Construct validation of analytic rating scales in a speaking assessment: Reporting a score profile and a composite

Yasuyo Sawaki

Language Testing 2007; 24; 355
DOI: 10.1177/0265532207077205

Construct validation of analytic rating scales in a speaking assessment: Reporting a score profile and a composite

Yasuyo Sawaki *Educational Testing Service, Princeton*

This is a construct validation study of a second language speaking assessment that reported a language profile based on analytic rating scales and a composite score. The study addressed three key issues: score dependability, convergent/discriminant validity of analytic rating scales and the weighting of analytic ratings in the composite score. Confirmatory factor analysis (CFA) and multivariate generalizability theory (G theory) were combined to analyze the responses of 214 admits to a study-abroad program to two role-play speaking tasks in a Spanish speaking assessment designed for student placement and diagnosis. The CFA and G theory approaches provided complementary information, which generally confirmed the key features of the assessment design: (1) the multicomponential and yet highly correlated nature of the five analytic rating scales: Pronunciation, Vocabulary, Cohesion, Organization and Grammar, (2) the high dependability of the ratings and the resulting placement decisions appropriate for the high-stakes decision-making context based on these analytic rating scales and the composite score, and (3) the largest contribution of Grammar to the composite score variance, which was consistent with the intention of program faculty members to reflect in the test design the relative importance of knowledge of grammar for students' academic success in the study-abroad program.

I Introduction

In second language performance assessment, analytic rating scales are often used to assess candidates' language ability within a single modality (e.g., speaking). Rationales in the literature for adopting analytic over holistic rating scales include the availability of rich information about examinees' language ability (Brown & Bailey, 1984; Pollitt & Hutchinson, 1987; Kondo-Brown, 2002; Bachman, Lynch & Mason, 1995); increased accuracy of ratings by drawing

Address for correspondence: Dr. Yasuyo Sawaki, Center for Validity Research, MS 10-R, Educational Testing Service, Rosedale Road, Princeton, NJ 08541, USA; email: ysawaki@ets.org

Language Testing 2007 24 (3) 355–390 10.1177/0265532207077205

judges' attention to specific criteria (Brown & Bailey, 1984); and consistency with the current multicomponential definition of language ability (Bachman, Lynch & Mason, 1995). Scores based on analytic rating scales can be reported in various forms. Multiple scores from individual scales can be reported separately as a language profile. They can also be accompanied by some sort of an overall score, which can be an additional rating obtained on an overall scale (e.g., Elder, 1993; McNamara, 1990, 1996); or a composite score obtained by averaging or summing across the scores on analytic scales by weighting all components equally (e.g., Brown & Bailey, 1984; Kondo-Brown, 2002) or differentially (e.g., Jacobs et al., 1981; Weigle, 1998).

In order for multiple scores reported within a language modality based on analytic rating scales to be useful for an intended purpose, empirical evidence must support the test design in three ways. First, empirical interrelationships among analytic rating scales must show that the scales are related to one another (convergent validity), and also distinct enough so that each scale provides information about a unique aspect of a candidate's language ability (discriminant validity). Second, when an overall score is reported in addition to analytic ratings, the empirical relationship of the analytic scales to the overall score, i.e., the weighting of individual analytic rating scales in an overall score, should be congruent with the relative importance of different aspects of language ability for a given purpose of assessment in a particular context. Finally, ratings provided by raters should be reliable, and decisions made based on such ratings should be dependable for analytic ratings as well as an overall score.

In reference to these three key issues, a number of previous studies have addressed score dependability, while only a few have examined convergent/discriminant validity, and relationships between analytic ratings and an overall score. As an attempt to narrow this gap, the present study investigated these issues for a Spanish speaking assessment by combining confirmatory factor analysis (CFA) and multivariate generalizability theory (G theory).

II Review of literature

The validation issue that has received the most attention in previous research into L2 performance assessments based on analytic rating scales is score dependability. Previous researchers addressed this topic extensively by employing G theory (Cronbach, Gleser, Nanda & Rajaratnam, 1972; Shavelson & Webb, 1991) and item response the-

ory approaches such as many-facet Rasch measurement (Linacre, 1989; Wright and Masters, 1982) as implemented by the computer program FACETS. Moreover, previous researchers who combined both in their analyses of second language performance assessments (Bachman, Lynch & Mason, 1995; Lynch & McNamara, 1998) advanced our understanding of the complementary role that these two analytic approaches play. On the one hand, G theory analyses of second language performance assessments (e.g., Brown & Bailey, 1984; Lynch & McNamara, 1998) have offered a principled approach to assessment design at a *global* level. In this approach, the information about score variability due to different sources of variation (e.g., candidates' true differences in their ability, rater severity, task difficulty and their interactions) guides test developers in determining, for example, how many tasks and ratings are required to achieve a desired level of score dependability to make various types of decisions. On the other hand, applications of many-facet Rasch measurement (e.g., Brown, 1995; Kondo-Brown, 2002; Lumley & McNamara, 1995; McNamara, 1990; 1996; Weigle, 1998; Wigglesworth, 1993) have functioned as a magnifying glass, where the analysis allowed score users to take a closer look at individual candidates, raters and tasks. This line of research boosted our understanding of individual rater behaviors, in particular. For example, information about a rater's overall severity, and rating consistency over time, as well as systematic scoring patterns observed when a particular rater encounters a particular rating scale, task or candidate (rater bias), contributed useful information for rater training and monitoring (e.g., Kondo-Brown, 2002; Lumley & McNamara, 1995; Weigle, 1998; Wigglesworth, 1993).

In contrast with the richness of previous research on score dependability, relatively little empirical evidence is currently available about the interrelationships, or convergent/discriminant validity, among analytic rating scales. Some relevant information can be found in two types of previous investigations. First, factor analytic studies showed that, for example, analytic rating scales for speaking assessments loaded together to form a distinct factor when modeled with other language ability measures in other modalities (e.g., Bachman, Davidson, Ryan, & Choi, 1995; Carroll, 1983; Kunnan, 1995; Shin, 2005). Second, recent applications of multivariate G theory (Cronbach et al., 1972; Webb, Shavelson & Maddahian, 1983; Brennan, 2001) to language assessments reported high universe-score correlations (a G theory analogue of correlations adjusted for measurement error) among analytic rating scales in language assessments (e.g., Lee, 2005; Lee &

Kantor, 2005; Sawaki, 2003, 2005; Xi, 2003). However, previous studies did not systematically attempt to falsify alternative explanations about interrelationships among analytic rating scales within a single modality, either because it was not the focus of a given investigation or because the analytic approach employed did not allow a close investigation of this issue.

With regard to the relationship of analytic rating scales to an overall score, even less information is available. Two studies by McNamara (1990) and Elder (1993) touched upon this in contexts where a separate overall rating and analytic ratings for different aspects of language ability were obtained. For example, in his FACETS analysis of the analytic scales in the Speaking and Writing sections of the Occupational English Test (OET) in Australia, McNamara (1990) identified an unexpected interdependence between two rating scales, *Overall Effectiveness* (overall rating scale) and *Resources of Grammar and Expression* (one of the analytic rating scales), by inspecting the fit measures for the scales. McNamara (1990) speculated that this finding might be explained by an overriding role that the grammar scale played in the raters' judgments on the overall scale. A multiple regression analysis showed that as much as 68–70% of the total variances on the overall ratings for both the Speaking and Writing sections were accounted for by the performance on the grammar rating scale devised in each section, confirming McNamara's hypothesis. In a similar vein, Elder (1993) also conducted a stepwise multiple regression analysis on the ratings given by content specialist vs. ESL specialist rating groups regarding English language behaviors of math teacher trainees in a classroom observation schedule and found that the two rater groups weighted different analytic scales differentially.

Besides the two studies above, there are few other recent language assessment studies that address the empirical relationship of analytic rating scales to an overall score. Some previous studies indicated how different analytic rating scales were weighted for obtaining a composite score. For example, Weigle (1998) reported the use of differential weighting of components in an ESL writing placement exam, where the score on one dimension (*Language*) was doubled before it was combined with two other ratings (*Content* and *Rhetorical Control*) to obtain a composite. Note, however, that these are *nominal weights*, or the weights that are used for calculating a composite score. Although nominal weights often represent the test developer's desired weights, reflecting the relative importance of different components, for example, the nominal weights are not necessarily the same as

effective weights (Wang & Stanley, 1970), namely, the degree to which individual rating scales empirically contribute information to a composite. Wang and Stanley (1970) pointed out that a common misconception is that nominal weights are equal to effective weights. The effective weight of a given rating scale is a function of three things: (1) the nominal weight given to the scale; (2) the variance of the scale; and (3) the covariances of the scale with the others (Bachman, 2005; Peterson, Kolen, & Hoover, 1988; Wang & Stanley, 1970). An important implication of this argument is that assigning a larger nominal weight to a given rating scale may not necessarily lead to a greater weighting of that scale than the others when (1) the variance of the scale is small, (2) the covariances of the scale with the others are small, or both. Accordingly, the empirical contribution of individual rating scales to a composite score should be monitored throughout test development and validation. Previous empirical studies in other fields addressed the issue of how best to weight different components to achieve optimum reliability or to maximize the degree of match between actual vs. desired weighting of different components in a composite as defined in test specifications (e.g., Jarjoura & Brennan, 1982, 1983; Li, 1990; Kane & Case, 2004; Marcoulides, 1994). However, it is perhaps fair to say that this important topic has not received the attention that it deserves in the language assessment literature so far.

