparability of the measure of examinees' writing performances across different administrations, "parallel" prompts that can be used interchangeably are required. Constructing parallel forms of a performance-based writing assessment is an extremely complicated issue because of intervening factors (e.g. raters, scoring procedures, scale structure), as well as the limited writing samples collected from each examinee in real testing contexts. Moreover, due to the diverse backgrounds of international students, the existence of potential differential prompt functioning in different subgroups (e.g. graduates versus undergraduates; science / engineering major versus business major versus liberal arts major; male versus female) also needs to be investigated.

The intended goals of this project are:

(a) construct parallel forms of the video-reading-based writing prompts on the basis of the current version of the *EPT Specification for Video-Reading Based Academic Essays.*

(b) construct a non-parallel form of the video-reading-based writing by adding additional interaction between the lecturer and audience.

(c) empirically validate the degree of score comparability across the new forms of writing prompts in the EPT context.

(d) investigate whether the degree of parallelism of the alternate forms of video-reading-based writing prompts is the same across subgroups of ESL writers in terms of their major (science / business / liberal arts), gender, and academic status (undergraduates / graduates) in the EPT setting.

(e) identify factors or facets which contribute to the degree of parallelism among different forms of video-reading-based writing prompts.

(f) provide test developers with empirical information for future development of criterion-referenced-based test specifications as well as for the future construction of parallel forms of performance-based writing prompts in academic settings.

Defining Parallelism

Theoretical Perspective

The fundamental concept involved in this project resides in the notion of the "parallelism" of writing prompts. Thus, it is essential to explicitly define "parallel" in the current context. Conceptually, "parallel" prompts can be viewed as any prompts that elicit the same skill or the same composite of multiple skills. Statistically, parallel prompts can be defined as prompts which will result in the same expected score for an individual examinee or the same score distribution for a group of examinees. Under the framework of Item Response Theory (IRT), if different writing prompts reveal the same step characteristics curves for ordered response categories (in the case of unidimensionality) or the same prompt characteristic surface or space (in the case of multidimensionality), they are considered to be "parallel" prompts. Another approach under IRT is based on the comparison of prompt information. If two prompts display the same information function curve for the same group of examinees, they are viewed as "parallel" prompts.

Operational Perspective

In the present context of the EPT, "parallelism" among writing prompts is defined at three levels -- the test specification level, the decision level and the skill profile level. At the test specification level, if a group of ESL and testing experts agree that both writing prompts are constructed in such a way that they completely follow the same criterion-referenced-based test specification, they are defined as parallel prompts. At the decision level, two writing prompts are considered parallel if they were developed based on the same criterion-referenced-based test specification, and examinees' responses to either of the two prompts will place them into the same level of ESL courses as does the other prompt (i.e., decision *reproducibility*). Parallelism at the skill profile level takes on a narrower definition. For prompts to be considered "parallel" at this level, they not only need to be developed on the basis of the same criterion-referenced-based test specification but also need to elicit the same profile of writing skills for ESL examinees (i.e., in terms of their writing strengths and weaknesses). In this project, all three levels of parallelism of writing prompts are investigated and used as empirical evidence for validating the comparability across new writing prompts in the EPT context. Moreover, the results of the comparison on prompt information curves and step characteristic curves across new writing prompts are used as additional evidence in determining the degree of prompt parallelism.

Research Questions

The principal research question of this project is defined as follows:

> Given the current purpose of the ESL Placement Test (EPT) (i.e., whether new international students need more ESL instruction, and if so, what level would be appropriate) and its specific essay test format (i.e., video-reading based prompt), to what extent are test results of different writing prompts constructed on the criterion-referenced-based test specification comparable to one another?

Due to the complexity involved, the principal question is investigated from four aspects, namely, overall prompt comparability, potential differential facet functioning in subpopulations, idiosyncratic responses or rating patterns, as well as the impact of the hypothesized factor, "lecturer-audience interaction", on overall prompt comparability. Under each aspect, specific sub-questions were examined to provide empirical information to answer the principal research question and described as follows:

Regarding overall prompt comparability:

(1) Based on prompt evaluators, to what extent are the three new writing prompts parallel to one another in terms of their connection with the *EPT Specification for Video-Reading Academic Essays*? And, what do prompt evaluators perceive to affect the connection?

(2) To what degree are the three writing prompts parallel to one another in terms of decision *reproducibility* (i.e. consistency in placing examinees into the same levels of ESL classes)?

(3) To what degree are the three writing prompts parallel to one another in terms of writing skill profiles?

(4) How well are the three writing prompts parallel to one another in terms of prompt information curves and step characteristic curves?

Regarding potential differential facet functioning in student groups:

(5) Do the three writing prompts display differential functioning across subgroups of ESL examinees in terms of their academic status (i.e., graduates vs. undergraduates), major (i.e., science / engineering vs. business vs. liberal arts majors) and gender (male vs. female)?

Regarding idiosyncratic behvior patterns (i.e., outlier analysis):

(6) Are there examinees or raters displaying aberrant behaviors from the majority? If so, who are they ?

Regarding the factor which is hypothesized to affect prompt comparability:

(7) Is the hypothesis that "lecturer-audience" interaction would affect the degree of "parallelism" across different writing prompts being supported by empirical evidence?

METHODS

Subjects

Prompt Evaluators

Four "experts" specializing in second language learning and teaching and in language testing participated in this project to evaluate the degree of congruence between the updated version of the *EPT Specifications for Video-Reading-based Academic Essays* and the three newly developed video-and-reading-based writing prompts. The tasks of the prompt evaluators were: (a) reading through the EPT test specification for constructing parallel video-and-reading-based academic essays, (b) watching all three of the video-lectures and read three academic-like texts related to the topics of the video-lectures, (c) assessing whether the three video-and-reading-based prompts follow the test specification they read and evaluating the degree of prompt comparability on the basis of their connection with the *EPT Specification for Video-Reading Academic Essays*, (d) identifying specific factor(s) contributing to their perception of prompt comparability and explaining how they arrived at their judgment, and (e) providing oral feedback for further editing of the reading passages after being given the information regarding how those prompts were deliberately constructed and what the intended goals are. The whole discussion process between prompt evaluators was recorded on a cassette tape.

Essay Writers / ESL Examinees

One hundred international students who took the EPT on August 27th or August 29th, 1996, were randomly sampled from each test date and served as the subjects for this project. All 15 undergraduate students who signed up for the EPT on August 27th, 1996, were selected while 85 graduate students were randomly sampled from the graduate examinee population. On August 29, 30 undergraduates and 70 graduates were randomly selected so as to maintain the regular ratio between the graduate and undergraduate students. Of the 200 students selected, 86 were women and 114 were men. Additionally, 75 of them were science or engineering major, 63 of them were business major, and 62 of them were liberal arts major. Each examinee responded to two video-reading-based writing prompts (i.e., one anchor prompt plus a second prompt) and produced one essay for each writing prompt.

Essay Raters

Eight raters participated in this project. All raters were both experienced EPT raters and experienced teachers of ESL service courses. Thus, the raters all had clear understandings of the rating rubrics, the course objectives in the level of ESL courses they have taught, as well as the level of students that were most likely to benefit from their classes. Except for one rater, all the raters were native English speakers. The ESL teaching experience of the raters ranged from one to eleven years and from elementary school to university level, including both private institutes and public schools.

Instruments / Materials

Essay Prompt Specification

The current version of the *EPT Specification for Video-Reading Academic Essays* was used as the basis to develop new forms of the EPT writing prompts. Specifically, it consists of five key components, general description (GD), sample item (SI), prompt attributes (PA), response attributes (RA), and supplementary specification (SS). Under the general description component, the abilities or skills being measured by the EPT essay test are explicitly specified. Under the sample item component, the information regarding the procedures and instructions of the test are given by means of a real example. Under the prompt attributes component, the required and desired characteristics of the stimuli (i.e., video lecture and reading passage) such as the length of the video-lecture and reading passage, the general difficulty level of the information presented, or the relation between video-lecture and reading passage, are specified. Under the response attributes component, descriptions of the expected writing response and format, as well as scoring procedure are stated in detail. Finally, under the supplementary specification component, additional comments or specifications that do not fit in any of the previous categories are specified here

Essay Prompts

Each EPT writing prompt consists of three components, a video-lecture by a university professor, an academic-like reading text, and a test booklet with written instructions and space for note-taking and an essay response. In this project, three new performance-based writing prompts were developed. Two writing prompts were constructed in such a way that they completely follow the EPT test specification and that were expected to be used as alternate forms in eliciting ESL examinees' writing performance. The third prompt; however, was developed in such a way that an additional "question-and-answer" exchange between the lecturing professor and one of the student audience was added to the video-lecture component and that it was not expected to elicit parallel writing responses from examinees.

Rating Rubrics

In this project, both a holistic rating procedure and a componential rating procedure were used to score the ESL examinees' essays. The examinee's essays were first rated using the holistic rating procedure, which has always been used for the operational EPTs. Each essay was rated independently by two raters and assigned one holistic score by each rater. As a result, each essay received two holistic ratings. Later, componential rating procedure was used to evaluate examinees' essays in terms of four key features (i.e., organization, content, grammar, and style) of the essays. One score was assigned to each particular feature of the essays and since each essay was judged by two raters, each key feature received two ratings and thus, four key features resulted in eight ratings. In short, each essay received ten ratings in total, including two holistic ratings and eight componential ratings, which later were used to calibrate the examinee's writing ability.

Subject Background Information Sheets (Examinees and Raters)

Personal information data sheets were developed and given to both ESL examinees and raters who participated in this project. These individual profiles were later used to provide qualitative interpretations of the results of "outlier analyses", in which examinees or raters who display idiosyncratic behavior patterns were identified and labeled as "outliers". In addition, because the international students who participated this project came from diverse backgrounds, the validation of new writing prompts was extended to the subgroup level in this project. Their background information was used as the basis to classify them into different subgroups for investigating whether new writing prompts function differentially across subgroups of examinees (i.e. bias analysis).

Research Design

Intended-to-be-parallel vs. Not-intended-to-be-parallel Prompts

To investigate what factors contributed to the degree of comparability across performance-based writing prompts, one attribute of the video-lecture component of the writing prompt (i.e., lecturer-audience interaction) was hypothesized as a factor that may affect prompt comparability and is intentionally used to induce prompt variation. It was similar to setting up a type of control group.

Regarding the construction of new writing prompt, first, two writing prompts were constructed in such a way that they were intended to be parallel and thus, completely follow the current version of the *EPT Specification for Video-Reading Academic Essays.* The video-lectures of the two intended-to-be parallel prompts consisted only of a lecture presentation from the professors (i.e., no students were allowed to pose any question in the lectures). Next, the third prompt was constructed in such a way that it was not intended to be parallel to the other two prompts. Additional lecturer-audience interaction was included in the lecture section of the intended-to-be nonparallel prompt through one exchange of question and answer at the end of the lecture (i.e., one question was allowed to be raised by one of the student audience regarding the content of the lecture and the lecturing professor would responded to the question briefly).