Considering the critical importance of investigation into convergent/discriminant validity and the interrelationships among analytic rating scales and a composite, the dearth of empirical studies that address effective weights of analytic rating scales is quite puzzling. A primary reason for this may be that previous investigations of L2 performance assessments have made extensive use of univariate analytic approaches that do not offer the full machinery to address this issue. Univariate analytic approaches such as univariate G theory and unidimensional IRT models (e.g., many-facet Rash measurement) assume, by definition, the presence of a *single* construct of interest. For this reason, in a typical FACETS analysis, for example, a single ability estimate is obtained per examinee, while analytic rating scales themselves are specified as a facet of measurement (i.e., part of the test method). This conceptualization of analytic rating scales does not seem to be fruitful, however, because what one is communicating loud and clear by employing such scales is the presence of more than one ability of interest. Multivariate analytic approaches allow not only specification of multiple constructs but also an investigation of convergent/discriminant validity among analytic rating scales,

which in turn serves as the basis for exploring the relationships of the scales to a composite score.

III The language ability assessment system (LAAS) Spanish test

In this study, data from the Speaking section of the Language Ability Assessment System (LAAS) Spanish test were analyzed. The LAAS is a criterion-referenced assessment of Spanish for academic purposes developed at the University of California, Los Angeles (UCLA) for assessing the readiness of University of California (UC) students who had already been admitted to a study-abroad program, the Education Abroad Program (EAP), sponsored by the University. There were two main purposes of the LAAS test: (1) to place EAP admits into either a full academic immersion program or a sheltered language program, and (2) to provide candidates with diagnostic feedback on their academic Spanish language ability in four modalities: reading, listening, speaking and writing. In this instrument, the construct, Spanish language ability, is defined as multicomponential. Accordingly, a single score is reported for each modality and, in addition, score profiles based on analytic rating scales for speaking and writing.

The LAAS design rationale and the test development procedure were discussed by Bachman, Lynch and Mason (1995). In order to serve the needs and context of its use in the EAP program, the test developers conducted an informal needs assessment, where they collected information about language needs in the EAP program from both EAP faculty members and previous attendees (Bachman, Lynch, Mason & Egbert, 1992).

The entire LAAS Spanish test was approximately two hours long and was administered via videotape in language laboratories. One unique feature of this test was the use of a common theme across all four sections. This test design reflected the results of the needs analysis, which suggested that EAP students were often required to process the same content in more than one modality by, for example, writing or speaking about what they had read and heard in their reading assignments and lectures (Bachman, Lynch & Mason, 1995). To simulate a real academic context, in Part 1 (Reading) candidates read materials directly related to the lecture to be presented in Part 2 and answered short-answer reading comprehension questions. In Part 2 (Listening), they first watched an introductory lecture delivered in simple Spanish, to provide some context for the academic lecture that

followed. After that, candidates watched a 10–12 minute segment of an actual academic lecture videotaped at an institution abroad and then responded to short-answer listening comprehension questions. In Part 3 (Speaking), candidates completed two speaking tasks based on the content of the academic lecture given in Part 2. Finally, in Part 4 (Writing), candidates had an opportunity to write an essay, which required them to integrate what they had read in the Reading section and what they had heard in the Listening section, and relate the information to their major field of study or their personal life.

The Speaking section consisted of two role-play tasks. Candidates were instructed to imagine that they were visiting the professor who delivered the introductory lecture in Part 2 and that the professor would ask them to first, summarize the lecture in their own words, and second, to elaborate on a point discussed in the lecture by relating it to their own experiences. For each of the speaking tasks, candidates were given one minute for preparation and three minutes to respond to the task. The candidates spoke into tape recorders to record their speech samples.

Reflecting the multicomponential definition of speaking ability in the LAAS, the scoring rubric for the Speaking section consisted of five analytic rating scales: Pronunciation, Vocabulary, Cohesion, Organization and Grammar (see Appendix A for the rating scales). Among the five rating scales, all except Grammar were rated on a 4-point scale, ranging from 1 (“no evidence”) to 4 (“good”). In contrast, Grammar was on a 7-point scale, ranging from 1 (“no systematic evidence of range and control of few or no structures; errors of all or most possible are frequent”) to 7 (“complete range and no systematic error, just lapses”).

Two independent ratings were obtained on the five analytic rating scales for each candidate’s response to each of the two speaking tasks. When a discrepancy was observed between the two ratings on any examinee response, another rating was provided by a third rater, and the closest two scores out of the three were used for score reporting. Six scores were reported to candidates for the speaking section. Each candidate received separate scores for the five rating scales, each of which was the mean across the four ratings after the adjudication (two independent ratings for each of the two speaking tasks). A composite score, which was the grand mean across all the 20 ratings, was reported as the Overall Speaking score as well.

The placement decisions were made in a non-compensatory manner. The test developers suggested cut scores of 3 (“moderate”) for Pronunciation, Vocabulary, Cohesion and Organization and 4 (“large, but not complete range and control of some structures used, but with

many error types") for Grammar for entering a full immersion EAP. Individual EAP program advisors from each UC campus placed candidates into either sheltered language courses or full immersion programs based on the lowest score on any of the five rating scales, along with any other information. The LAAS placement decisions were rather high stakes. Since sheltered language programs were not available at some hosting institutions, differential placements sometimes meant that students had to move to different universities.

One point to note in this rating scale design is the use of the increased score points for Grammar compared to the others. A primary reason for the test developers' decision to do so was to adequately differentiate among different levels of grammar knowledge. Another important reason was to reflect the EAP Spanish instructors' perception of the relative importance of grammar knowledge for EAP admits' success in the program (Bachman, personal communication, 2006). Despite this, the intended effective weights of the analytic rating scales in the Overall Speaking score were not specified during the test development process. This is perhaps because the actual placement decisions were based not on the Overall Speaking score but on the analytic rating scales. However, investigating the effective weights of the analytic rating scales in the composite is still crucial for monitoring the functioning of the analytic rating scales in relation to the composite score.

Given the design rationale for the LAAS Spanish speaking section above, it is important to empirically investigate the three key issues identified above. Thus, the present study addressed the following research questions:

1) Is the underlying multicomponential trait factor structure assumed in the LAAS Spanish speaking test design supported?
2) How reliable are the LAAS Spanish speaking ratings?
3) How dependable are the high-stakes placement decisions based on the LAAS Spanish speaking test?
4) Do the empirical contributions of each of the LAAS ratings scales to the composite score variance differ?

V Method

1 Participants

The data for the speaking section of the LAAS Spanish test obtained from 214 EAP admits who participated in the operational Spring

1993 administration were analyzed in the present study. Most of the participants were sophomores from eight UC campuses, where about 75% of them were going to attend the study-abroad program in Spain, and the remaining 25% in Mexico. This is essentially the same data analyzed by Bachman, Lynch and Mason (1995). However, whereas those researchers analyzed only the Grammar rating scale, the present study included all five of the analytic scales for the Speaking section. Since the data from the third ratings were not available to this study, only the first two ratings were included in the analyses.

2 Raters

The raters were 15 graduate students and faculty members at the Department of TESL/Applied Linguistics and the Department of Spanish and Portuguese at UCLA. All of them were native or near-native speakers of Spanish. Bachman et al. (1992) described the rater training process employed for the speaking portion. The training began with raters familiarizing themselves with the project and the rating scales, which included reviewing the test procedure and directions as well as studying and discussing the rating scales. This was followed by norming of the raters, where each of them independently rated six speaking tapes at home and then reconvened to discuss the rating scales further. Each rater then rated four additional tapes for discussion at the final norming session for the speaking section.

3 Data analysis methods

Two multivariate analytic approaches—confirmatory factor analysis (CFA) and multivariate G theory—were combined in this study in order to address the research questions above. First, CFA was employed in order to test relative goodness of fit of CFA models that offer competing explanations of the structural relationships among the five rating scales. In this study, a special type of CFA model for multitrait-multimethod (MTMM) analysis (Jöreskog, 1974; Marsh, 1988, 1989; Marsh & Grayson, 1995; Widaman, 1985) was employed. The CFA approach to MTMM, which was also applied to language assessment by Bachman and Palmer (1981, 1982), Llosa (2005) and Sawaki (2003) is currently the most commonly used alternative to Campbell and Fiske's (1959) original

MTMM analysis based on an inspection of an observed correlation matrix for a set of measures.

The second approach employed was multivariate G theory (Cronbach et al., 1972; Brennan, 2001). As a broad analytic framework that subsumes the univariate theory previously applied to language assessment studies, a multivariate G theory analysis yields all the information that is available in a univariate analysis, including variance component estimates for different sources of score variation and various summary indices of score dependability. The additional information available in multivariate G theory that is particularly relevant to this study is the interrelationships among a set of analytic rating scales as well as a comprehensive composite score analysis. In conventional approaches not based on G theory (e.g., Wang & Stanley, 1970), effective weights are obtained to investigate the extent to which individual measures account for the *observed* composite score variance, which contains both true-score and error variances. However, the particular advantage of the multivariate G theory approach to the composite score analysis is that the effective weights of analytic rating scales can be obtained separately for the parts of the composite score variance contributing to the true score variance (composite universe-score variance) and different types of measurement error (e.g., composite absolute-error variance for a criterion-related score interpretation; see Brennan 2001, pp. 305–306). Thus, for example, effective weights of analytic rating scales for a composite true-score variance tell us how much information they contribute to differentiate among examinees based on their true differences in a given ability represented by a composite.

In a sense, CFA and multivariate G theory yield overlapping information associated with the research questions. However, combining the two approaches is advantageous for this study because of the strengths of these approaches in different areas. First, with regard to the investigation of convergent/discriminant validity, CFA offers a sequential model testing framework for explicitly supporting or rejecting competing explanations about the relationships among analytic rating scales, while multivariate G theory only allows "eyeballing" of the universe-score correlations among analytic rating scales. Second, because the primary interest of this study is a criterion-referenced interpretation of the LAAS Spanish speaking test for the EAP admits, it is useful to have information available about the score dependability estimates for absolute decisions applicable to criterion-referenced assessment in multivariate

G theory. Finally, the comprehensive composite score analysis available in multivariate G theory adequately addresses the empirical weighting of the LAAS analytic rating scales in a composite score.

In the CFA-based MTMM analysis in this study, the covariance matrix for the 20 scores obtained for each EAP Spanish examinee (scores on two tasks rated by two raters on each of the five rating scales) was analyzed. Each CFA model included five latent (unobservable) factors associated with the five analytic rating scales (Pronunciation, Vocabulary, Cohesion, Organization and Grammar) and four latent factors related to the measurement design: two for the two ratings, and two for the two speaking tasks. A series of models that depicted different relationships among these latent factors were tested to primarily address Research Questions 1 and 2. Following the procedure suggested by Rindskopf and Rose (1988), this sequential testing proceeded from least restrictive to more restrictive models. Maximum-likelihood (ML) was used as the model parameter estimation method. Multiple criteria below were employed in order to assess the overall goodness of fit of the CFA models:

- Model chi-square statistic: A statistically non-significant model chi-square statistic indicates an adequate model fit.
- *The ratio of model chi-square to model degrees of freedom ($\chi^2_{S\text{-}B}/df$):* Because the model chi-square statistic is sensitive to sample size, this ratio is often used as a model fit criterion. In this study the ratio of 1.5 or below was considered as an indication of good model fit.
- Three incremental fit indices, which compare relative improvement in the explanation of the covariations among the measures in the target model against the baseline model that assumes that the measures are completely uncorrelated (Hu & Bentler, 1995). In this study, the comparative fit index (CFI), Bentler-Bonnet normed fit index (NFI), and the Bentler-Bonnet non-normed fit index (NNFI) of .90 or above were used as indicators of adequate model fit.
- Two goodness-of-fit indices that address model parsimony: Akaike Information Criterion (AIC) and Consistent Version of this Statistic (CAIC). AIC adjusts for the number of parameters estimated, while CAIC takes account of both the number of parameters estimated and sample size (Kline, 1998). The lower

the value, the better the model fit given by the complexity of the model.

- Root Mean Square Error of Association (RMSEA): A RMSEA indicates the extent to which the model approximates the data, taking into account the model complexity. A RMSEA of .05 or below is considered as an indication of good model fit.

All the CFA analyses were conducted using EQS 6.0 Beta (Bentler, 1985–2002).

The multivariate G theory analysis was conducted primarily to address Research Questions 2, 3 and 4. Persons were treated as the objects of measurement (a G theory term for the target of measurement). The five LAAS analytic rating scales were modeled as the fixed facet. Representing the five Spanish speaking abilities of primary interest, the rating scales were not exchangeable with others. Each rating scale was associated with a two-facet crossed design (denoted as *p x r x t*), where persons (*p*) were completely crossed with ratings (*r*) and tasks (*t*), i.e., each person completed both tasks and were rated twice. The ratings were treated as a random facet because the two ratings were considered as samples from a universe of admissible observations, i.e., interchangeable ratings provided by EAP raters, who had similar backgrounds as applied linguists and completed the same EAP rater training. The tasks were also treated as a random facet. Conceptually, treating this facet as fixed may be more appropriate because the two speaking tasks were different in nature. However, the results from this two-facet crossed design are reported here because the results for the mixed effects design that specifies the tasks as the fixed facet can be obtained from the results of the present fully-crossed design. Furthermore, the results of the mixed effect design were almost identical to those of the completely-crossed design.

As a first step, a generalizability study (G study) was conducted in order to estimate the relative contribution of seven sources of score variation in the LAAS ratings for the present G study design (Person ability, Rating severity, Task difficulty, Person by Rating interaction, Person by Task interaction, Rating by Task interaction, and residual) for a situation where only one rater and one task are used for assessment on each rating scale. Then, a decision study (D study) was conducted by setting the numbers of ratings and tasks to two each in order to reflect the actual LAAS test design. In the D study the observed variances and covariances among the five analytic rating scales were decomposed into different parts called variance-covariance component

estimates. The key measures among them were (1) the universe-score variance-covariance component estimates, which represent the part of the observed score variances and covariances attributable to the true ability differences among the persons, and (2) the absolute-error variance-covariance component estimates, which represent the part attributable to all the sources of score variation contributing to the absolute error (Appendix B). These results served as the basis for the analyses of score dependability for the LAAS analytic rating scales and the composite score analysis discussed below. Throughout the composite score analysis, equal weights were assigned to the analytic rating scales (0.2 for all, so that the sum equals 1) to conform to the nominal weights actually used for the calculation of the Overall Speaking score in the operational LAAS. The computer program mGENOVA (Brennan, 1999) was used for the multivariate G theory analyses.

4 Outliers, missing scores and normality of distributions

Out of the 214 examinees, seven who did not complete the entire test and thus had missing ratings for the speaking section were deleted list-wise. Another case had one missing score for an Organization rating that was imputed by using the EM (estimation maximization) algorithm and retained in the subsequent analyses. None of the remaining 207 cases was found to be a univariate outlier on the 20 LAAS ratings, while two were identified as multivariate outliers based on Mahalanobis distance. The listwise deletion of these two cases resulted in a final sample size of 205. An investigation of randomly-selected scatterplots on various combinations of the ratings suggested that the variables were roughly linearly related. Inspection of the histograms as well as standardized skewness and kurtosis values showed that all the distributions were univariate normal except for the four Pronunciation ratings, which were significantly negatively skewed. Bachman et al. (1992) reported the relatively high mean ratings that examinees obtained for the Pronunciation rating scale in an earlier LAAS pilot study as well. Unfortunately, the language background information of the EAP admits in the present sample and the pilot study sample were not available to this study. One possible explanation, given the large Spanish-speaking population in California, is that the EAP admits involved in the pilot test and the 1993 operational test administration might have represented heritage Spanish speakers and/or learners of Spanish who had developed good pronunciation skills with frequent exposure to the Spanish language.[1]

Multivariate normality of the score distribution for all the 20 ratings was examined by the normalized Mardia's coefficient. The value of 17.62 suggested considerable non-normality of the multivariate score distribution. This non-normality of the data was accommodated in the CFA analysis by using the robust statistics available in EQS. Moreover, chi-square difference tests conducted to statistically compare relative fit of competing models were also based on the Satorra-Bentler Scaled chi-square statistic (Satorra & Bentler, 1999). In the multivariate G theory analyses, a regular MANOVA-based procedure for variance-covariance component estimation, as implemented in mGENOVA, was used due to its robustness to nonnormality of score distributions (Brennan, 2001).[2]

VI Results

1 Research Question 1: Is the underlying multicomponential trait factor structure assumed in the LAAS Spanish speaking test design supported by empirical results?

Both the CFA and the multivariate G theory analysis yielded relevant information to address this research question. In the multivariate G theory analysis, the universe-score correlations, i.e., the G theory analogue of true-score correlations, of the five LAAS analytic rating scales were obtained in the D study. As can be seen in Table 1, the universe-score correlations were extremely high, ranging from .85 to .98, especially those among Vocabulary, Cohesion and Grammar, suggesting that an examinee that scored high on one of these three rating scales tended to score high on the other two as well. Although this

---

[1]Another possibility might be the relative leniency of the raters on the Pronunciation rating scale. However, this explanation is not straightforward because previous findings on rater behavior regarding their harshness on rating of pronunciation are mixed. Earlier studies reported that ratings that involved pronunciation were not particularly harsh (e.g., McNamara, 1990), while others that compared rater groups with different backgrounds reported that rater harshness on this dimension depended on the rater background (e.g., Brown, 1995).

[2]Previous authors (Jöreskog, 1974; Linn & Werts, 1979; Marcoulides, 1996, 2000; Raykov & Marcoulides, 2006) have indicated that some variance component estimates needed for G theory analyses can be obtained through a CFA, and Schoonen (2005) applied this approach to an analysis of a norm-referenced language assessment. However, the results based on the CFA approach are not reported in this paper due to technical difficulties associated with estimation of some variance-covariance components for sources of score variation not involving the objects of measurement. Although the use of the MANOVA-based approach as implemented in mGENOVA with the CFA in this study makes the results from the two approaches look "disjoined," the CFA and ANOVA-based variance-covariance component estimation produce essentially identical results (e.g. Linn & Werts, 1979; Marcoulides, 1996, 2000).

Table 1 Universe-score correlations among the LAAS analytic rating scales[a]

Rating scale	Universe-score correlations				
	Pronunciation	Vocabulary	Cohesion	Organization	Grammar
Pronunciation	1.00				
Vocabulary	0.91	1.00			
Cohesion	0.91	0.98	1.00		
Organization	0.85	0.93	0.94	1.00	
Grammar	0.90	0.97	0.97	0.92	1.00

[a]Based on the D study for 2 ratings and 2 tasks.

result demonstrates the high intercorrelations among the rating scales, multivariate G theory framework does not allow one to systematically test the extent to which the convergent/discriminant validity of the LAAS rating scales was tenable. Thus, this issue was explored further within CFA.