Data Collection Procedures

The following procedures were used to collect necessary information for answering the research questions:

Step 1: Three U. of I. professors from different academic fields were invited to be the lecturing professors in the video-lecture of the new writing prompts.

Step 2: New performance-based writing prompts of selected topics (i.e., ethics, economics, brain specialization) were developed based on the *EPT Specification for Video-Reading Academic Essays*.

Step 3: Based on the actual content of the video-lecture, the reading articles that were initially provided by the professors or obtained from the library search were selected and revised to meet the requirement of the test specifications in terms of the length of the reading text, the difficulty level of the information presented, as well as the relation between the reading text and the video-lecture (i.e., they should be of same theme but different viewpoints).

Step 4: A pilot study was conducted in the Intensive English Institute (IEI) at the University of Illinois at Urbana-Champaign during the summer of 1996 to test the effects of the newly-constructed writing prompts on ESL students. Three advanced IEI classes of about 12 to 16 students participated in the pilot study. Since the purpose of extra question-and-answer interaction was to induce prompt variation, the writing prompt with the question that best facilitates students' comprehension of the corresponding lecture (based on students' feedback in the pilot study) was selected to be the not-intended-to-be prompt. As a result, the prompts of the ethics and the brain specialization topics were selected as the intended-to-be-parallel prompts (i.e., without "question and answer") and the prompt of the economics topic was determined as the "not-intended-to-be parallel" prompt.

Step 5: Prompt evaluators, who were experts in second language teaching and assessment, were asked to evaluate the degree of parallelism between *the EPT Specification for Video-Reading Academic Essays* and the three selected video-and-reading-based writing prompts.

Step 6: The finalized writing prompt sets were administered to the international students who were required to take the EPT in fall 1996. Since the first two EPT administrations in the fall semester usually cover the majority of new international students who are required to take the test, those who signed up for the first two administrations (i.e., August 27th and August 29th) were used as the subjects (essay writers) in this project. Specifically, each ESL examinee was required to respond to two writing prompts (first, the anchor prompt and then, the other selected prompt) and thus, produced two essays. The design is presented graphically in Figure 1:

Group 1 (Aug. 27th)	Group 2 (Aug. 29th)
Ethics prompt (without-question-version)	Ethics prompt (without-question-version)
Multi-item structure test	Multi-item structure test
Economics prompt (with-question-version)	Brain specialization prompt (without-question-version)

Figure 1. The EPT administration plan

Step 7: Immediately after the test, examinees' essays on the anchor prompt (i.e., the ethics prompt) were first scored using the holistic rating rubrics. Each essay was evaluated by two experienced raters.

Step 8: After the operational scoring of the EPT was completed, the essays of the non-anchored prompts (i.e., the economics prompt and the brain specialization prompt) were then scored by experienced raters using the same holistic rating rubrics. Each essay was again rated by two raters on a four-point scale.

Step 9: After the holistic ratings were completed, examinees' essays on all three writing prompts were then scored using the componential rating rubrics. However, to avoid the "interaction effect", the holistic scoring and the componential scoring on the same essay were performed by different raters. That is, it was controlled so that one rater would not assign both componential and holistic ratings to the same essay. Each essay was componentially rated on a five-point scale independently by two experienced raters.

Data Analysis Strategies

The first research question (i.e., Based on expert judgment, to what extent are the three new writing prompts parallel to one another in terms of their connection with the EPT test specification?) was analyzed by means of experts' judgments on the connection between the new writing prompts and the *EPT Specification for Video-Reading Academic Essays* (i.e., the degree of fit-to-specification).

The second research question (i.e., How well are the three writing prompts parallel to one another in terms of decision reproducibility?) was examined by comparing the placement decisions made across the three writing prompts. First, a multi-faceted Rasch model was used to calibrate the examinee's writing ability because it can accommodate the complicated relation between the unobserved latent trait (i.e., the writing ability) and the observed scores (i.e., raw ratings) in performance-based assessment by taking the "intervening" factors (e.g., prompt difficulty, rater severity, scale structure) into account when estimating the ability trait. Consequently, the variability caused by the intervening factors can be controlled and consistent ability estimates can be obtained for the same individuals regardless of the specific raters or tasks involved in a test. Specifically, a four-faceted Rasch model was adopted to calibrate the ESL examinees' writing ability, analyze the variability caused by rater severity, prompt difficulty, and aspect difficulty (i.e. organization, content, grammar, and style). The model is formally expressed as:

$$\log (P_{nhijk} / P_{nhijk-1}) = B_n - D_i - C_j - E_h - F_{ijhk} \qquad \text{(Equation 1)}$$

where

P_{nijhk} = probability of examinee n, on aspect h, being rated k on prompt i by judge j

$P_{nijhk-1}$ = probability of examinee n, on aspect h, being rated k-1 on prompt i by judge j

B_n = the ability of examinee n

D_i = the difficulty of prompt i

C_j = the severity of judge j

E_h = the difficulty of aspect h (e.g. organization, content, style)

F_{ijhk} = the difficulty of the step up from category k-1 to category k for each prompt i/judge j/aspect h combination.
(*note: this step difficulty parameter is not labeled as a facet)

The model states that the probability of examinee n receiving a score category k rather than category k-1 is a function of the difference between the examinee's latent ability and the difficulty of the writing prompt after adjusting for the variability caused by rater severity, aspect difficulty, and scale structure. After the examinee's writing ability was calibrated, they were then placed into different levels of ESL service courses based on their calibrated scores (in logits) against the cut-scores, which were determined by converting the cut-scores in raw score units into the logit scale. Finally, the degree of decision reproducibility was determined by comparing the placement distributions across the three writing prompts. The more similar the placement distributions, the higher the degree of decision consistency achieved.

The third research question (i.e., How well are the three writing prompts parallel to one another in terms of writing skill profiles?) was studied by analyzing examinees' skill profiles generated by the multiple-trait scoring procedure (i.e., componential rating) in terms of aspect or subskill difficulty. Again, a four-faceted Rasch model was used to calibrate the examinee's writing ability, prompt difficulty, rater severity, aspect difficulty, and scale structure. After the estimates of the facet parameters were obtained, the three writing prompts were compared in terms of their calibrated aspect difficulty. Since the elicited profile of the examinee's writing strengths and weaknesses is the result of the interactive functioning between the examinee's ability in different aspects (i.e., organization, content, grammar, and style) and the difficulty levels of those aspects, comparing the calibrated aspect difficulties across the three writing prompts is the same as comparing the examinee's writing profiles elicited by the three writing prompts.

The fourth research question (i.e., To what degree are the three writing prompts parallel to one another in terms of prompt information curves and step characteristic curves?) was studied by comparing the prompt's information curves and step characteristics across the three writing prompts. The same four-faceted Rasch model was used to calibrate all the parameters in the specified measurement model (i.e., the model statement in Equation 1). On the basis of the calibrated step difficulties for each "item" of the writing prompt (note that in this study, each component rated such as organization, content, grammar, style is viewed as an "item" and the discrimination parameter is treated as equal across items in the Rasch model), step characteristic curves and information curves for each "item" were first constructed by graphic computer software. Next, information for the entire writing prompt was computed by summing up the information previously obtained for individual "items" and finally, the prompt information curves of the three writing prompts were constructed. By comparing the overall shape of the information curves across the three writing prompts, the answer to this research question was obtained.

The answer to the fifth research question (i.e., Do the three writing prompts display differential functioning across subgroups of ESL examinees in terms of their academic status, major, and gender?) was investigated by means of the interaction analysis in the multi-faceted-Rasch-model (i.e., the bias analysis in the FACETS program). In the framework of multi-faceted Rasch measurement, the study of differential prompt functioning is like the study of interaction effects between various facets specified in the multi-faceted model. The investigation of potential differential facet functioning can be achieved by including the suspected facet in the measurement model, and then examining the relationship between the suspect facet and other facets in the measurement model to see whether significant interactions exist between them. In the specific measurement model for bias analysis, the examinee facet in the original measurement model (i.e., Equation 1) was replaced by a subgroup facet to investigating potential differential prompt functioning due to subgroup effects. The measurement model used is formally expressed as follows:

$$\log (P_{rhijk} / P_{rhijk-1}) = G_r - D_i - C_j - E_h - F_{ijhk} \qquad \text{(Equation 2)}$$

where

P_{nijhk} = probability of group r, on aspect h, being rated k on prompt i by judge j

$P_{nijhk-1}$ = probability of group r, on aspect h, being rated k-1 on prompt i by judge j

G_r = the ability of subgroups defined (e.g., male vs. female)

D_i = the difficulty of prompt i

C_j = the severity of judge j

E_h = the difficulty of aspect h (e.g. organization, content, style)

F_{ijhk} = the difficulty of the step up from category k-1 to category k for each prompt i/judge j/aspect h combination.
(*note: this step difficulty parameter is not labeled as a facet)

After all the parameters are calibrated and the interaction between facets are checked, if significant interactions are found between the subgroup facet and other facets in the model, the bias analysis in the multi-faceted Rasch model then calculates a bias measure based on the data involved in the elements of the facets where the significant interaction occurs. The size of the estimated bias is reported in logits, and its significance is reported in terms of mean squared residuals (i.e., Infit and Outfit statistics) and a standard z-score (Linacre, 1996). And, finally the procedures of parameter calibration and the check on between-facet interaction were repeatedly for each classification of subgroups (i.e., graduate vs. undergraduate; science vs. business vs. liberal arts majors; male vs. female) and the results of potential differential prompt functioning due to examinees' subgroups were acquired.

The sixth research question (i.e., Are there examinees or raters displaying aberrant behaviors from the majority?) was examined by means of the outlier analysis in the multi-faceted Rasch model. In the multi-faceted measurement, the identification of examinees and raters who behave differently from the expected majority (i.e., outliers) is determined based on the person fit statistics (i.e., local model-to-data fit in terms of person parameters). Therefore, to determine whether there are individual examinees or raters whose response patterns are different from the majority, the observed essay ratings were fit by a four-faceted Rasch model (i.e., Equation 1) to obtain the estimates of parameters in the model, as well as the local model-to-data fit, including both item-fit statistics and person-fit statistics. Here, the focus was on person-fit statistics, which is an index for evaluating the extent to which an individual's behavior pattern corresponds to the predicted overall behavior pattern for the individual. Significant misfit suggests that an individual's behavior pattern does not correspond to the individual's expected behavior pattern such that an individual's performance on one item can not be predicted from the individual's performance on other items (McNamara, 1996). Person-fit statistics are computed based on the difference between the modeled probability and the actual observed value for an individual's performance (i.e., residuals). If an unexpectedly large residual is found for a particular individual, that individual is identified as a misfitting person (Linacre, 1989). Depending on measurement purpose and the use of test results, different range of acceptable person-fit statistics can be established by individual researchers.