In the CFA approach, relative goodness of fit of CFA models that offered competing explanations about the trait factor structure of the LAAS Spanish speaking section were compared. The path diagram for the initial model (henceforth, Correlated Trait Factor Model, or CTF Model) is shown in Figure 1. This initial model depicted the multicomponential and yet correlated nature of the language ability assessed in the LAAS Spanish speaking section assumed by the test developers. The 20 rectangles in the center of Figure 1 represent the 20 observed variables, i.e., the 20 LAAS ratings awarded to each candidate as all possible combinations of the five rating scales, two ratings and two tasks. The ovals in the diagram represent latent factors that are hypothesized to predict examinees' observed scores. The five ovals to the left are for the traits of interest: Pronunciation, Vocabulary, Cohesion, Organization and Grammar. The four ovals to the right are for the latent factors associated with the test method: the two ratings and the two tasks. In this model, each of the 20 observed variables was specified as related to one trait factor and two method factors. For example, the first observed variable was the Pronunciation rating by Rater 1 on Task 1 (labeled as PROR1T1 in the figure). That is, the rating on this variable can be predicted by a candidate's ability on the Pronunciation rating scale as well as the severity of Rating 1 and the difficulty of Task 1. These predictive relationships between the latent factors and the observed variable are thus indicated by the arrows in the figure and were estimated as path coefficients. Another important feature of this model is the interrelationships specified among the five trait factors. This was done to reflect the

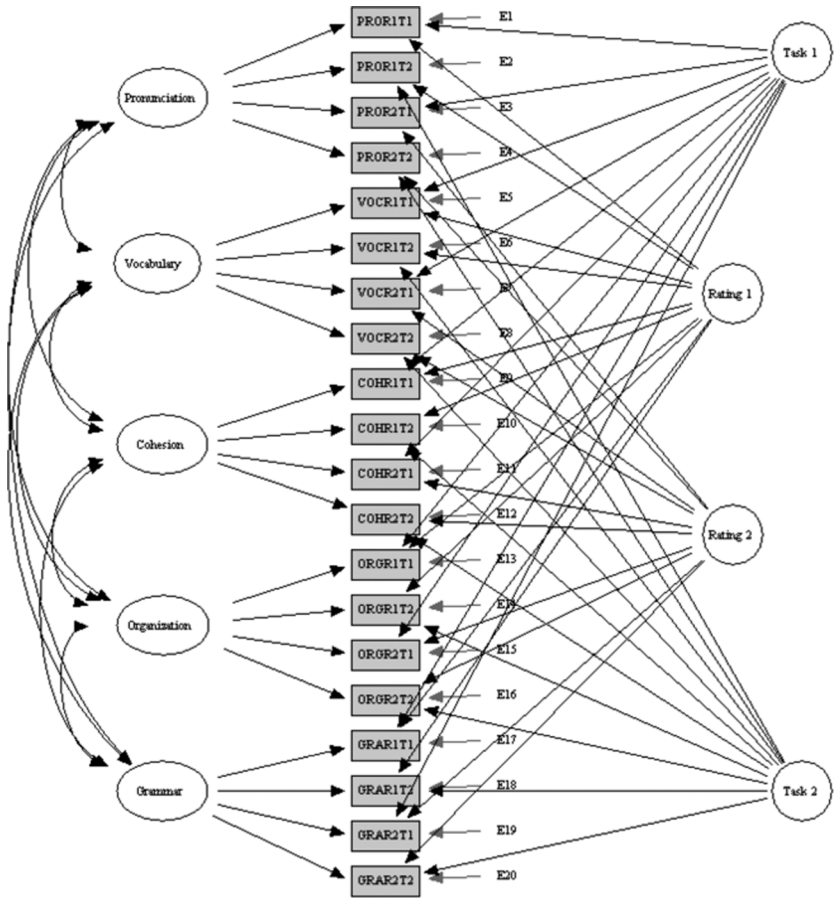

Figure 1 Initial CFA model (Correlated Trait Factor Model)

expected substantial correlations among the rating scales, which was also suggested by the universe-score correlations among the rating scales mentioned above. The correlations among the trait factors indicated by the curved two-headed arrows connecting pairs of the trait factors were freely estimated.

The reader may notice that this model does not fully represent the situation of the LAAS where a composite score and ratings on the five rating scales are reported because a trait factor that represents the Overall Speaking ability is "missing" from this diagram. A model that better corresponds to this situation is the Higher-order Trait Factor Model (or HTF for short) to be discussed in detail later in this section. The ultimate goal is to demonstrate a satisfactory fit of that

model; however, the CTF Model was introduced first in this study because a satisfactory fit of the CTF model was the prerequisite for an adequate fit of the more restrictive HTF model.

The obtained model fit indices based on the robust estimation for the CTF Model are shown in Table 2. The Satorra-Bentler Scaled chi-square statistic for the CTF Model was statistically significant ($df = 120$; $\chi^2 = 203.89$; $p < .05$), but this model showed an excellent fit to the data. The obtained values of the CFI, NFI and NNFI met the pre-determined criteria of model fit laid out in Section V, while the RMSEA and the ratio of the Satorra-Bentler Scaled chi-square to the degrees of freedom were slightly worse than the suggested criteria. Thus, overall, this result supported the distinct and yet correlated nature of the traits as defined by the five rating scales.

Although the satisfactory fit of the CTF Model partially supports the convergent/discriminant validity of the rating scales, a stronger test is needed to falsify alternative explanations for the underlying interrelationships among the rating scales. For this reason, three alternative CFA models with varying trait factor structures were developed, while keeping the method factor structure constant. The specifications of the three alternative models, all being more restrictive versions of the CTF Model and nested with the CTF Model, are as follows:

Orthogonal Trait Factor (OTF) Model: This model specified the five traits representing the rating scales as being uncorrelated with

Table 2 Assessment of CFA model fits

Model description	Correlated Trait Factor (CTF) Model	Orthogonal Trait Factor (OTF) Model	Unitary Trait Factor (UTF) Model	Higher-Order Trait Factor (HTF) Model
Model df	120	130	130	125
Normal theory chi-square	232.01	668.08	406.39	237.44
Satorra-Bentler chi-square	203.89	624.86	371.41	209.14
S-B chi-square/df	1.70	4.81	2.86	1.67
CFI	0.98	0.89	0.95	0.98
NFI	0.96	0.87	0.92	0.96
NNFI	0.97	0.84	0.92	0.97
AIC	−36.11	364.86	111.41	−40.86
CAIC	−554.87	−197.13	−450.58	−581.24
RMSEA	0.06	0.14	0.10	0.06
RMSEA CI	.04–.07	.13–.15	.08–.11	.04–.07

Mardia's normalized multivariate kurtosis: 17.62.

one another. In order to demonstrate the convergent validity of the LAAS rating scales, the CTF model must fit significantly better than this model.

Unitary Trait Factor (UTF) Model: This model specified the five traits representing the rating scales as being essentially indistinguishable from one another. In order to demonstrate the discriminant validity of the LAAS rating scales, the CTF model must show significantly better fit than this model.

Higher-Order Trait Factor (HTF) Model: In this model a higher-order factor structure was imposed on the correlations among the five trait factors in the CTF Model. This model not only specifies the five traits as being intercorrelated but also assumes the presence of an underlying higher-order factor that can account for a common variance across the first-order trait factors representing the five rating scales (Rindskopf & Rose, 1988). This trait factor structure reflects the assumption underlying the policy of reporting a single composite score.

The goodness-of-fit indices for these three alternative models are shown in Table 2. The results suggest that the fit of the UTF and the HTF Models was satisfactory, while that of the OTF Model was poor.

Next, the relative goodness of fit of these three alternative models was compared against that of the CTF model by conducting chi-square difference tests. The results are summarized in Table 3. First, a chi-square difference test showed that the fit of the CTF model was significantly better than that of the OTF model ($\chi_{dff}^2 = 1809.41$, $df = 10$), supporting the convergence of the rating scales. Second, the CTF model fit the data significantly better than the UTF model ($\chi_{dff}^2 = 308.10$, $df = 10$), supporting the hypothesis that the five trait factors are psychometrically distinct from one another. The better fit of the CTF model than those of the OTF and UTF models is also suggested by the considerably low AIC and CAIC values for the CTF model.

Table 3 Chi-square difference test results[a]

Models compared	df	Chi-square difference	Significance ($p < .05$)
CTF vs. OTF	10	1809.41	Significant
CTF vs. UTF	10	308.10	Significant
CTF vs. HOF	5	0.83	Non significant

[a]Based on Satorra-Bentler scaled chi-square with adjustments by Satorra & Bentler (1999).

Regarding the comparison of the CTF and the HTF models, although the chi-square difference test suggested that these models fit the data equally well ($\chi_{dff}^2 = .83$, df = 5), the lower AIC and CAIC values for the HTF model indicated that, when model complexity is taken into consideration, the HTF model fit better than the CTF model. Moreover, the HTF model was substantively more interpretable than the CTF model because the HTF model explicates the relationships among the five rating scales and the overall construct represented by the Overall speaking Score: Spanish speaking ability. For these reasons, the HTF model was selected as the final model (see Figure 2).