Finally, the answer to the last research question (i.e., Is the hypothesis that lecturer-audience interaction would affect the degree of parallelism among the three writing prompts being supported by empirical evidence?) was based on the findings and conclusions of the first four research questions. If the two intended-to-be parallel writing prompts are indeed found to be parallel to each other in terms of "fit-to-specification" agreement, decision reproducibility, skill profile consistency, as well as prompt information curves and if the degree of comparability between the "not-intended-to-be-parallel" writing prompt and the other two prompts is found to be lower, then it can be concluded that the empirical evidence does support the hypothesis that lecturer-audience interaction affects the comparability of the three prompts. But if the results of the comparability of the three writing prompts do not meet the expectation (for example, all three writing prompts are found to be functioning similarly, despite the lecturer-audience interaction factor), then the hypothesized impact of lecturer-audience interaction is determined to be not supported by empirical evidence.

Data Analysis Results And Discussion

Due to the complexity involved in the principal research, the collected data were analyzed from four perspectives, namely prompt comparability in the target population, potential differential prompt functioning in student groups, idiosyncratic response patterns (i.e., outlier analysis), as well as the hypothesized "lecturer-audience interaction" impact on prompt comparability.

Comparability / Parallelism Across Writing Prompts

Content-Wise Comparability

Four prompt evaluators were asked to evaluate the degree of comparability of the three newly-developed writing prompts based on their connection with the *EPT Specification for Video-Reading Academic Essays*. Based on the definition of parallelism at the test specification level, if prompt evaluators agreed that the three writing prompts were constructed in such ways that they completely followed the EPT test specifications, the three prompts are validated as parallel forms. Regarding the ethics prompt, although all four evaluators agreed that the prompt completely follows the test specification, two evaluators raised concerns about the "culturally neutral" criterion specified in the test specification for different reasons. For example, one evaluator worried that the reference to Mussolini and Nazis in the video-lecture of the ethics prompt might affect some international examinees because they may have different views about these historical figures. Another evaluator had a reservation about the use of boycotts as the theme for the reading passage. He claimed that based on his experiences in teaching international students, perhaps not too many of them know enough about the kind of boycotts going on in the U.S. to understand why a boycott is even an issue. Additionally, one evaluator was concerned about the cognitive load of this writing task. He felt that the content of the video-lecture was abstract and conceptually-oriented. As a result, it would be a difficult task for ESL examinees to process the information presented both in video and in reading passage because they not only have to understand the language used but also have to perform the cognitive task. Another evaluator agreed that the arguments in the reading passage were difficult to follow but considered the information in the video-lecture to be general.

With regard to the economics prompt, at first all of the prompt evaluators except one found instances of violations to the test specification in regards to the use of technical terms and information presented in the video-lecture. Two evaluators claimed that the information level in the video-lecture violated the "not too technical" criterion in the test specification in that students may need some background knowledge to understand the economic symbols and graphs used by the lecturer. Nevertheless, they also agreed that the principal viewpoint in the video-lecture was clearly presented and the link between the video and the reading passage was very good so that once students grasp the main ideas, it should not be too difficult for them to write an opinion-type of essay. In addition, three of the four evaluators directly pointed out that the extra "question-and-answer" section at the end of the video-lecture violated the specification of "no lecturer-audience interaction." After this was pointed out by other evaluators, the evaluator who initially did not notice the violation also agreed that it was an obvious discrepancy between the prompt and the test specification. However, he still maintained his position in terms of the technical level of this prompt that -- some examinees may not be familiar with some of the technical terms or information presented by the lecturer. Nevertheless, because the definitions of those technical terms were shown on the screen long enough, he felt that examinees could simply copy them down. Moreover, because a lot of symbols and graphs were used, one evaluator reported that he thought those would facilitate examinees' processing of "specialized" knowledge.

As far as the brain specialization prompt is concerned, all four prompt evaluators agreed that the degree of fit-to-spec. was quite good and the delivery of the lecture was very clear and understandable. They also liked the fact that the camera focus was always on the lecturer. However, one evaluator pointed out that there was a difference in the use of visual aids between this prompt and the other two prompts. In this prompt, only single words / terms, but not the whole definitions, were put down on the paper and shown on the screen. As a result, students could not simply copy down the definitions of terms; they had to acquire this information through *listening* to the lecture in real time, which may increase the difficulty level of the task. Another evaluator; however, contended that students do not need special definitions to understand those terms. In fact, the definitions are embedded in the context. Students should be able to follow the main gist of the lecture even without complete definitions of those terms. Nonetheless, both of them commented that the particular way of handling visual aids does not violate any criterion in the test specification. In addition, all four evaluators agreed that the link between the reading passage and the video-lecture was rather clear and most examinees should be able to grasp the opposing viewpoints portrayed in the two sources of information and successfully complete the writing task.

Summing up the prompt evaluators' opinions, the answer for the first research question (i.e., Based on expert judgment, to what extent are the three new writing prompts parallel to one another in terms of their connection with the EPT test specification?) would be -- at the test specification level, although the degree of "fit-to-specification" is generally high for all three writing prompts, the degree of prompt comparability is a little bit higher between the ethics prompt and the brain specialization prompt than between the economics prompt and either of the two prompts.

Decision-Wise Comparability

Based on the calibration results, it was found that the three writing prompts of interest indeed differed slightly in their prompt difficulty level (see Table 1) such that the ethics prompt imposed the least challenge on ESL examinees (difficulty = -.29 logit), followed by the economics prompt (difficulty = .08 logit), and the brain specialization prompt was the most difficult writing task (difficulty = .37 logit). However, generally, the differences in calibrated prompt difficulties for the three writing prompts were found to be only minor, especially after taking the standard error of measurement (i.e., .04 logit, .06 logit, .06 logit) into account, because the differences in prompt difficulties would not result in differences in observed ratings under normal circumstances. In other words, since each level in raw rating scale covers at least six units in the logit scale; the differences in prompt difficulty across the three prompts were considered minor and lack of practical significance.

Table 1
Calibrated Prompt Difficulty in Logit Scale

Prompt	Measure	Model S.E.	Infit MnSq	Outfit MnSq
	(less difficult)			
Ethics	−.29	.04	1.1	1.1
Economics	−.08	.06	1.1	1.0
Brain	.37	.06	0.9	0.9
	(more difficult)			
Mean	.00	.05	1.0	1.0
S.D.	.27	.01	0.1	0.1

The fit statistics (i.e., Infit and Outfit) in terms of mean square residuals provide information on how well the observed ratings of individual prompt fit the four-faceted Rasch model. Infit statistics focus on the degree of fit of the most typical observations in the data matrix while outfit statistics evaluate the model-to-data fit based on all observations (i.e., including outliers). Mean square values have an expected value of 1, and when the observed values show greater variation than the measurement model expects, mean square values will be larger than 1 and when less variation than expected, mean square values less than 1 will be found. A useful rule of thumb for acceptable mean square values ranges approximately from 0.75 to 1.3 (McNamara, 1996). More precisely, for n size of 30 or more, the range is the mean ± twice the standard deviation of the mean square statistic. Based on this criterion, in this case, values greater than 1.2 show significant misfit and values below 0.8 show significant overfit. Apparently, all three prompts are found to fit the specified measurement model reasonably well.

In addition, based on the results of calibrated rater severity, raters, overall, were found to differ in their overall severity in essay-rating. Specifically, it was found that rater 5 (severity= -1.95 logit) and rater 4 (severity = -1.19 logit) adopted the most lenient standards in rating examinees' essays while rater 8 (severity = 1.49 logit) and rater 1 (severity = 1.20 logit) judged the essays the most harshly.

Moreover, regarding the overall difficulty of different subcomponents or aspects of essays (i.e., organization, content, grammar, and style) across the three writing prompts, it was found that generally the overall aspect (i.e., result of holistic scoring) imposed the greatest challenges (difficulty = .64 logit) on ESL examinees, followed by the content aspect (difficulty =

.13 logit) and then the organization aspect (difficulty = .07 logit). At the same time, the grammar aspect (difficulty = -.57 logit) was revealed to be the easiest subskill to be mastered by ESL examinees, followed by the style aspect (difficulty = -.27 logit). Nonetheless, the differences found among different writing aspects were overall minor. Additionally, the fit statistics, including both infit and outfit statistics, of the calibrated aspect difficulty showed that all writing aspects fit the measurement model quite well in that the values are either equal to or very close to the expected 1.0.

Furthermore, due to a number of intervening variables (e.g., rater severity, task difficulty, domain difficulty, and scale structure) between the underlying latent trait, and the observed ratings, it is particularly important to ensure that proper adjustments are made for the differences caused by unintended factors before examinees' estimated ability measures are used as a valid and reliable source for decision-making (e.g., placement, certification, promotion, etc.). When the same decisions are reached regardless of the particular combination of raters and tasks involved in the test, the decisions are viewed as "reproducible" (Lunz, Stahl & Wright, 1994). In this project, the parallelism defined at the decision level was based on the same idea of "decision reproducibility" -- that is, when ESL examinees' responses to any of the writing prompts placed them into the same level of ESL courses, the writing prompts were considered as parallel / comparable. The essential element to the "decision reproducibility" is reliable / reproducible ability estimates, which can be calibrated from observed ratings after the variabilities of observed ratings due to unwanted factors such as rater severity, prompt difficulty, scale structure are statistically controlled. The FACETS program was used in this project to obtain the desired reproducible estimates of ESL examinees' writing ability measures.

Since the FACETS program reports ability estimates and all other calibrated parameters on a logit scale and there are negative values in the logit scale, it is not a convenient nor descriable scale for reporting test results to students, advisors, and school administrators. In addition, the current scale for reporting essay test results, which matches students' ability levels with different ESL courses offered by the Division of English as an International Language, has been used for years. Therefore, instead of reporting the logit-unit score, a converted table was constructed to transform the ability measures in logit scale to the corresponding level scale. Presently, essay test results of the EPT are reported on a one-to-four scale, with level 1 as the lowest level and level 4 as the highest level. Since the required courses corresponding to different score levels differ for graduate students and undergraduate students, different sets of "cut scores" underlying the rating rubrics were used for the two groups. Based on their performance on the ethics prompt (i.e., the operational prompt), examinees' raw ratings were matched with their ability measures (estimated by the FACETS program) and the best matching points which resulted in the greatest consistency between the two measures were noted and used as converting cut-off points. For example, suppose that the great majority of graduate examinees who received rating 2 for the ethics prompt had ability measures in logit-unit in the range of -6.56 to -3.96 logits, then those two logit points become the matching points for the scale conversion table.

The complete sets of scale matching points for graduates and undergraduates are reported in Table 2. It is clearly shown that score scale structures underlying the observed ratings are different for graduate and undergraduate students. Overall, higher writing ability required of graduate students to receive the same level score as the undergraduates. Moreover, for graduate students, the range of the logit-unit score corresponding to level 2 is almost the same as that of the logit-unit score corresponding to level 3, but for undergraduates, the range of logit-unit score corresponding to level 2 is comparatively much narrower.

After the scale conversion table was established on the basis of the operational prompt (i.e., the ethics prompt), prompt comparability at the decision level can then be evaluated in the logit scale in terms of the number of examinees in each placement level across the three writing prompts of interest. However, because the total number of graduate and undergraduate examinees who took the EPT on different administration dates was different, direct comparison of placement consistency across the three writing prompts required the transformation of the number of examinees into the proportion of total examinees who responded the particular prompt. Table 3 shows the proportion of examinees in each placement level in terms of the three writing prompts they responded to. The proportion data shows that for the undergraduate group, the decisions based on examinees' performance on the ethics prompt and the brain specialization prompt

Table 2
Conversion Table for Level Scale and Logit Scale

	Logit Scale	
Level Scale	Undergraduates	Graduates
Level 1	< –5.21	< –6.56
Level 2	–5.21 to –3.96	–6.56 to .08
Level 3	–3.95 to 4.02	0.09 to 6.60
Level 4	> 4.02	> 6.60

Table 3
Proportion of Examinees in Each Placement Level

	Proportion of examinees in							
Prompt	level 1		level 2		level 3		level 4	
	UG	G	UG	G	UG	G	UG	G
Ethics	0	.019	.022	.341	.644	.574	.333	.064
Economics	.133	.094	0	.447	.600	.400	.267	.059
Brain	.100	.157	.033	.371	.567	.386	.300	.086
Mean	.078	.090	.018	.386	.604	.453	.329	.070
S.D.	.069	.069	.017	.055	.039	.105	.033	.014

generally come closer to each other than the economics prompt. The only large discrepancy occurred in placement level 3, in which higher "decision reproducibility" results from the economics and brain specialization prompts. Similarly, for the graduate group, it is the ethics and brain specialization prompts that tend to render comparatively similar decisions regarding the levels of ESL courses that examinees are required to take (if necessary) when compared with the decisions made on the basis of the economics prompt. The only exception occurs in placement level 3. In this level, it is found that examinees' performances on the ethics and brain specialization prompts result in comparatively more inconsistent decisions and the economics and brain specialization prompts result in more consistent decisions. Moreover, the standard deviation of each academic sequence in each placement can be computed and the average of the sum of the standard deviation is the overall averaged error rate, which indicates the degree of placement discrepancy across the three writing prompts generally. The result shows that the overall averaged error rate of placement decisions is around 5% with the highest decision consistency occuring in the highest placement level for graduate students (i.e., standard deviation = .014) and the highest decision inconsistency occuring in the placement level 3 for graduate students (i.e, standard deviation = .105).

The decision consistency reported above focused on the prompt comparability in terms of the whole examinee group. Regarding the decision consistency at the individual examinee level (i.e., Whether different prompts would place the individual examinee in the same placement level?), it was found that 76% of graduate examinees and 74% of undergraduate examinees who responded to the ethics prompt and the economics prompt would be placed in the same levels of ESL courses. When responded to the ethics prompt and the brain specialization prompt, it was found that 86% of graduate examinees and 77% of undergraduate examinees would be placed in the same levels of ESL courses. Therefore, it was concluded that the degree of prompt comparability between the ethics and the brain specialization prompt was a little bit higher than that between the economics prompt and the ethics prompt in terms of the placement consistency at the individual examinee level.

Summing up the findings for the second research question (i.e., How well are the three writing prompts parallel to one another in terms of decision reproducibility?) -- the overall degree of prompt comparability of the three writing prompts in terms of decision consistency is generally good, although comparatively, the overall degree of parallelism is slightly higher between the ethics prompt and the brain specialization prompt than the degree between the economics prompt and either of the other two prompts. Because of the minor decision inconsistency found in specific score levels, these three writing prompts cannot be considered as "strictly parallel" (i.e., resulting in total decision consistency). Nevertheless, the overall averaged error rate in terms of placement decisions (i.e., decision discrepancy) was found to be around 5%; therefore, if that is acceptable for decision-makers in the current context, these three writing prompts can still be considered as parallel forms at the decision level.

Skill-Wise Comparability

The analysis of this study treats different aspects of the essay scoring as test items. Since there are four aspects being evaluated in componential rating (i.e., organization, content, grammar and style) and one aspect in holistic rating (i.e., the holistic aspect), there are a total of five aspects / items for each writing prompt. By examining ESL examinees' writing performance in these specific aspects, profiles of ESL examinees' writing subskills in terms of strengths and weaknesses can be compiled. Therefore, evaluating prompt comparability at the skill can be approached by comparing the estimates of aspect difficulty across the three writing prompts.

In terms of the ethics prompt (see Table 4), the holistic aspect (difficulty = 2.38 logit) was found to be the most challenging aspect for ESL examinees, followed by the content aspect (difficulty = -.13 logit) , and the easiest aspect was found to be the grammar aspect (difficulty =-.94 logit). Similarly, in terms of the brain specialization prompt, it is found that the holistic aspect (difficulty = 1.12 logit) was the most difficult aspect, followed by the organization aspect (difficulty = .15 logit) and the content aspect (difficulty = -.10 logit), and the easiest aspect was found to be the grammar aspect (difficulty = -.81 logit). However, in terms of the economics prompt, it is the holistic aspect (difficulty = -.43 logit) that was found to be the easiest aspect, followed by the grammar aspect (difficulty = -.03 logit), and the content aspect (difficulty = .25 logit) was the aspect that imposed the greatest challenges for ESL examinees. The comparability among the three prompts in terms of their rank order in aspect difficulty was summarized in Table 5.

In addition, some variation of the calibrated aspect difficulty were found to exist across the three writing prompts in terms of the value of aspect difficulty estimates. In regard to the holistic aspect, the ethics prompt was reported as the most difficult one, followed by the brain specialization prompt, and the easiest prompt was the economics prompt. When organization aspect was the focus, the brain specialization prompt was found to be the most difficult one, followed by the economics prompt, and the easiest prompt was the ethics prompt. With respect to the content aspect, however, the economics prompt was found to be of the highest difficulty, followed by the brain specialization prompt and then, the ethics prompt. Similarly, as far as grammar is concerned, the economics prompt was found to be the one with highest difficulty, followed by the brain specialization prompt and then, the ethics prompt. In terms of style, the economics prompt, once more, was found to be the most difficult one and the easiest one was the ethics prompt. Furthermore, when aspect difficulty is compared across the three prompts, the overall aspect is found to be the aspect that results in the largest discrepancy in difficulty estimates. For example, the largest difference in all aspect difficulty estimates occurs between the ethics prompt (difficulty = 2.38 logit) and the economics prompt (difficulty = -.43 logit).

Table 4
Calibrated Aspect Difficulty for the Ethics Prompt

Aspect	Aspect Difficulty (logits)	Model S.E.	Infit MnSq	Outfit MnSq
	(less difficult)			
Grammar	–.94	.11	1.2	1.5
Style	–.87	.11	1.1	1.1
Organization	–.44	.10	1.0	1.0
Content	–.13	.10	1.0	1.0
Holistic	2.38	.14	0.9	0.9
	(more difficult)			
Mean	.00	.11	1.1	1.1
S.D.	1.23	.01	0.1	0.2

Table 5
Rank Order of Aspect Difficulty in the Three Writing Prompts

Prompts	Rank Order (* ">" represents more difficult)
Ethics	Holistic > Content > Organization > Style > Grammar
Economics	Content > Style > Organization > Grammar > Holistic
Brain	Holistic > Organization > Content > Style > Grammar

Summing up for the third research question (i.e., How well are the three writing prompts parallel to one another in terms of writing skill profiles?) -- it was found that the three writing prompts generally elicited similar writing skill profiles form ESL examinees because, except for the overall aspect, the three writing prompts of interest were found to be quite comparable to each other in terms of aspect difficulty, especially when measurement errors were taken into account. Specifically, the degree of parallelism between the ethics prompt and the brain specialization prompt at the writing profile consistency level was found to be slightly higher than either the degree of parallelism between the ethics prompt and the economics prompt or that between the economics prompt and the brain specialization prompt.

Step Characteristic Curves and Information Curves

The information curves for the three writing prompts constructed based on the sum of "item" / aspect information are presented in Figures 2-4 (see appendix). By comparing the information curves across the three writing prompts, it is clearly shown that the overall shapes of the information curves for the ethics prompt and the brain specialization prompt are remarkably similar in that both prompts elicit the most information from the lower ability group and the higher ability group with respect to ESL examinees' writing ability. Comparatively, both prompts also elicited the least information from the intermediate ability group in terms of ESL examinees' writing ability. However, when the information curve for the economics prompt is compared with either the ethics or brain specialization prompts, minor differences are found in the middle range of the latent ability scale in that the economics prompt elicited slightly more information from the intermediate ability group while a little less information from the lower or higher ability groups.

To further investigate the potential factors underlying the minor differences found in the information curves of the three writing prompts, additional information curves were constructed and compared on the basis of the componential ratings (i.e., by summing item information across the four componentially scored items) and the holistic ratings (i.e., by using the item information curve of the only holistically scored item). The results of comparing the three writing prompts' information curves based on componentially scored items clearly show that not only the overall shape of the curves are quite similar between the ethics prompt and the brain specialization prompt, the previously found minor differences between the economics prompt and either of the two other prompts are found to decrease greatly in that the economics prompt is found to elicit less information from the intermediate ability group but more information from the lower or higher ability groups. These changes result in higher degree of parallelism among the three writing prompts in terms of the information they elicit. Moreover, in regard to the prompt comparison in terms of the information of the holistically scored item, it is found that the overall shape of information curve of the brain specialization prompt differs greatly from the other two prompts in that it elicits most information from either extremely low or extremely high ability groups and at the same time, elicits little information from the intermediate ability group with respect to ESL examinees' latent ability. However, the overall shape of the information curves for the ethics prompt and the economics prompt is found to be almost identical with the only slight difference in the "peak" of the information curve.

In addition, step characteristic curves and information curves of each aspect of the three writing prompts are constructed. The comparison results show that in terms of the organization aspect, all three prompts elicited rather similar information, with the highest degree of parallelism found between the ethics prompt and the economics prompt. In terms of the content aspect, the three prompts once more elicited generally similar information; however, the highest degree of parallelism of the item information curves was found to occur between the ethics prompt and the brain specialization prompt. In terms of the grammar aspect, it was found that all three prompts elicited almost identical information and thus, overall the degree of parallelism across the three prompts was extremely high. Similarly, with respect to the style aspect, the item characteristic and information curves across all three prompts were again found to be almost completely identical.

Summing up for the fourth research question (i.e., To what degree are the three writing prompts parallel to one another in terms of step characteristic curves and information curves?) -- it was found that overall the three writing prompts elicited quite similar amount of information from ESL examinees' with respect to their academic writing ability. Specifically, the degree of parallelism between the ethics prompt and the brain specialization was slightly higher than that between the economics prompt and either of the other two prompts in terms of the information elicited.