The standardized model parameter estimates for the final model (HTF Model) are presented in Table 4. Because they are adjusted for

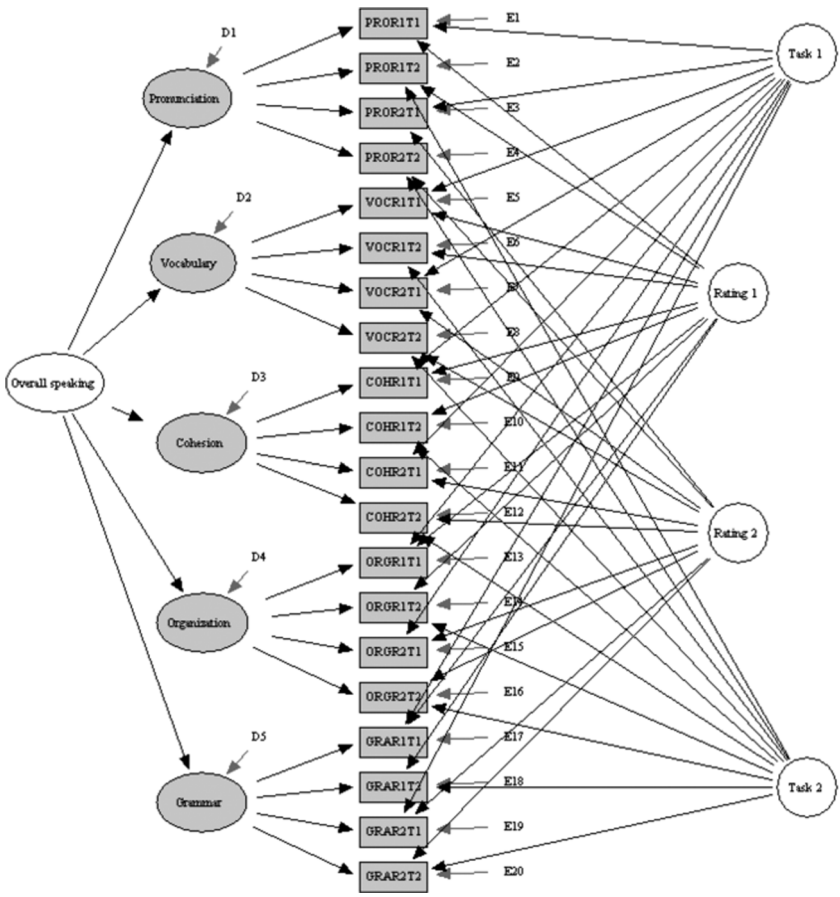

Figure 2 Final CFA model (Higher-order Trait Factor model)

Table 4 Standardized parameter estimates for the final model (HTF Model)

Variable	Trait					Method				Residual	% Variance accounted for
	Pronunciation	Vocabulary	Cohesion	Organization	Grammar	Rating 1	Rating 2	Task 1	Task 2		
Pronunciation											
Rating 1 Task 1	.93**	.00*	.00*	.00*	.00*	.09	.00*	.00	.00*	.35	.88
Rating 1 Task 2	.90	.00*	.00*	.00*	.00*	.01	.00*	.00*	.13	.43	.82
Rating 2 Task 1	.86	.00*	.00*	.00*	.00*	.00*	.33	.15	.00*	.36	.87
Rating 2 Task 2	.84	.00*	.00*	.00*	.00*	.00*	.33	.00*	.04	.42	.82
Vocabulary											
Rating 1 Task 1	.00*	.91**	.00*	.00*	.00*	.21	.00*	−.08	.00*	.36	.87
Rating 1 Task 2	.00*	.91	.00*	.00*	.00*	.09	.00*	.00*	.27	.31	.91
Rating 2 Task 1	.00*	.89	.00*	.00*	.00*	.00*	.18	.27	.00*	.32	.90
Rating 2 Task 2	.00*	.82	.00*	.00*	.00*	.00*	.49	.00*	.15	.25	.94
Cohesion											
Rating 1 Task 1	.00*	.00*	.93**	.00*	.00*	.21	.00*	−.12	.00*	.30	.91
Rating 1 Task 2	.00*	.00*	.90	.00*	.00*	.07	.00*	.00*	.34	.27	.93
Rating 2 Task 1	.00*	.00*	.88	.00*	.00*	.00*	.22	.23	.00*	.35	.88
Rating 2 Task 2	.00*	.00*	.85	.00*	.00*	.00*	.42	.00*	.13	.30	.91

Organization											
Rating 1 Task 1	.00*	.00*	.00*	.86**	.00*	.20	.00*	−.02	.00*	.47	.78
Rating 1 Task 2	.00*	.00*	.00*	.88	.00*	.14	.00*	.00*	.22	.40	.84
Rating 2 Task 1	.00*	.00*	.00*	.84	.00*	.00*	.16	.31	.00*	.42	.82
Rating 2 Task 2	.00*	.00*	.00*	.82	.00*	.00*	.34	.00*	.06	.47	.78
Grammar											
Rating 1 Task 1	.00*	.00*	.00*	.00*	.86**	.48	.00*	.10	.00*	.15	.98
Rating 1 Task 2	.00*	.00*	.00*	.00*	.90	.24	.00*	.00*	.14	.28	.92
Rating 2 Task 1	.00*	.00*	.00*	.00*	.95	.00*	.12	.16	.00*	.24	.94
Rating 2 Task 2	.00*	.00*	.00*	.00*	.95	.00*	.17	.00*	.03	.26	.93

Regression from higher-order to lower-order factors

	Speaking	*Disturbance*	*% Variance explained*
Pronunciation	.90	.44	.81
Vocabulary	.99	.13	.98
Cohesion	.99	.15	.98
Organization	.94	.34	.89
Grammar	.96	.30	.91

*Factor loading fixed to zero.
**Factor loading fixed to one for scale identification.
Significant factor loadings are underscored in the table.

scale differences, the path coefficients are directly comparable among themselves as indicators of the strengths of relationships between the factors and the observed variables as well as between the first-order factors and the higher-order factor. The path coefficients of the observed ratings to the corresponding first-order trait factors presented in the left half of the table were high and significantly different from zero, ranging from .82 to .95. Moreover, the path coefficients of the five first-order trait factors on the higher-order speaking factor presented at the bottom right of Table 4 were extremely high, ranging from .90 to .99. These results indicate strong linear relationships between the first-order trait factors and the observed ratings, and between the higher-order factor and the first-order factors, respectively.

In summary, the multicomponential and yet highly correlated nature of the traits being assessed in the LAAS Spanish speaking test was confirmed by the selection of the Higher-order Trait Factor Model. The final CFA model included a single higher-order factor strongly associated with the five LAAS rating scales, suggesting the presence of a substantial common underlying dimension. Moreover, the support for the final model that specified separate factors for the individual rating scales provided support for the distinctness of the five rating scales.

2 Research Question 2: How reliable are the LAAS Spanish speaking ratings?

a Reliability of the individual rating scales Three pieces of information from the CFA and the multivariate G theory analysis provided empirical evidence on the reliability of the LAAS Spanish speaking test for the individual rating scales. The first was the relative magnitudes of the path coefficients for the trait factors as opposed to those for the method factors. As mentioned in the previous section, the path coefficients of the observed variables for the five first-order trait factors in the HTF model presented in Table 4 were generally high, while those for the method factors presented in the upper right half of Table 4 were low to moderate (−.12 to .49). This suggests that the observed LAAS ratings were strongly associated with the trait factors, whereas their relationships with the method factors were relatively weak. It is worth noting, however, that there was considerable variation in the magnitudes of the path coefficients for the method factors across the 20 LAAS ratings. That is, some of the path coefficients for the rating and task factors were sizable and statistically significantly different from zero (e.g., the

rating path coefficient of .49 and the task path coefficient of .15 for Rating 2 for Task 2 on Vocabulary), while others were not (e.g., the rating path coefficient of .01 and the task path coefficient of .13 for Rating 1 for Task 2 on Pronunciation). This indicates that the individual LAAS ratings were affected by the method factors to different degrees.

Second, the relative contribution of the different sources of score variation to the LAAS rating variability was examined by means of the D study variance component estimates for the individual analytic rating scales obtained in the G theory analysis. As can be seen in Table 5, variance component estimates for the Persons were by far the largest across the rating scales, with more than 88% of the proportions of the scale variances accounted for by the ability differences among the persons. In contrast, the proportions of variance accounted for by the facets of measurement were uniformly very small: only those for the Person by Rating and Person by Task interactions, ranging from .2% to 7.1%, were of any size. In contrast, all the others except the residual variance components were virtually zero across all the rating scales. The non-zero variance component estimates for the Person by Rating and Person by Task interactions show that there were differences in the rank-orderings of the candidates

Table 5 Estimated variance components and proportion of variance accounted for by facets of measurement[a]

Source of variation	Estimated variance component				
	Pronunciation	Vocabulary	Cohesion	Organization	Grammar
Person (p)	0.42	0.48	0.52	0.41	1.79
	88.0%	91.6%	91.5%	88.7%	93.2%
Rating (R)	0.00	0.00[b]	0.00	0.00	0.00
	0.1%	0.0%	0.0%	0.0%	0.0%
Task (T)	0.00	0.00	0.00	0.00	0.00
	0.1%	0.2%	0.3%	0.3%	0.2%
p × R	0.03	0.02	0.02	0.02	0.08
	7.1%	3.1%	3.4%	4.0%	4.0%
p × T	0.01	0.00	0.00	0.00	0.01
	1.1%	0.5%	0.8%	0.2%	0.5%
R × T	0.00[b]	0.00[b]	0.00	0.00[b]	0.00[b]
	0.0%	0.0%	0.0%	0.0%	0.0%
pRT,e	0.02	0.02	0.02	0.03	0.04
	3.4%	4.7%	4.0%	6.7%	2.0%

[a]Based on the D study for 2 ratings and 2 tasks.
[b]Negative variance component fixed to zero after calculation.

across first and second ratings as well as across the tasks. The nonzero residual variance components across the rating scales suggest the presence of (1) Person by Rating by Task interaction, (2) sources of variability due to error that was not captured by the present two-facet crossed D study design, or both.