Differential Prompt Functioning

The identification of potential systematic sub-patterns of behavior regarding prompt functioning is achieved through the "bias analysis" of the multi-faceted Rasch measurement. Under this framework, the study of bias is like studying potential "interaction" effects between facets specified in the measurement model. Specifically, the bias analysis works in such a way that all elements of the facets specified in the model are first calibrated in the main analysis and then those estimates are fixed to obtain the most likely scores (i.e., the expected scores) under particular combination of elements from the facets under investigation. The difference between expected scores and observed ratings divided by the number of observations contributing to that difference (i.e., the averaged residuals) is then used as the criterion to evaluate the actual difference between expected scores and observed ratings (i.e., the actual residuals). If the difference between the averaged residuals and the actual residuals is within a "small" range, no bias is reported; otherwise, the estimates of the size of the bias will be reported for further investigation. The acceptable range for the discrepancy between the averaged residuals and the actual residuals is normally determined by individual researchers and is specified in the format of z-scores in the program specification. Conventionally, values outside the range of approximately +2 to -2 z-score suggest significant bias (Linacre, 1996; McNamara, 1996). And, the direction (i.e., sign) of the bias depends on how it is defined in the program specification. If bias direction is defined along with the direction of task difficulty or rater severity, when the observed score is larger then expected score, the element of interest is interpreted as less difficult or more lenient than expected (i.e., the overall behavior pattern of that particular element). When the observed score is less than expected score, the element of interest is interpreted as more difficult / more severe than expected.

In this project, ESL examinees were divided into subgroups in terms of examinee characteristics such as academic status (i.e., graduates vs. undergraduates), field of study (i.e., science vs. business vs. liberal arts), and gender (male vs. female). The functioning of the three writing prompts of interest as well as the raters' judging severity were evaluated at different subgroups based on three examinees' characteristics.

Undergraduates vs. Graduates

For the undergraduate group, an interaction was found to occur between the subgroup facet and the aspect difficulty facet as well as the rater by aspect combination. However, the highest level of statistically significant interaction reported involved all three remaining facets in the measurement model, namely examinees' subgroup, rater severity and aspect difficulty. Because the highest-way interaction subsumes the lower-way interaction in that it provides the most detailed information regarding specific bias sources, only the size of the bias estimates and fit statistics of the highest way interaction are discussed here. Based on the z-score standard (i.e., ± two standard deviations), four statistically significant bias were reported by the FACETS program. It turned out that the grammar aspect of both the brain and ethics prompt when judged by rater 7 imposed less difficulty on undergraduate examinees than expected (i.e., the overall difficulty based on this particular combination of facet elements). At the same time, the organization aspect of both the brain specialization and ethics prompts imposed more difficulty on the undergraduate examinees than expected. However, because all four bias measures (bias = -1.07 logit, -0.70 logit, 0.79 logit, and 0.81 logit) are less than or about the size of one logit when measurement errors (standard error of measurement = 0.27 logit, 0.21 logit, 0.22 logit, and 0.18 logit) are taken into consideration, these reported statistically significant bias terms do not really affect the raw ratings examinees received and thus are considered not to be practically significant.

For the graduate group, an interaction was found to occur between the subgroup facet and the rater facet, the subgroup facet and the prompt by rater combination, the subgroup facet and the rater by aspect combination, as well as the subgroup facet and the prompt rater by aspect combination. Five statistically significant bias terms, which result from the highest way interaction, were identified. It was found that the grammar aspect of the ethics prompt when judged by rater 7 imposes less difficulty (i.e., bias =-1.34 logit) on the graduate examinees than expected (i.e., the overall difficulty based on this particular combination of facet elements). On the contrary, the overall aspect of the ethics prompt when judged by raters 4 and 5 imposed more difficulty (bias = 0.82 logit and 0.83 logit, respectively) to the graduate examinees than expected. In addition, both the grammar aspect of the economics prompt and the overall aspect of the brain specialization prompt when scored by rater 1(bias = 0.31 logit and 0.68 logit, respectively) turned out to be more difficult than expected for graduate examinees. Despite the statistical significance of these reported biases, the size of all biases is less than or close to 1 logit unit, which implies they do not have practical significance.

Science vs. Business vs. Liberal Arts Majors

In terms of differential facet, in total there were eight statistically significant sources of interaction between different facets in the four-faceted Rasch model being identified in the bias analysis. For the science-major group, it was found that the reported biases were caused by the impact of rater by aspect combination as well as prompt by rater by aspect combination. For the business-major group, the combined influence of prompt by rater, rater by aspect, and prompt by rater by aspect were found to be the sources of identified differential facet functioning. For the liberal arts-major group, the same bias sources as those for the business-major group, in terms of the facets involved, were identified. The size of all bias measures, standard error of estimate, and item fit statistics in the highest-way interaction found in the three different majors of subgroups are presented in Table 6.

Table 6
Size of Statistically Significant Bias Based on Examinees' Major

Interaction Sources	Bias+ (logit)	Z-score	Fit (MnSq)
Science x Ethics x Rater 7 x Grammar	–.93	–3.96	1.1
Science x Brain x Rater 7 x Organization	.90	3.24	1.0
Business x Brain x Rater 7 x Grammar	–1.34	–2.21	0.8
Liberal Arts x Ethics x Rater 7 x Grammar	–.86	–3.27	1.1
Liberal Arts x Ethics x Rater 3 x Style	–.81	–3.14	0.9
Liberal Arts x Ethics x Rater 7 x Organization	.94	4.30	0.7

Specifically, for the science subgroup, it was found that the grammar aspect of the ethic prompts when judged by rater 7 imposed less difficulty on examinees than expected overall difficulty level (bias = -0.93 logit), which is computed based on the joined product of the ethics prompt's overall difficulty and rater 7's overall severity and the grammar aspect's overall difficulty. At the same time, when the organization aspect of the brain prompt is evaluated by rater 7, the challenge of the task for the science group was greater than expected (bias = 0.90 logit). For the business group, the results of bias analysis show that the grammar aspect of the brain prompt when judged by rater 7 was less difficult than expected (bias = -1.34 logit). For the liberal arts group, the grammar aspect of the ethics prompt when rated by rater 7 was less difficult than expected (bias = -0.86), while the organization aspect of the ethic prompt when rated by rater 7 was more difficult than expected (bias = 0.94 logit). Additionally, it was found that the style aspect of the ethics prompt when judged by rater 3 was less difficult than expected (bias = -0.81 logit). However, these reported biases can be considered to be lacking practical significance in that after standard error of measurement being taken into account, the size of the reported bias was not large enough to affect the raw ratings examinees received.

Male vs. Female

Similar to the findings of the study of other subgroups (examinees' academic status and field of study), the size of all bias measures reported here was less than or around 1 logit and the item fit statistics were generally found to be acceptable. Based on the results of the bias analysis, regarding the female group, the grammar aspect of the ethic prompt when judged by rater 7 was less difficult than expected (bias = -0.95 logit) while the organization aspect of the ethics prompt when rated by rater 7 was more difficult than expected (bias = 0.64 logit). Regarding the male group, similar results were found in terms of the response patters based on the joint effect of the ethics prompt, the grammar and organization aspect, and rater 7 in that the grammar aspect of the ethics prompt were less difficult than expected (bias = -0.88 logit) and the organization aspect of the ethics prompt was more difficult than expected (bias = 0.72 logit) when judged by rater 7. Moreover, the same response pattern was found in terms of the brain prompt. That is, when the grammar aspect of the brain prompt was judged by rater 7, it imposed less difficulty for the male group than expected (bias = -1.02 logit) but more difficulty than expected (bias = 0.89 logit) for male examinees' performance on the organization aspect of the brain prompt scored by rater 7. In addition, the overall aspect of the brain prompt when judged by rater 1 was found to be more difficult than expected (bias = 0.94 logit) for the male examinees. However, all the bias estimates, though identified as statistically significant, were of small values, suggesting that they had no practical significance in the sense that little difference in raw ratings would result from them.

Summing up for the fifth research question (i.e., Do the three writing prompts display differential functioning across subgroups of ESL examinees in terms of their academic status, major, and gender?) -- although a few statistical biases were found as a result of differential prompt functioning in ESL examinees' subgroups in terms of academic status, major, and gender, none of the identified bias are serious enough to cause differences in observed ratings, and thus, are considered as practically insignificance. In other words, practically speaking, the three writing prompts are functioning equivalently in terms of these investigated examinees subgroups.

Outlier Analysis

Multi-faceted Rasch analysis enables researchers to examine the "coherence" of an individual's response with all other individuals' responses of interest (McNamara, 1996) and that information is provided through "person-fit" statistics by the FACETS program (Linacre, 1996). Person-fit statistics can, thus, be used to identify examinees whose writing performance responses differ from the normal pattern of most examinees (i.e., outliers or "misfitting" persons). The same interpretation can be used on raters. If an unexpectedly large fit statistic (i.e., mean square of residuals) is found to be associated with a particular rater, this is an indication that the rater does not judge examinees' writing the same way as the other raters.

Examinees

There are four examinees (i.e., examinee 45, examinee 75, examinee 175, and examinee 181) whose ability measures calibrated based on both prompts responded were found to be associated with "unacceptably" large mean square values (i.e., beyond two standard deviation range). That is, these four examinees are identified as outliers in that their response patterns are not "congruent" with that of the other 196 examinees. Moreover, at the individual prompt level, five examinees (i.e., examinee 68, examinee 90, examinee 45, examinee 20 and examinee 23) are identified as outliers based on their performance on the ethics prompt; two examinees (i.e., examinee 48 and examinee 64) are found to be outliers based on their performance on the economics prompt and finally, two other examinees (i.e., examinee 125 and examinee 129) turn out to be outliers based on their performance on the brain prompt.

Overall, no consistent pattern was found among the identified outlier examinees. Based on their personal profiles, they came from different first language backgrounds, received different training in English, differed in their general English proficiency, were majoring in different academic fields, and were interested in different issues. It was, therefore, concluded that no systematic bias was found to underlie their aberrant response patterns. Rather, their unexpected response patterns were more likely to be as a result of individual idiosyncrasies.

Raters

The results of the examination of person-fit statistic show that none of the eight raters was found to be associated with "unacceptable" mean square residuals at the all-prompt level (see Table 7). In other words, it was found that all raters displayed rating patterns that were generally congruent with one another despite that the degree of fit varies across individual raters. For instance, the rating patterns of both rater 3 and rater 6 fit perfectly with the general rating trend while rater 7's rating patterns departs, comparatively, further from the general trend.

Table 7
Rater Calibration Results: All Prompts Combined

Raters	Measure (logits)	Model S.E.	Infit (MnSq)	Outfit (MnSq)
5	−1.95	.19	1.1	1.0
4	−1.19	.19	1.2	1.3
2	−.86	.14	1.3	1.1
6	−.61	.18	1.0	0.9
3	.73	.07	1.0	1.0
7	1.19	.09	1.4	1.4
1	1.20	.05	1.2	1.1
8	1.49	.05	0.9	0.9

However, at the individual prompt level, rater 5 was found to be the only "misfitting" rater in terms of her rating patterns on the economics prompt. That is, in terms of this particular prompt, rater 5 somehow judged the examinees' essay performance differently from the rest of the raters. Since both the infit and outfit statistics associated with rater 5 were found to be below acceptable range (i.e., 0.6), significant overfit is clearly indicated. Significant overfit, simply speaking, suggests that a dependency exists between the ratings assigned by rater 5 on the essays of the economics prompt. In this project, rater 5 rated essays from all three prompts; however, she only rated them holistically but not componentially. Her background profile is described as follows: rater 5 was a native speaker of English and an experienced teaching assistant for both the ESL Service courses and the Intensive English Institute (IEI) here at UIUC. She had taught both graduate (i.e., ESL 401) and undergraduate (i.e., ESL 113) ESL courses, as well as the level 2 (i.e., advanced level) of composition component in IEI. She has graded the EPT essays for the past three semesters. Although she attended the rater training session before, she has never been a rater trainer. She reported that she felt quite confident in interpreting the level descriptors in the holistic rating rubrics and believed that she interpreted those levels the same as other raters. Nevertheless, she also reported that she thought that the holistic rating guidelines should not be literally followed in real scoring scenario. Rather, she claimed that she believed in using the rating rubrics only as general guidelines. Other than rater 5, all other seven raters were found to rate essays in a similar fashion even at the individual prompt level.

In summary for the sixth research question (i.e., Are there examinees or raters displaying aberrant behaviors from the majority?) -- the answer is "yes". Overall, four examinees were found to display idiosyncratic response patterns and a few more examinees are identified as outliers in terms of their response patterns in individual writing prompts. Nevertheless, they consisted of only a tiny fraction of all examinees. Similarly, it was found that all eight raters overall interpreted the rating rubrics in a rather homogeneous way. Only one rater was found to use less score categories than other raters when rating the economics-prompt-based essays holistically.

The Hypothesized Interaction Factor Affecting the Parallelism Across Performance-Based Writing Prompts

The effects of lecturer-audience interaction on prompt parallelism were evaluated at test specification level, decision reproducibility level, skill profile level, as well as prompt information level. At the test specification level, three out of the four prompt judges directly pointed out the "extra" question-and-answer section at the end of the video lecture in the economics prompt was a clear violation to *the EPT Specification for Video-Reading Academic Essays.* After this was pointed out by three of the judges, the fourth prompt judge also acknowledged that lecturer-audience interaction was one of the factors that affected the connection between the economics prompt and the EPT specification. Thus, the lecturer-audience interaction is indeed an influencing factor for prompt comparability in the present EPT format at the test specification level. At the decision level, it is found that although the three writing prompts of interest cannot be considered as "strictly parallel" forms in that the degree of prompt comparability is found to be higher between the ethics prompt and the brain specialization prompt than between the economics prompt and each of the other two prompts, these three writing prompts can at least be viewed as "weakly" parallel in the sense that these three writing prompts resulted in fairly consistent placement decisions, with only "small" error rate (i.e., 5%) being identified. Therefore, the hypothesis that lecturer-audience interaction affected prompt comparability is supported by empirical evidence; nevertheless, its influence on prompt comparability is only minor because the three writing prompts were still found to be generally comparable to one another. Furthermore, at the skill profile level, in terms of overall aspect (via holistic scoring), it was found that the skill profiles elicited by the ethics prompt and the brain prompt were more similar compared with those elicited by the economics prompt. This finding suggests that lecturer-audience interaction did affect prompt parallelism when essays were holistically scored. However, except for the overall aspect, the three prompts of interest were found to be comparable to each other in terms of aspect difficulty in four subcategories (via componential scoring) when measurement errors were taken into account, and thus elicited similar skill profiles of examinees' writing strengths and weaknesses. The finding implied that the lecturer-audience interaction had little or no effect on prompt comparability when essays were componentially scored. Finally, when the prompt information curves were compared across these three writing prompts based on both holistic and componential essay scores, it was found that the overall shapes of the information curves for the ethics prompt and the brain prompt were remarkably similar and, comparatively, the overall shape of the information curve of the economics prompt was more different from the other two prompts. Therefore, it was concluded that the lecturer-audience interaction indeed exhibited some impact on prompt comparability in terms of the information elicited from ESL examinees with respect to their latent writing ability.

Summing up for the last research question (i.e., Is the hypothesis that lecturer-audience interaction would affect the degree of parallelism among the three writing prompts being supported by empirical evidence?) -- the empirical evidence basically supports that the hypothesized impact of the lecturer-audience interaction on prompt comparability but the degree of influence was only minor.

Final Remarks

Conclusions

Based on the findings of the investigation on prompt comparability, it can be concluded that overall, despite of the differences between the three prompts in terms of content technicality and the connection with the *EPT Specification for Video-Reading Academic Essays*, the degree of comparability among the three new EPT writing prompts was quite high in terms of both the placement reproducibility and the skill profile agreement. It appeared that the "lecture-audience interaction" did not exhibit the kind of expected impact on the degree of comparability among the three writing prompts. At best, its impact was minor. Nonetheless, if the three new writing prompts were used as equivalent forms of the EPT essay test, overall about five percent of error rate in terms of placement consistency would still exist. If this amount of placement discrepancy is acceptable by test users, the three new writing prompts can very well be treated as equivalent forms for the future EPT use. Moreover, because it was found that componential scoring resulted in very similar skill profiles of examinees' across the three writing prompts, it was concluded that if the three writing prompts are to be used interchangeably, it was better that the componential rating method be used to score examinees' essays. The use of the holistic rating method alone did not guarantee the interchangeability of the three writing prompts. In addition, it was found that when examinees were divided into subgroups in terms of their academic status (i.e., graduates vs. undergraduates), field of specialization (i.e., science major vs. business major vs. liberal arts major), and gender (i.e., female vs. male), none of the few reported biases in all three different sets of subgroups is of practical concern in terms of their size, despite of their statistical significance. Therefore, it is concluded that no bias exists as a result of examinees' subgroups and that the three new writing prompts function equivalently when responded to by examinees from the different subgroups studied here. Furthermore, the results of the outlier analyses on both examinees and raters revealed that basically the three new writing prompts were valid test instruments for the majority of examinees in eliciting their writing performance that is representative of their underlying writing ability. And, the EPT raters also interpreted the rating rubrics, including both holistic and componential rating rubrics, in a similar manner, due to the rating training as well as their teaching experience in ESL service courses.

Implications

This empirical study clearly demonstrates that the multi-faceted Rasch model is an effective measurement model for polytomously scored items where judgment is involved in assigning scores. And, since the multi-faceted measurement model provides valuable diagnostic information concerning how the various individual aspects of a writing assessment program are functioning (i.e., the individual raters, examinees, prompts or topics, rating scale, etc.), test developers or program directors can use this information to improve the accuracy, accountability and fairness of an assessment system. For example, The results of the multi-faceted measurement models provide useful diagnostic information for rater training in that individualized rater profiles are obtained. The individual rater profiles include not only the information regarding rating severity of individual raters, which is reflected on the rating scale individual raters actually apply to score examinees' test performance, but also the way in which individual raters interpret the rating rubrics, compared with other raters. With individual profiles, rater trainers can assess the results of rater training as well as specifically determine what part of the rater training program needs to be improved along with which raters need more training. If some raters are found to exercise a harsher standard on scoring examinee's performance, experienced raters or rater trainer can discuss more about the rating rubrics with those particular raters by checking on their interpretation of the specified rating rubrics, going through more exemplary essays and score more essays against the rating rubrics. If the difference in rating severity arises from misinterpreting the criteria listed in the rating rubrics, once the misconception is cleared, raters should interpret the rating rubrics in a more or less homogeneous way. Alternatively, if there is more than one feature in the rating rubrics, then it is likely that the individual rater difference comes from assigning different weights to different features, consciously or unconsciously. Through extensive discussion and practice, raters who are found to employ individual rating scales can be trained to operationalize the rating rubrics consistently in scoring real essays. Moreover, the individual rating behavior can also provide valuable feedback for revising the rating rubrics such that the rating scale is more strictly defined to avoid unnecessary misinterpretation.

The findings also suggest that criterion-referenced-based test specifications provide an important basis for developing and validating parallel forms of performance-based items in the sense that test specifications include not only the specification of the construct(s) to be measured but also the specification of the attributes of the test instruments that are to measure the construct(s). On the basis of the specifications, content experts or testing experts can easily evaluate the comparability of different forms of performance-based writing test in terms of the specified construct(s) as well as the test instruments, and also ensure the validity of the score inferences. Messick (1995) contended that a construct-driven rather than a task-driven approach to performance assessment should be adopted because the meaning of the construct guides the selection or construction of relevant tasks. Criterion-referenced-based test specification is intended to be such a construct-driven approach. Through the iterative process of constructing criterion-referenced-based test specification, the intended construct underlying the test can be explicitly defined in terms of what examinees know and can do, which is essential in criterion-referenced measurement where the goal for a criterion-referenced test is to be able to interpret test results in terms of the observable behaviors of the intended latent trait (Linn, 1995). On the other hand, the design of relevant tasks for operationalizing the specified construct can be based on the specifications of prompt attributes. Furthermore, when the construct being measured and the measurement instruments are comparable across different test forms, they are highly likely to elicit similar examinee behaviors representing the same underlying trait.

Consider the three new writing prompts from this project. Examining the prompts themselves, it is hard to evaluate their comparability because the topics and content of the three prompts apparently are different. However, examining the *EPT Specification for Video-Reading Academic Essays*, from which the three prompts were generated, reveals that the three prompts are actually measuring the same construct (i.e., examinee's writing ability). Additionally, the specified attributes of these three prompts are quite similar with the only difference in the extra "lecturer-audience interaction" attribute of the economics prompt. As a result, the degree of comparability across the three prompts can be evaluated based on the extent they follow the EPT test specification. In other words, as long as high-quality criterion-referenced based test specifications with clear and sound definition on the construct of interest and the type of test instruments that can adequately measure the intended construct are available, constructing and validating parallel forms of performance-based writing prompts has a theoretical basis upon which to fall. Furthermore, the definition of academic writing competence may vary from one academic setting from another. By specifying the operational definition of the construct and the types of test instruments that most adequately elicit representative behvariours of the construct, whether the results of the performance-based writing assessment in one academic setting can be generalized to another academic setting can be better determined.

Finally, although test developers attempt to construct test forms that are as similar as possible to one another in both the construct and the test instrument according to the test specification, the forms that are intended to be parallel typically differ somewhat in difficulty. Equating is intended to statistically adjust for such differences in difficulty, thus allowing the forms to be used interchangeably. In the current context, since difficulty parameters across the three writing prompts were calibrated on the same scale (i.e., logit scale) through the multi-faceted-Rasch-measurement-based FACETS program, differences in prompt difficulties can easily be adjusted statistically, if desired. Additionally, the unintended bias caused by the interaction between elements of facets in the specified measurement model such as rater severity, prompt difficulty, domain difficulty, and rating scale structure is also easy to adjust statistically, if necessary, because it is estimated on the common logit scale as are the other parameters in the model.