Finally, two types of summary indices on the reliability of LAAS analytic ratings were obtained, each based on the CFA and the multivariate G theory analysis (see Table 6). The first is the intraclass reliability coefficient (e.g., Bae & Bachman, 1998; Bagozzi, 1991; Werts, Linn & Jöreskog, 1974) obtained from the path coefficients of the observed ratings for the trait and method factors in the final CFA model. In general, high path coefficients for a trait factor as opposed to low path coefficients for method factors of observed variables associated with a given trait factor resulted in a high intraclass reliability estimate for that trait factor. Overall, the intraclass reliability estimates were high, ranging from .83 to .89. Second, the index of dependability (phi coefficient denoted as Φ; Brennan & Kane, 1977a, b), for the rating scales were obtained in the D study in the multivariate G theory analysis. As can be seen in Table 6, the five rating scales were ranked similarly by the two methods, while the dependability coefficients from the multivariate D study were consistently higher than the corresponding intraclass reliability coefficients. In both types of coefficients, however, the estimates for the Grammar rating scale were the highest of all, while the estimates for the Pronunciation and Organization rating scales were the lowest.

Taken together, the dependability of the individual LAAS analytic rating scales was supported by the three pieces of evidence: (1) the relatively high path coefficients for the trait factors vs. the relatively low path coefficients for the method factors of the observed ratings in the CFA analysis; (2) the large proportions of variances of the individual rating scales accounted for by the Person effect in the multivariate G theory analysis; and (3) the generally high intraclass reliability indices as well as the phi coefficients for the rating scales.

Table 6 CFA intraclass reliability and multivariate G theory dependability estimates

	Pronunciation	Vocabulary	Cohesion	Organization	Grammar
Intraclass rel.	0.85	0.86	0.87	0.83	0.89
Phi[a]	0.88	0.92	0.92	0.89	0.93

[a]Based on the D study for 2 ratings and 2 tasks.

b Reliability of the composite score The composite score analysis based on the D study in the multivariate G theory analysis yielded the index of dependability (phi coefficient, Φ, for the LAAS Overall Speaking score) as well. The phi-coefficient for the composite score discussed here is an extension of the index of dependability, which was applied to individual rating scales in addressing Research Question 3, to the analysis of the composite score, as described by Brennan (2001). It is a function of (1) the composite universe-score and absolute-error variances obtained from the universe-score and absolute-error variance-covariance components for the five rating scales and (2) the nominal weights given to the rating scales (0.2 for all in this case). The obtained composite universe-score and absolute-error variances with the equal nominal weights were .62 and .04, respectively. The phi coefficient for the composite, which is the ratio of the composite universe-score variance to the sum of itself and the composite absolute-error variance was high (Φ=.95, i.e., .62/(.62+.04) =.95). This suggests that as much as 95% of the variance in the composite was accounted for by the universe score variance, or the score variance due to the true differences among examinees in terms of their language ability.

3 Research question 3: How dependable are the placement decisions made based on the LAAS Spanish speaking test?

This research question was addressed by examining estimates of phi-lambdas (Φ_λ) for the individual rating scales obtained in the D study within the multivariate G theory analysis. A phi-lambda is an agreement index for the dependability of decisions made at a predetermined cut score (Brennan & Kane, 1977a, b). The cut scores recommended by the test developers for the placement decision-making and the phi-lambda for each rating scale are presented in Table 7. As can be seen in the table, the phi-lambdas for the individual rating scales (.89 to .94) were generally high, suggesting the high dependability of decisions made at the cut scores on the rating scales. Although still satisfactory, the relatively lower phi-lambda's

Table 7 Dependability of decisions at predetermined cut scores[a]

	Pronunciation	Vocabulary	Cohesion	Organization	Grammar
Cut score	3	3	3	3	4
Phi-lambda	0.89	0.92	0.91	0.89	0.94

[a]Based on the D study for 2 ratings and 2 tasks.

for the Pronunciation and Organization rating scales correspond to the relatively lower dependability of the rating scales associated with the large variance component estimate for Person by Rating interaction on the Pronunciation rating scale and the large variance component estimate for residuals on the Organization rating scale, respectively.

4 Research Question 4: Do the empirical contributions of each of the LAAS rating scales to the composite score variance differ?

Effective weights of the individual rating scales were obtained separately for the composite universe-score and absolute-error variances as part of the D study in the multivariate G theory analysis. The effective weight of a given rating scale for the composite true-score (or absolute-error) variance is determined by (1) the nominal weight for the rating scale, (2) universe-score (or absolute-error) variance of the rating scale, and (3) the universe-score (or absolute-error) covariances of the rating scale with the others (Brennan, 2001). Table 8 shows the effective weights of the five rating scales for the universe score variance and the absolute error variance. The results showed that the Grammar rating scale accounted for as much as 33.63% of the composite universe-score variance, while the other four explained only 15.37% (Pronunciation) to 18.13% (Cohesion). The contribution of the Grammar rating scale to the composite absolute-error variance (29.20%) was the largest as well, while those of the others ranged from 16.91% (Pronunciation) to 18.40% (Organization). These results suggest that the Grammar rating scale contributed relatively more information to both the composite universe-score and absolute-error variances compared to the other rating scales.

Table 8 Composite score analysis results[a]

Contributions to:	Variance and covariance components				
	Pronunciation	Vocabulary	Cohesion	Organization	Grammar
a priori weights	0.20	0.20	0.20	0.20	0.20
Effective weights contributing to:					
Universe score variance (%)	15.37	17.37	18.13	15.49	33.63
Absolute error variance (%)	16.91	17.58	17.90	18.40	29.20

[a]Based on the D study for 2 ratings and 2 tasks.

VII Discussion and conclusions

The results of the CFA and the multivariate G theory analyses in this study answered all the four research questions affirmatively. A few points are worthy of more discussion.

First, the present study provided some empirical support for the view that the highly correlated and yet multicomponential nature of language ability is tenable not only for language abilities across different modalities, as found in previous factor analytic studies (e.g., Bachman, Davidson, Ryan & Choi, 1995; Bachman & Palmer, 1981, 1982; Bae & Bachman, 1998; Kunnan, 1995; Llosa, 2005; Shin, 2005), but also for language measures (analytic rating scales) within a single modality. One explanation for the extremely high intercorrelations among the LAAS analytic rating scales indicated in the multivariate G theory analysis is the overlap of the constructs across the analytic rating scales as shown in the scoring rubrics given to the raters (See Appendix A). For instance, the universe score correlation between the Vocabulary and Cohesion ratings (.98) was the highest of all. On the one hand, the criteria suggested to the raters for assigning Cohesion ratings were coordination, subordination, reference and topicalization, all of which are realized by appropriate use of words in appropriate contexts. On the other hand, the rating guide for vocabulary draws the attention of raters to three points: use of false cognates, code switching and sophistication of lexical choice. Thus, a candidate who demonstrated sophistication of lexical choice in marking cohesion and coherence, for example, might have received high ratings on both cohesion and vocabulary.

In addition, given that the raters were allowed to assign ratings for all five rating scales at the same time in the LAAS rating process, we cannot rule out the possibility that the correlations among the scales were further inflated due to a halo effect. That is, the raters might have had difficulty in differentiating among the abilities assessed by the rating scales and thus awarded similar scores across them. Further revisions of the subscales and/or a change in the rating design (e.g., allowing raters to provide only one rating at a time) might change the patterns of interrelationships among the subscales at least to some extent, provided such a halo effect is in fact present in the LAAS ratings.

The complementary use of CFA in the present study allowed us to not only confirm the convergent validity of the LAAS rating scales suggested in the multivariate G theory analysis above, but also go a step further to examine the discriminant validity of the scales by the sequential testing of competing models that offered alternative explanations about the interrelationships among the scales. The adoption of

the Higher-order Trait Factor model as the final model provided empirical support for the psychometric distinctness of the constructs assessed by the LAAS rating scales, despite the high correlations among themselves.

Second, the notable contribution of multivariate G theory was to examine the relative weighting of analytic rating scales through composite score analysis. As mentioned above, one unique feature of the LAAS Spanish speaking assessment is the use of more score points for the Grammar rating scale than those for the others. Because the use of a longer Grammar rating scale may lead to a larger empirical weight given to the scale, this decision may appear controversial when one considers the debate surrounding the role of grammar in speaking performance assessments as exemplified by the criticisms of the ACTFL Oral Proficiency Interview (OPI) for placing undue weight on structural accuracy in the rating criteria (e.g., Bachman & Savignon 1986; Savignon, 1985) and the attempts to de-emphasize structural accuracy in L2 speaking performance assessment design (McNamara, 1990, 1996). Moreover, in a recent needs analysis survey given to U.S. undergraduate and graduate faculty and students on academic English ability (Rosenfeld, Leung, & Oltman, 2001), structural accuracy of speaking while performing various academic language use tasks received generally low ratings on its relative importance for successful completion of academic courses and the academic success of nonnative speakers of English.