Limitations of the Current Study

Except for the essays that were scored during the EPT operations (i.e., essays of the ethics prompt), the rest of the essays (i.e., essays of the economics prompt and the brain prompt) were taken home by raters who participated in this study. Unlike the operationally rated essays, raters evaluated the non-operational essays without a time limit and know that the results of the scoring would have no impact on the ESL examinees or ESL service courses. In addition, although all raters had received proper training in using the rating rubrics (both holistic and componential), not all of them had chances to recalibrate the rating rubrics (two of the eight raters did) before they rated the non-operational essays, as they had done during the operational scoring. The differences in rating conditions and perceived consequences of the scoring may lead to unknown impact on the ratings assigned by the raters, which may have indirectly affected the score comparability across the three writing prompts.

Since multiple language skills were involved in the current format of the video-reading-based EPT essay test, the existence and identification of a dominant trait in academic writing tasks was extremely important both statistically and substantially. In this project, the existence of a dominant trait was statistically evaluated based on the investigation on the psychometric unidimensionality and the identification of the dominant trait is based on the EPT test specification, where the major construct is defined. However, more and stronger empirical evidence can be obtained regarding the dominant trait underlying the video-reading-based academic essays if multi-trait-multi-method analysis is conducted. Unfortunately, since there was no reliable, independent measure of each of the language skills involved in the academic tasks, multi-trait-multi-method analysis could not be performed.

Moreover, due to time constraints, comprehensible input, the hypothesized factor that may affect prompt comparability, was operationalized in this project through one-time "question-and-answer" exchange between the professor lecturer and one of the student audience. Whether that additional exchange actually facilitated the comprehensibility of the video-lecture and thus, changed the difficulty of the writing prompt is unknown because no independent measure was conducted. If the impact of the additional question-and-answer exchange on examinees' comprehensibility of the video-lecture is empirically validated, the conclusion regarding the affect of comprehensible input on prompt comparability could then have been generalized to other testing contexts with more confidence.

Future Research Directions

Although the feasibility and utility of the multi-faceted Rasch model in the performance-based writing tests has been empirically established, it can only be applied to datasets that fit the one-parameter (i.e., Rasch) model, where item difficulty is assumed to be the only estimated item parameter, with discrimination being the same across items. To incorporate the analysis for data which display different degrees of item discrimination, the multi-faceted Rasch model needs to be extended to a multi-faceted two-parameter model in which both item difficulty and item discrimination are included in the IRT model. Additionally, future research on performance-based writing tests in academic settings should also explore the possibility of extending the unidimensional multi-faceted Rasch model to the multidimensional model, in case there is more than one dominant trait underlying performance-based academic writing tasks.

With well-defined scoring rubrics and high-quality rater training, past studies as well as this study have demonstrated that quite high levels of generalizability across raters can be achieved (Dunbar, Koretz & Hoover, 1991; Shavelson, Gao, & Baxter, 1993; Linn & Burton, 1994). On the other hand, it has also been observed that performance-based tasks are generally of high degree of task specificity; that is, limited degree of across-task generalizability (Linn, 1995; van der Vleuten & Swanson, 1990; Linn & Burton, 1994), which implies that it is very difficulty to develop completely parallel forms in performance-based format. Nevertheless, the need for alternate performance-based test forms does exist due to the desired positive washback on curriculum as well as test security and fairness. Consequently, instead of focusing on overall writing task generalizability in different contexts, perhaps the more efficient and practical approach in developing and validating parallel forms of performance-based writing tasks is to focus on only the level of needed generalizability, depending on the purpose of the test and the importance of the decisions to be made (i.e., decision consistency). Therefore, it is suggested that more future studies look into the issue of prompt comparability in performance-based writing assessment in terms of decision consistency, rather than equivalent score distribution.

Moreover, although the emphasis in authenticity in performance-based writing assessment in academic settings comes from the desire to make writing tasks as similar to academic writing tasks as possible such that the utility and predictability of tests will increase, the format of performance-based assessment itself does not automatically guarantee higher validity of the assessment of performance compared to tests with multiple choice questions. Therefore, on the one hand, the correlation between the placement decisions made on the basis of performance-based academic writing assessment and the students' general academic performance in their own fields needs to be studied so that the degree of predictability can be determined. On the other hand, more empirical studies are needed to investigate the nature and the strength of the interaction between various language skills involved in academic writing tasks, as well as its impact on the measure of examinee's ability in writing academic essays because test validation processes become more complicated for performance-based writing tests due to the additional complexity introduced by various potential language skills involved in the authentic academic writing tasks.

Finally, it is recommended that future studies on prompt comparability should always include differential prompt functioning analyses, which check for potential bias caused by unintended factors or traits because if differential prompt functioning is identified, the validity of score comparability across performance-based writing prompts is threatened. Due to the complexities and skills involved in the performance-based writing assessment, the existence of potential differential prompt functioning is likely and should always be carefully monitored, especially in large scale testing contexts where examinees may come from a variety of backgrounds.

REFERENCES

Bode, R. (1994). Controlling for demographic characteristics in person measures using a many-faceted Rasch model. Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, LA.

Brand, A. (1991). Constructing tasks for direct writing assessment: A frontier revisited. (ERIC Document Reproduction Service No. ED 340 037).

DeMauro, G. (1992). An investigation of the appropriateness of the TOEFL as a matching variable to equate TWE topics (TOEFL Research Report No. 37). Educational Testing Service.

Dunbar, S., Koretz, D., & Hoover, H. (1991). Quality control in the development and use of performance assessment. Applied Measurement in Education, 4, 239-304.

Hamp-Lyons, L. (1990). Second language writing: Assessment issues. In Kroll, B. (Ed.), Second language writing: Research insights for the classroom (pp.69-87). Cambridge: Cambridge University Press.

Lehmann, R. (1993). Rating the quality of student writing: Findings from the IEA study of achievement in written composition. In A. Huhta et al. (Eds.), Language testing: New openings (pp.186-204). University of Jyvaskyla Press.

Linacre, J. (1989). Many-faceted Rasch Measurement. Unpublished Ph.D. Dissertation. University of Chicago.

Linacre, J. (1996). A user's guide to FACETS: Rasch measurement computer program (Version 2.9). MESA Press:Chicago.

Linn, R. & Burton, E. (1994). Performance-based assessment: Implications of task specificity. Educational Meausrement: Issues and Practice, 13(1), 5-8.

Linn, R. (1995). High-stakes uses of performance-based assessments: Rationale, examples, and problems of comparability. In T. Oakland & R. Hambleton (Eds.), International perspectives on academic assessment (pp. 49 -73). Kluwer Academic Publishers: Boston.

McNamara, T. (1996). Measuring Second Language Performance. Longman: London and New York.

Messick, S. (1995). Standards of validity and the validity of standards in performance assessment. Educational Measurement: Issues and Practice, 14 (4), 5-8.

Petersen, N., Kolen, M., & Hoover, H. (1993). Scaling, norming, and equating. In R. Linn (Ed.), Educational measurement (3rd ed., pp. 221-262). National Council on Education: ORYX Press.

Shavelson, R., Gao, X., & Baxter, G. (1993). Sampling variability of perfromance assessments. Journal of Educational Measurement, 30, 215-232.

van der Vleuten, C., & Swanson, D. (1990). Assessment of clinical skills with standardized patients: The state of the art. Teaching and Learning in Medicine, 2, 58-76.