However, the LAAS rating scale design itself does not tell us the whole story as to whether Grammar indeed played an overriding role in the placement decisions. As mentioned above, the EAP staff adopted a non-compensatory placement approach. Because the decision based on the lowest score point on *any* rating scale does not allow a high score on one scale to compensate for a low score on another, this system itself does not take account of the relative importance of the five components. The decision to use the non-compensatory approach was made based on a criterion-group study with second-year EAP students and those who had just completed the program (Bachman, personal communication, 2006).

Alternatively, if EAP wants to explore ways for the placement decisions to better reflect the relative importance of grammar knowledge as perceived by the EAP faculty members, a compensatory approach based on the Overall Speaking score can be pursued. In this case effective weights such as those obtained in the present study would help EAP monitor the functioning of the analytic rating scales. If the current equal nominal weights for the five components are

maintained, the program staff would know that the Grammar scale accounts for about 33% of the true-score variance of the Overall Speaking score. Moreover, if EAP concludes that the effective weight of 33% for Grammar is too high or too low, new nominal weights that allow them to obtain the desired effective weights of the five components could be calculated by using a weight re-adjustment procedure such as Bachman's (2005).

The combined use of CFA and multivariate G theory in this study allowed a focused investigation of some critical issues associated with the construct validity of L2 speaking performance assessments involving analytic rating scales—score dependability, convergent/discriminant validity of analytic rating scales, and the empirical relationship of analytic rating scales to a composite. While the LAAS Spanish speaking test was in use in the EAP program in the early 1990s and was discontinued afterwards, the use of analytic rating scales for student placement and diagnosis in order to better link assessment to instruction is well aligned with the current growing interest in diagnostic language assessments (Alderson, 2005a, b; Alderson & Huhta, 2005). Accordingly, many validation issues surrounding the LAAS are applicable to various current and future L2 performance assessments, and thus this line of investigation should be an integral part of the development as well as construct validation of L2 speaking performance assessments.

Acknowledgements

An earlier version of this paper was presented at the 24th Annual Language Testing Research Colloquium held in Hong Kong in December 2002. This study was completed while the author was a doctoral student at the University of California, Los Angeles. Special thanks are due to Lyle F. Bachman, Noreen Webb and George A. Marcoulides for their guidance throughout the completion of this study. In addition, the author thanks the three anonymous reviewers, as well as Lorena Llosa and Emily Midouhas, for their comments on earlier versions of the paper.

VIII References

Alderson, J.C. 2005a: The challenge of (diagnostic) testing: Do we know what we are measuring? Plenary paper presented at the 27th Annual Language Testing Research Colloquium. Ottawa, Canada, 20–22 July.

—— 2005b: *Diagnosing foreign language proficiency: The interface between learning and assessment.* London: Continuum.

Alderson, J.C. and **Huhta, A.** 2005: The development of a suite of computer-based diagnostic tests based on the Common European Framework. *Language Testing* 22, 301–20.

Bachman, L.F. 2005: *Statistical analyses for language assessment.* Cambridge: Cambridge University Press.

Bachman, L.F., Davidson, F., Ryan, K. and **Choi, I.-C.** 1995: *An investigation into the comparability of two tests of English as a foreign language: The Cambridge-TOEFL comparability study.* Cambridge: Cambridge University Press.

Bachman, L., Lynch, B. and **Mason, M.** 1995: Investigating variability in tasks and rater judgments in a performance test of foreign language speaking. *Language Testing* 12, 238–57.

Bachman, L., Lynch, B., Mason, M. and **Egbert, M.** 1992: *EAP Language Ability Assessment System: Interim report*. Los Angeles: University of California, Los Angeles, Department of TESL and Applied Linguistics.

Bachman, L.-F. and **Palmer, A.** 1981: The construct validation of the FSI oral interview. *Language Learning* 31, 67–86.

—— 1982: The construct validation of some components of communicative proficiency. *TESOL Quarterly* 16, 449–65.

Bachman, L.F. and **Savignon, S.J.** 1986: The evaluation of communicative language proficiency: A critique of the ACTFL Oral Interview. *The Modern Language Journal* 70, 380–97.

Bae, J. and **Bachman, L.F.** 1998: A latent variable approach to listening and reading: Testing factorial invariance across two groups of children in the Korean/English two-way immersion program. *Language Testing* 15, 380–414.

Bagozzi, R.-P. 1991: Further thoughts on the validity of measures of elation, gladness and joy. *Journal of Personality and Social Psychology* 61, 98–104.

Bentler, P.-M. 1985–2002: EQS for Windows 6.0 Beta Build 94 Version. [Computer software]. Encino, CA: Multivariate Software, Inc.

Brennan, R.L. 1999: mGENOVA, Version 2.0. [Computer software].

—— 2001: *Generalizability theory.* New York: Springer-Verlag.

Brennan, R.L. and **Kane, M.T.** 1977a: An index of dependability for mastery tests. *Journal of Educational Measurement* 14, 277–89.

—— 1977b: Signal/noise ratios for domain-referenced tests. *Psychometrika* 42, 609–25.

Brown, A. 1995: The effect of rater variables in the development of an occupation-specific language performance test. *Language Testing* 12, 1–15.

Brown, J.-D. and **Bailey, K.M.** 1984: A categorical instrument for scoring second language writing skills. *Language Learning* 34, 21–42.

Campbell, D.T. and **Fiske, D.W.** 1959: Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin* 56, 81–105.

Carroll, J.B. 1983: Psychometric theory and language testing. In Oller, J.W., Jr., editor, *Issues in language testing research,* Rowley, MA: Newbury House, 80–107.

Cronbach, L.J., Gleser, G.C., Nanda, H. and **Rajaratnam, N.** 1972: *The dependability of behavioral measurements: Theory of generalizability for scores and profiles.* New York: Wiley.

Elder, C. 1993: How do subject specialists construe classroom language proficiency? *Language Testing*10, 235–54.

Hu, L. and **Bentler, P.M.** 1995: Evaluating model fit. In Hoyle, R.H., editor, *Structural equation modeling: Concepts, issues, and applications.* Thousand Oaks, CA: Sage Publications, 76–99.

Jacobs, H.L., Zinkgraf, S.A., Wormuth, D.R., Hartfiel, V.F. and **Hughey, J.B.** 1981: *Testing ESL composition: A practical approach.* Rowley, MA: Newbury House.

Jarjoura, D. and **Brennan, R.** 1982: A variance components model for measurement procedures associated with tables of specifications. *Applied Psychological Measurement* 6, 161–71.

—— 1983: Multivariate generalizability models for tests developed from tables of specification. In Fyans, L.J., Jr., editor, *Generalizability theory: Inferences and practical applications*, San Francisco, CA: Jossey-Bass, 83–101.

Jöreskog, K.G. 1974: Analyzing psychological data by structural analysis of covariance matrices. In Atkinson, R.C., Krantz, D.H., Luce, R.D. and Suppes, P., editors, *Contemporary developments in mathematical psychology: Measurement, psycho-physics, and neural information processing* (Vol. 2), San Francisco: Freeman, 1–56.

Kane, M. and **Case, S.M.** 2004: The reliability and validity of weighted composite scores. *Applied Measurement in Education* 17, 221–40.

Kline, R.B. 1998: *Principles and practice of structural equation modeling*. New York: Guilford Press.

Kondo-Brown, K. 2002: A FACETS analysis of rater bias in measuring Japanese second language writing performance. *Language Testing* 19, 3–31.

Kunnan, A.J. 1995: *Test taker characteristics and test performance: A structural modeling approach.* Cambridge: Cambridge University Press.

Lee, Y.-W. 2005: *Dependability of scores for a new ESL speaking test: Evaluating prototype tasks.* (TOEFL Monograph Series, MS-28). Princeton, NJ: Educational Testing Service.

Lee. Y.-W. and **Kantor, R.** 2005: *Dependability of new ESL writing test scores: Evaluating prototype tasks and alternative rating schemes.* (TOEFL Monograph Series, MS-31). Princeton, NJ: Educational Testing Service.

Li, W. 1990: *Multivariate generalizability of hierarchical measurement.* Unpublished doctoral dissertation, University of California, Los Angeles.

Linacre, J.M. 1989: *Many-facet Rasch measurement.* Chicago: MESA Press.

Linn, R.L. and **Werts, C.E.** 1979: Covariance structures and their analysis. In Traub, R., editor, *New directions for testing and measurement: Methodological development* (No. 4). San Francisco: Jossey-Bass 53–73.

Llosa, L. 2005: *Building and supporting a validity argument for a standards-based classroom assessment of English proficiency.* Unpublished PhD dissertation, University of California, Los Angeles.

Lumley, T. and **McNamara, T.** 1995: Rater characteristics and rater bias: Implications for training. *Language Testing* 12, 54–71.

Lynch, B. and **McNamara, T.** 1998: Using G theory and Many-facet Rasch measurement in the development of performance assessments of the ESL speaking skills of immigrants. *Language Testing* 15, 158–80.

Marcoulides, G.A.1994: Selecting weighting schemes in multivariate generalizability studies. *Educational and Psychological Measurement* 54, 3–7.

—— 1996: Estimating variance components in generalizability theory: The covariance structure analysis approach. *Structural Equation Modeling* 3, 290–99.

—— 2000: *Advances and new developments in G-theory: Applications of structural equation modeling.* Workshop presented at the 22nd Language Testing Research Colloquium, Vancouver, Canada, March 8–11.

Marsh, H.W. 1988: Multitrait-multimethod analysis. In Keeves, J.P., editor, *Educational research methodology, measurement, and evaluation: An international handbook,* Oxford: Pergamon, 570–80.

—— 1989: Confirmatory factor analysis of multitrait-multimethod data: Many problems and a few solutions. *Applied Psychological Measurement* 13, 335–61.

Marsh, H.W. and **Grayson, D.** 1995: Latent variable models of multitrait-multimethod data. In Hoyle, R.H., editor, *Structural equation modeling: Concepts, issues and applications,* Thousand Oaks, CA: Sage, 177–98.

McNamara, T. 1990: Item response theory and the validation of an ESP test for health professionals. *Language Testing* 7, 52–75.

—— 1996: *Measuring second language proficiency.* London: Longman.

Petersen, N.S., Kolen, M.J. and **Hoover, H.D.** 1989: Scaling, norming, and equating. In Linn, R.L., editor, *Educational Measurement*, third edition. New York: American Council on Education and Macmillan, 221–62.

Pollitt, A. and **Hutchinson, C.** 1987: Calibrating graded assessments: Rasch partial credit analysis of performance in writing. *Language Testing* 4, 72–92.

Raykov, T. and **Marcoulides, G.A.** 2006: Estimation of generalizability coefficients via a structural equation modeling approach to scale reliability evaluation. *International Journal of Testing* 6(1), 81–95.

Rindskopf, D. and **Rose, T.** 1988: Some theory and applications of confirmatory second-order factor analysis. *Multivariate Behavioral Research* 23, 51–67.

Rosenfeld, M., Leung, S. and **Oltman, P.K.** 2001: *The reading, writing, speaking, and listening tasks important for academic success at the undergraduate and graduate levels.* (TOEFL Monograph Series, MS-21). Princeton, NJ: Educational Testing Service.

Satorra, A. 1990: Robustness issues in structural equation modeling: A review of recent developments. *Quality & Quantity* 24, 367–86.

Satorra, A. and **Bentler, P.** 1999:. A scaled difference chi-square test statistic for moment structure analysis. *UCLA Statistics Series, No. 260,* University of California, Los Angeles.

Savignon, S.J. 1985: Evaluation of communicative competence: The ACTFL Provisional Proficiency Guidelines. *The Modern Language Journal* 69, 129–42.

Sawaki, Y. 2003: *A comparison of summarization and free recall as reading comprehension tasks in web-based assessment of Japanese as a foreign language.* Unpublished doctoral dissertation, University of California, Los Angeles.

—— 2005: The generalizability of summarization and free recall ratings in L2 reading assessment. *JLTA Journal* 7, 21–44.

Schoonen, R. 2005: Generalizability of writing scores: an application of structural equation modeling. *Language Testing* 22, 1–30.

Shavelson, R.-J. and **Webb, N.-M.** 1991: *Generalizability theory: A primer.* Newbury Park, CA: Sage Publications.

Shin, S.-K. 2005: Did we take the same test? Examinee language proficiency and the structure of language tests. *Language Testing* 22, 31–57.

Wang, M.W. and **Stanley, J.C.** 1970: Differential weighting: A review of methods and empirical studies. *Review of Educational Research* 4, 663–705.

Webb, N.-M., Shavelson, R.-J. and **Maddahian, E.** 1983: Multivariate generalizability theory. In Fyans, L.J., Jr., editor, *Generalizability theory: Inferences and practical applications*, San Francisco, CA: Jossey-Bass, 67–81.

Weigle, S.-C. 1998: Using FACETS to model rater training effects. *Language Testing* 15, 263–87.

Werts, C.E., Linn, R.L. and **Jöreskog, K.G.** 1974: Intraclass reliability estimates: Testing structural assumptions. *Educational and Psychological Measurement,* 34, 25–33.

Widaman, K.F. 1985: Hierarchically nested covariance structure models for multitrait-multimethod data. *Applied Psychological Measurement* 9, 1–26.

Wigglesworth, G. 1993: Exploring bias analysis as a tool for improving rater consistency in assessing oral interaction. *Language Testing* 10, 305–335.

Wright, B.D. and **Masters, G.N.** 1982: *Rating scale analysis.* Chicago: MESA Press.

Xi, X. 2003: *Investigating language performance on the graph description task in a semi-direct oral test.* Unpublished doctoral dissertation, University of California, Los Angeles.

Appendix A: LAAS Spanish speaking test: Ratings of language ability in speaking

Pronunciation

The pronunciation rating should reflect the subject's accuracy of intonation, stress, and segmental sounds and appropriateness of pace.

1. NO EVIDENCE: Pronunciation problems make speech virtually incomprehensible, or the subject speaks too little to judge.
2. POOR: Pronunciation problems are numerous and interfere with comprehension.
3. MODERATE: Pronunciation problems sometimes interfere with communication.

4. GOOD: Pronunciation may mark the subject as a non-native speaker, but does not interfere with communication.

Vocabulary

The vocabulary rating should reflect the subject's range and accuracy of vocabulary used. Consideration should be given to the following:

Use of false cognates
Code switching (where *not* appropriate)
Sophistication of lexical choice

[N.B. These specific elements may vary from language to language.]

1. NO EVIDENCE: *Extremely limited vocabulary.* **Range:** Vocabulary limited to a few words and set phrases; **Communication:** not possible for subject to discuss topic due to limited vocabulary.
2. POOR: *Small vocabulary.* **Range:** Limited range of lexicon, evidenced by frequent repetition of Spanish words; **Accuracy:** inaccurate usage of Spanish words and/or frequent use of false cognates; **Communication:** difficult for subject to discuss topic because of vocabulary limitations.
3. MODERATE: *Vocabulary of moderate size.* **Range:** Some variety of word choice; **Accuracy:** Some Spanish words may be inaccurately used; occasionally relies on false cognates and/or English words; **Communication:** frequently misses or searchers for Spanish words.
4. GOOD: *Large vocabulary:* **Range:** Uses a variety of lexical choices, demonstrating a wide range; **Accuracy:** choice almost always appropriate; **Communication:** seldom misses or searches for words.

Cohesion

The cohesion rating will be made in terms of 1) inclusion of appropriate means for marking relationships among utterances and ideas and 2) variety of appropriate means used. Means for marking relationships among utterances include, but are not limited to, the following:

Coordination (e.g., English: and, but, also; Spanish: pero, y, o, que)
Subordination (e.g., English: although, while, since; Spanish: cuando, donde, aunque, si, que)

Reference (e.g., English: these, this, those, such; Spanish: este, ese, aquel)
Topicalization of information (passive, pseudo-cleft, etc.)

[N.B. These specific elements may vary from language to language.]

Cohesion scale levels

1. NO EVIDENCE: Utterances completely disjoined, or the discourse is too short to judge.
2. POOR: Connections between utterances not adequately marked; frequent confusing relationships among ideas.
3. MODERATE: Connections among utterances sometimes adequately marked; sometimes confusing relationships among ideas.
4. GOOD: Uses a variety of appropriate means for connecting utterances; hardly ever confusing relationship among ideas.

Organization

The organization rating will be made in terms of the inclusion, distinctness and appropriateness of sequencing of the following parts:

A) Identification of topic
B) Development of topic
   i. Distinguishing of main points
   ii. Appropriate sequencing of main points
   iii. Supporting details
C) Conclusion/closure

Organization scale levels

1. NO EVIDENCE of conscious attempt to organize presentation, or discourse too short to judge
2. POOR: Parts poorly developed; main points not clearly distinguished and not appropriately sequenced.
3. MODERATE: Some parts appropriately developed; main points sometimes clearly distinguished sometimes and appropriately sequenced.
4. GOOD: Parts appropriately developed; main points clearly distinguished and appropriately sequenced.

Grammar

The grammar rating should reflect both the morphological and syntactic structures used by the subject. The levels of grammatical competence are defined in terms of two aspects:

1) *RANGE* of morphologic and syntactic structures and
2) *ACCURACY* or degree of control of morphological and syntactic structures.

Grammar scale levels

RANGE		ACCURACY
1 No systematic evidence	AND	Control of *few or no* structures; errors of all or most possible are frequent.
2 Limited range, but with systematic evidence	AND	Control of *few or no* structures; errors of all or most possible are frequent.
3 Limited range, but with systematic evidence	AND	Control of *some* structure used; with *many* error types.
4 Large, but not complete range	AND	Control of *some* structure used; but with *many* error types.
5 Large, but not complete range	AND	Control of most structure used, with few error types.
6 Complete range	AND	Control of *most* structure used, with *few* error types.
7 Complete range	AND	No systematic errors, just lapses.

NB: If you feel the candidate's grammatical range and accuracy are split across levels more than accommodated in the above ratings, give an "s" rating, and indicate the split. For example, if the candidate is a "2" in range and "4" in accuracy, you would give a rating of "3s" and indicate that the split is "2/4".

Appendix B: D study universe-score and absolute-error variance-covariance component estimates[a]

Source of variation	Variance and covariance components				
Universe score	Pronunciation	Vocabulary	Cohesion	Organization	Grammar
Pronunciation	.416				
Vocabulary	.404	.476			
Cohesion	.421	.487	.519		
Organization	.351	.412	.434	.408	
Grammar	.776	.898	.935	.784	1.791
Absolute error	Pronunciation	Vocabulary	Cohesion	Organization	Grammar
Pronunciation	.056				
Vocabulary	.024	.044			
Cohesion	.022	.032	.048		
Organization	.019	.023	.025	.051	
Grammar	.024	.029	.028	.041	.130

[a]Based on the D study for 2 ratings and 2 tasks.