APPENDIX

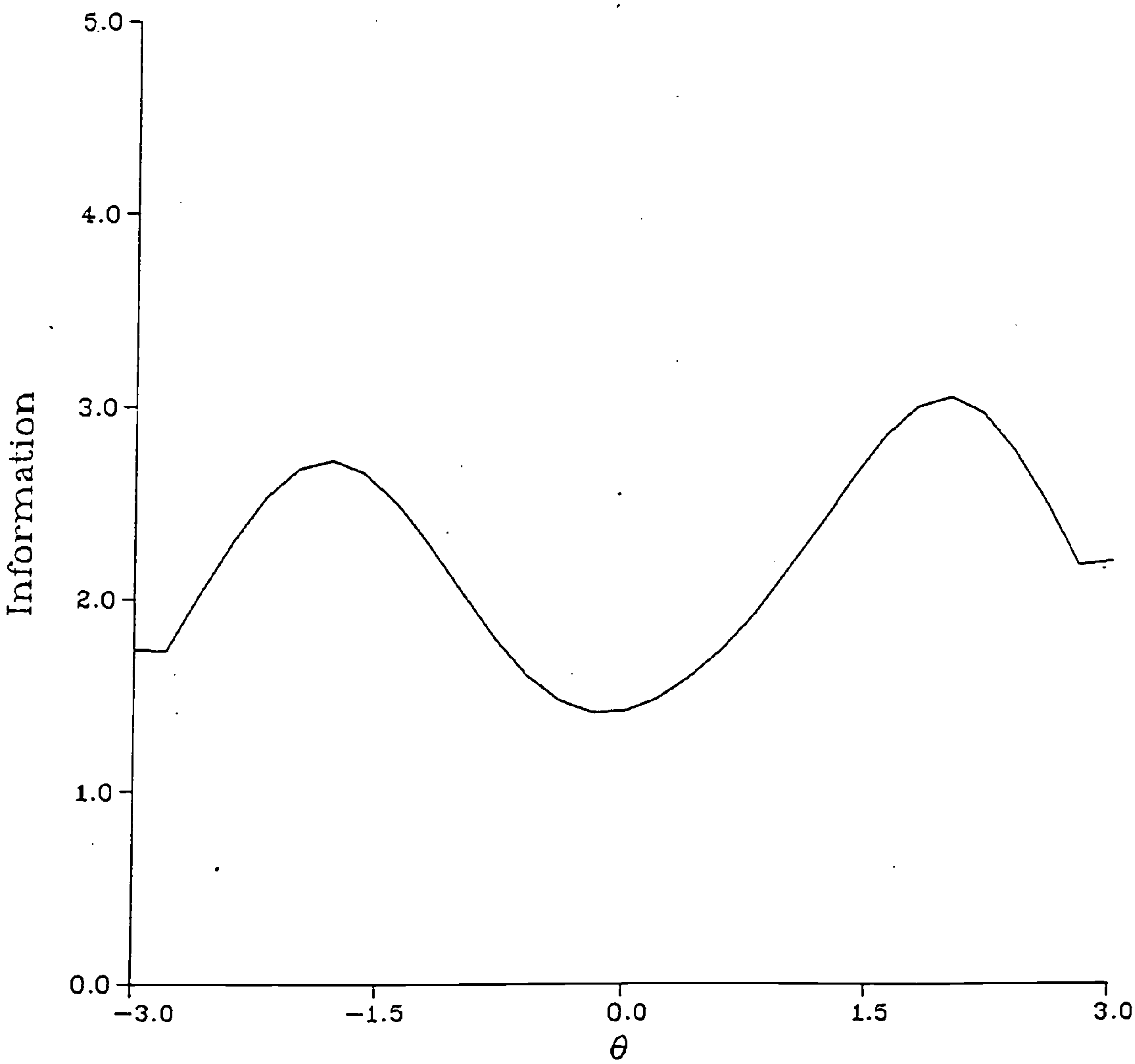

Figure 2. The information curve for the ethics prompt

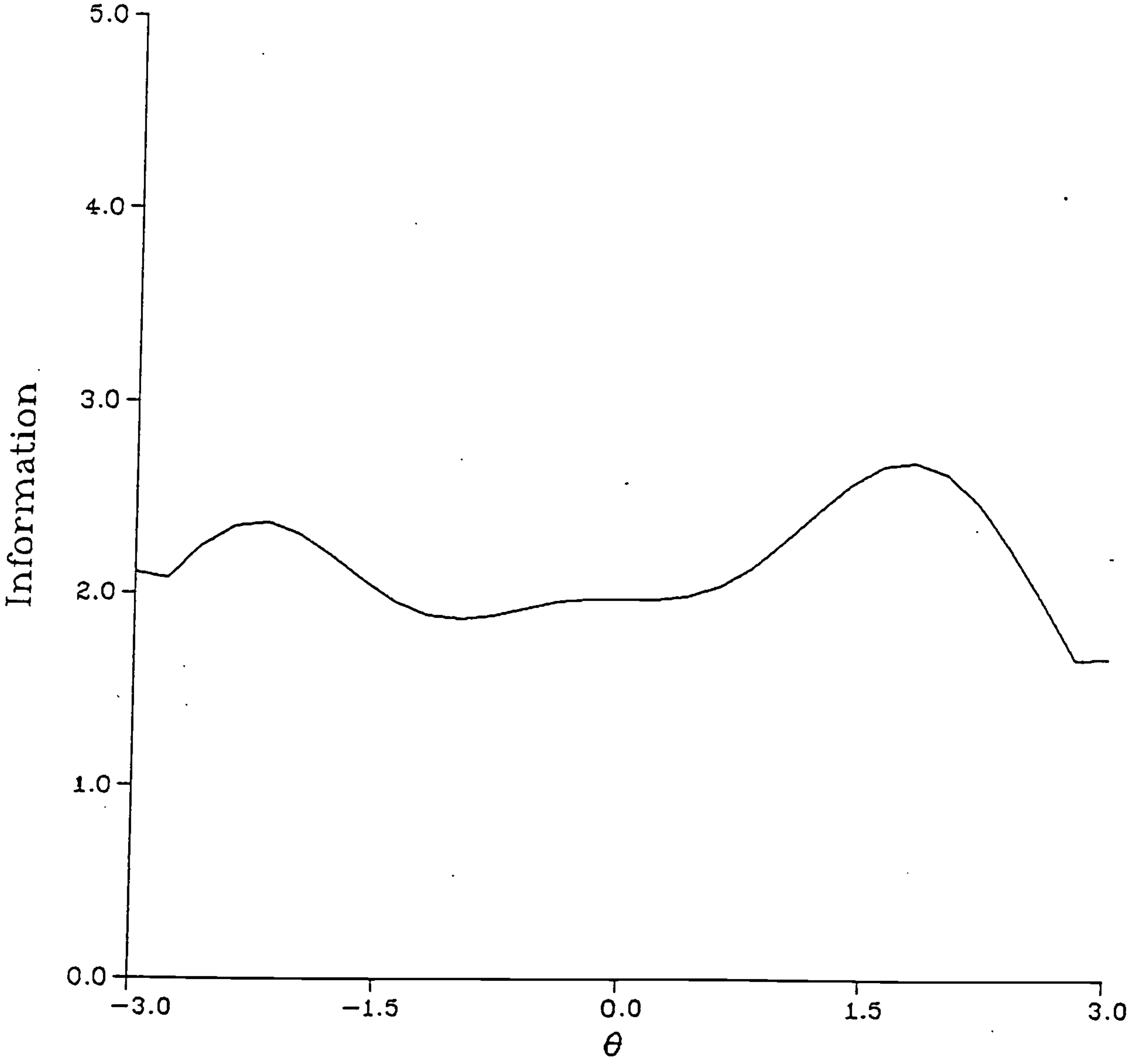

Figure 3. The information curve for the economics prompt

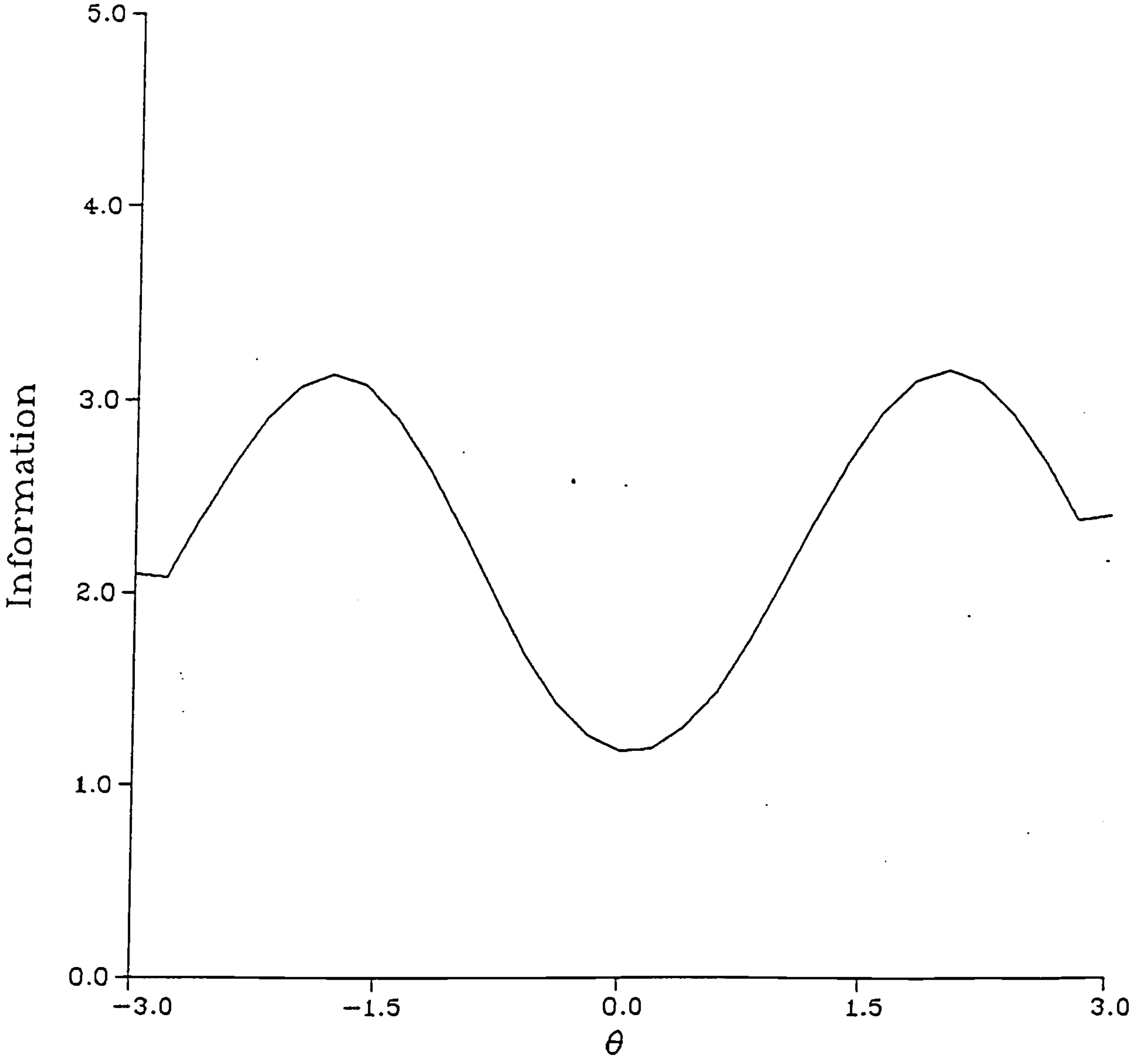

Figure 4. The information curve for the brain specialization prompt

REPRODUCTION RELEASE

(Specific Document)

FL025411

I. DOCUMENT IDENTIFICATION:

Title: Constructing and Validating Parallel Forms of Performance-based Writing Tasks in Academic Settings

Author(s):

LTRC presentation? [✓ yes] —no If no, was this presented elsewhere? —yes —no Specify:

Publication Date: N/A

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic/optical media, and sold through the ERIC Document Reproduction Service (EDRS) or other ERIC vendors. Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following two options and sign at the bottom of the page.

☒

Check here
For Level 1 Release:
Permitting reproduction in microfiche (4" x 6" film) or other ERIC archival media (e.g., electronic or optical) *and* paper copy.

The sample sticker shown below will be affixed to all **Level 1** documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

Level 1

The sample sticker shown below will be affixed to all **Level 2** documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN OTHER THAN PAPER COPY HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

Level 2

□

Check here
For Level 2 Release:
Permitting reproduction in microfiche (4" x 6" film) or other ERIC archival media (e.g., electronic or optical), but *not* in paper copy.

Documents will be processed as indicated provided reproduction quality permits. If permission to reproduce is granted, but neither box is checked, documents will be processed at **Level 1**.

"I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche or electronic/optical media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries."

Sign here→ please

Signature: Angie Liu

Printed Name/Position/Title: Angie H.C. LIU, Ph.D.

Organization/Address:
Texas Education Agency
1701 N. Congress Ave., Austin,
Tx 78701

Telephone: (512)463-9199

FAX: (512)249-5963

E-Mail Address: aliu@tmail.tea.state.tx.us

Date: 07/30/98

Office of Policy Planning and Evaluation

(over)

III. DOCUMENT AVAILABILITY INFORMATION (FROM NON-ERIC SOURCE):

If permission to reproduce is not granted to ERIC, *or*, if you wish ERIC to cite the availability of the document from another source, please provide the following information regarding the availability of the document. (ERIC will not announce a document unless it is publicly available, and a dependable source can be specified. Contributors should also be aware that ERIC selection criteria are significantly more stringent for documents that cannot be made available through EDRS.)

Publisher/Distributor: N/A

Address:

Price:

IV. REFERRAL OF ERIC TO COPYRIGHT/REPRODUCTION RIGHTS HOLDER:

If the right to grant reproduction release is held by someone other than the addressee, please provide the appropriate name and address:

Name: N/A

Address:

V. WHERE TO SEND THIS FORM:

Send this form to the following ERIC Clearinghouse:

ERIC Clearinghouse on
Languages & Linguistics
1118 22nd Street NW
Washington, D.C. 20037

However, if solicited by the ERIC Facility, or if making an unsolicited contribution to ERIC, return this form (and the document being contributed) to: