EDUCATIONAL AWAKENING

Journal of the Educational Sciences

International Islamic University Malaysia

Volume 3 , Number 2

2006

INTERACTION BETWEEN TEST-TAKER CHARACTERISTICS, TASK FACETS AND L2 ORAL PROFICIENCY TEST PERFORMANCE

Noor Lide Abu Kassim, Ainol Madziah Zubairi*

Abstract

The nature of test-taker characteristics and their impact on second language test performance has drawn a considerable amount of research interest among language testers in the last two decades. This is hardly surprising given the potential influence of test-taker attributes on second language (L2) learning and assessment. However, examining the influence of these attributes on test performance alone is insufficient. There are other elements in the assessment framework which not only influence the variability of test scores but which also yield a significant amount of influence on test-taker reactions towards certain tasks and test performance. It is, therefore, necessary to further investigate the nature of these other factors in the assessment framework and examine how they relate to test-taker reactions and how they impact test performance. In this study, the interaction between three main factors in the assessment framework–test task facets, test-taker personal attributes, and language ability–and their simultaneous effect on oral proficiency test performance are examined through the use of structural equation modeling which allows the effects of the different factors in the assessment framework to be directly observed and modeled, and their individual effect on test performance estimated.

* Dr. Noor Lide Abu Kassim is a lecturer and Dr. Ainol Madziah Zubairi is an assistant professor at the Centre for Language and Pre-University Academic Development (CELPAD), International Islamic University Malaysia.

Introduction

Some of the common areas of interest or interfaces in Second Language Acquisition (SLA) and Language Testing (LT) are in describing and explaining the variability in language acquisition and test performance (Bachman & Cohen, 1998). Much research has been conducted and an array of sources of variability have been identified and classified. It is also recognized that sources of variability central to language acquisition and test performance are not only made up of differences that are due to individuals' language abilities but also differences in individuals' personal attributes, the strategies employed in performing particular tasks and differences in the test tasks and context (Bachman & Cohen 1998).

In the context of oral proficiency testing, the influence of these potential sources of variability particularly those related to test task and test-taker characteristics and their impact on test performance, among other things, have also been studied (for example, Brown, 1993; O'Loughlin, 2001; Fulcher, 1996; Lazarton, 1996; Hill, 1998; O'Sullivan, 2002; Elder et al., 2002). The results of these studies are somewhat varied but in many ways encouraging as we believe they have helped resolved some of the important issues in oral proficiency testing and have also given rise to some observations that need to be deliberated upon.

In relation to task difficulty, much has been studied to investigate the effect of task difficulty on performance in oral proficiency testing. Fulcher (1996) investigated the reactions of test takers to different task types (a picture description, an interview and a group discussion) in an attempt to look at, among others, the relationship between a range of test-taker-related facets (including anxiety) and test takers' performance and reactions. The study reported that the affective factor related to anxiety was associated with the interlocutor, presence of equipment in the test situation, and self-confidence. Anxiety was also reported to be associated with task type. It was found that test task involving group discussions generated less anxiety, and was perceived to be the least difficult.

Research into task difficulty in speaking tests that utilizes test-taker feedback (e.g., in Elder et al., 2002) has also acknowledged that various factors might affect perceived task difficulty. Skehan (1998a, 1998b) suggests the different categories that affect difficulty which include familiarity of information and complexity of the task. It was also found that task difficulty affected performance in the speaking test in terms of fluency, accuracy and the complexity of language used.

One important observation that has consistently come up in most studies in performance tests is that no single dimension of a task or facet in the assessment setting could be attributed to the difficulty of a task and hence test performance (e.g. conclusions made by O'Loughlin, 2001; Elder et al., 2002; Brindley & Slatyer, 2002). It is "the particular combinations of item characteristics" (Brindley & Slatyer, 2002, p. 387) and the "complex and unstable interactions between task features and different test-taker attributes" (Elder et al., 2002, p. 364) that "either accentuate or attenuate the effect of difficulty" (Brindley & Slatyer, 2002, p. 387). Hence, there is a need to further explore the nature of the interaction of these elements in the assessment setting and examine the relationships that exist between them.

This study is an attempt to investigate the nature and the strength of relationships that exist between task characteristics (namely, topical knowledge and examiner/interlocutor) and test taker affective variables (namely, level of confidence and test anxiety) and their simultaneous effect on test performance. Specifically this study seeks to answer the following research questions:

- Do test taker affective reactions have a direct effect on test performance?
- Does perceived task difficulty have a direct impact on test performance?
- How does perceived task difficulty affect test takers' affective reactions towards the test?

- To what extent do test taker affective reactions and perceived task difficulty affect test performance?

Though a number of studies have been conducted to investigate the relationships between test taker attributes, task characteristics, perceived task difficulty and test performance, very few have actually investigated the nature and the strength of the relationships or associations that exist between these elements. The present study is an attempt to gather a better understanding of the interrelated relations between test-taker affective reactions and test task characteristics and their impact on test performance.

METHOD

The Speaking Tasks

The speaking tasks (Short Talk, Question Time and Extended Conversation) used in the study are part of the actual speaking test which is part of a placement battery administered by the International Islamic University Malaysia (IIUM). Examinees were first gathered in groups for a twenty-minute briefing where the three tasks were explained, and a few topics for a 'short talk' were given for students to choose before they entered the speaking test venues. In the examination venue, examinees were first asked to deliver a short presentation on the chosen topic followed by a 'question time' session. The third task is a one-to-one interview where the interlocutor engaged candidates in an extended discourse. Each candidate spent about 15-20 minutes on the three tasks.

The Sample

The sample for this study consisted of 100 new-intake students for the 2003/04 academic year. 79% of these students were Malaysians while the remainders were international students who came from various countries (mainly from Maldives, Indonesia, Bangladesh and China). Of the 100 students, 46 were males while 54 were females.

The students who took the speaking placement test had earlier taken a general proficiency test which is Part 1 of the placement test battery. The mean score of the sample on the test was 64.1, with a standard deviation of 10.4 and a range of 50 to 90 points. Therefore, students in the sample consisted of those who are in the upper intermediate level of proficiency. One of the main limitations of the study is that the sample was comprised of those who scored above 50% in the English Placement Test (EPT) Part 1. In view of that, the findings have to be treated with caution so as not to be generalized to examinees of lower language ability groups.

The Questionnaire

The questionnaire instrument is a Likert-type scale questionnaire with seven possible responses to each of the items. The scale ranged from 1 (strongly disagree) to 7 (strongly agree) including a mid-point 'neutral' category. The self-report questionnaire was in part adapted from Fulcher (1996) with a few items taken from O'Loughlin (2001) and in part developed by the researchers. The first part of the questionnaire asked for some personal details. Part 2 and 3 asked candidates about their feelings and impressions regarding the three test tasks in the speaking test. The final part asked some questions concerning aspects related to oral proficiency test tasks in general (task-related, student-related, and interlocutor-related questions).

Data Collection Procedure

The subjects in this study took the speaking test in May 2003 under actual test conditions. Their performances on the three test tasks were scored separately using a holistic scoring scale. Fourteen examiners were involved in the test and consisted of language teachers and instructors who had had experience in conducting the placement speaking test at least twice.

After the test, examinees were requested to complete the questionnaire constructed for the purpose of the study. The questionnaires were later checked for missing responses. Questionnaires with large number of missing responses, especially those involving items that would be used in the SEM analysis, were excluded from the final data set leaving a final sample size of 100 cases.

Data Analysis Procedure

The method used in this study is the testing of a hypothesized structural model through the use of the Structural Equation Modeling (SEM). The SEM, also referred to as covariance structure analysis or causal modeling, is a set of multivariate, analytic procedures that allows for the investigation of relationships between observed and latent variables, and among latent variables based on substantive theory or previous empirical research (Purpura, 1999).

In the SEM, there are two basic parts involved: the measurement model and structural model. The measurement model examines the relationships between latent variables and their corresponding indicator variables or measured variables, whereas the structural model involves the estimation of regression or path coefficients and determination of whether the relationships indicated by the paths are significant. This method of analysis is preferred over other types of multivariate analyses as it enables the assessment of relationships and paths simultaneously without being confounded by other variables in the model, and it provides fit indices that indicate the model fits a given data set (Tremblay & Gardner, 1995).

RESULTS

Preliminary Statistical Analyses

As in other statistical procedures, data preparation is extremely important and need to be conducted. With regard to SEM data-related problems may affect model-fit (Kline, 1998). Hence, several preliminary statistical analyses were conducted on the data set used in this study. Descriptive statistics were calculated and an internal consistency reliability check was computed. The descriptive statistics indicate that the items in the questionnaire follow the normal distribution as all values for skewness and kurtosis were within acceptable range. Item-total correlations were also calculated and it was found that most of the items had item-total correlations within the range of .02 to .39, which is low. The Cronbach's coefficient alpha for the questionnaire as a whole was .71. Though this value is not as high as we expected, it is still within acceptable range (as indicated in Hair, et al., 1998).

Analysis: The Structural Equation Modeling Approach

In carrying out the analysis, four basic steps were followed: model specification, model identification, model estimation, and testing model fit. As the fit indices showed acceptable model fit, the hypothesized model was not re-specified.

Step 1: Model specification

The model developed in this study is a structural model (Hair et al., 1998) formulated to investigate and test the causal relationships between task facets (namely, topical knowledge and examiner or interlocutor), test taker's attitudes (namely, test-related anxiety and confidence), language ability and oral proficiency test performance. The hypothesized model is in keeping with what has been proposed in the language testing literature that some interaction exists between task attributes and test takers' attitude or affective reactions, and test performance (e.g., Fulcher, 1996; O'Loughlin 2001; Brindley & Slayter, 2002; Elder et al., 2002; O'Sullivan 2002).

Figure 1
Hypothesized model of the relationships between task facets, test taker attitudes, language ability and test performance

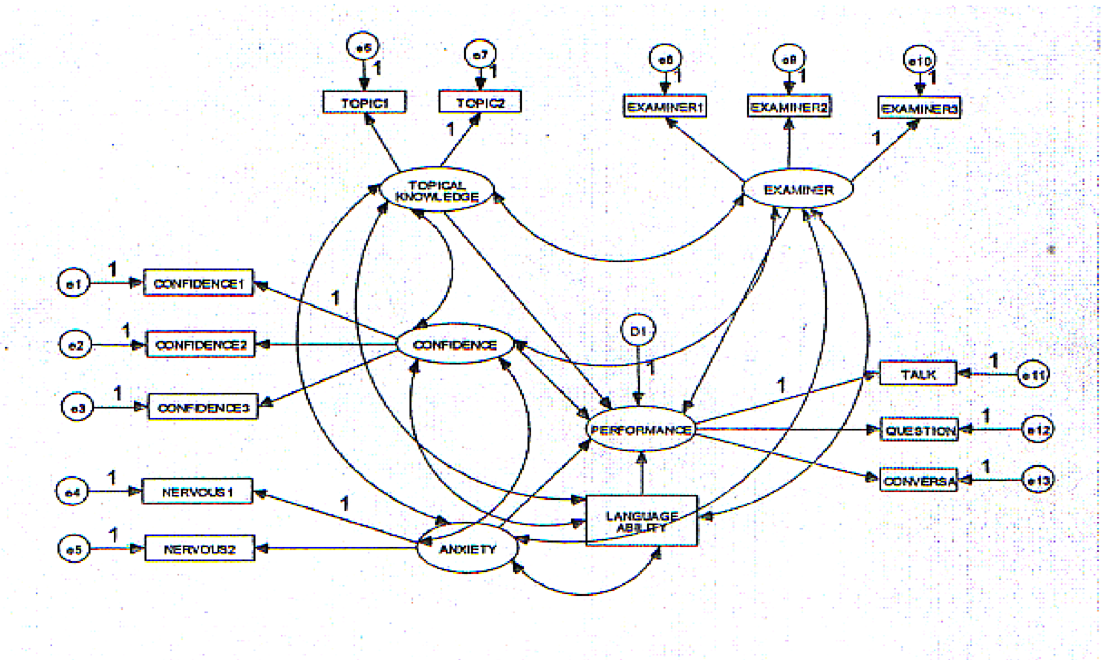


Figure 1 presents the hypothesized model. The latent variables or factors in this model are CONFIDENCE, ANXIETY, EXAMINER, TOPICAL KNOWLEDGE, and PERFORMANCE. Measured variables or observed variables are TOPIC1, TOPIC2, EXAMINER1, EXAMINER2 AND EXAMINER3, CONFIDENCE1, CONFIDENCE2, CONFIDENCE3, NERVOUS1, NERVOUS2, TALK, QUESTION, CONVERSA, and LANGUAGE ABILITY. TOPIC1 – TOPIC2 and EXAMINER1 – EXAMINER3 are individual items used in the questionnaire whereas CONFIDENCE1 – CONFIDENCE3 and NERVOUS1 and NERVOUS2 are composites of the same items which has been consfructed to elicit examinees' responses on the three different test tasks. TALK, QUESTION, and CONVERSA are examinees' test scores on the three speaking test tasks. LANGUAGE ABILITY is examinees' test score on a general English language proficiency test, which they took prior to the speaking test. Table 1 gives the detailed description of the latent and observed (measured) variables in the model.

Table 1
Labels and descriptions of latent and observed variables

Label	Description
TOPIC1	I perform better on topics that I know about
TOPIC2	Talking about topics I know about makes me feel more confident
EXAMINER1	I worry when I have to ask the examiner questions
EXAMINER2	Tasks that require me to ask an examiner questions are difficult
EXAMINER3	Tasks that require me to interact with an examiner are difficult
CONFIDENCE1	I felt confident when I did the short talk/question time/extended conversation
CONFIDENCE2	I liked doing the short talk/question time/extended conversation
CONFIDENCE3	I believe I did well on the short talk/question time/extended conversation
NERVOUS1	I felt nervous before the short talk/question time/extended conversation.
NERVOUS2	I felt nervous while I was doing the short talk/question time/extended conversation
PERFORMANCE	Oral Proficiency test performance
LANGUAGE ABILITY	Language ability /general English language proficiency
CONFIDENCE	Level of confidence
ANXIETY	Test-related anxiety
TOPICAL KNOWLEDGE	Topical knowledge
EXAMINER	Attitude towards examiner/interlocutor

The single-headed arrows from a factor or latent variable (e.g., CONFIDENCE) to a measured variable represent regression coefficients or factor loadings. These indicate the degree to which an underlying factor or latent variable is measured by the observed variables (e.g., CONFIDENCE1 – CONFIDENCE3). Single-headed arrows from one factor (or latent variable) to another factor represent regression coefficients or path coefficients and indicate the impact of one variable on another. For example, the direct effect of CONFIDENCE on PERFORMANCE. A single-headed arrow from an error term (in this case those marked as e1 to e13) to a variable represents the error associated with that variable. In the SEM, error represents much more than random variations in a particular variable due to measurement error. Error also represents a composite of other aspects on which the particular variable may depend, but which was not measured by the variable in the questionnaire (Arbuckle & Wothke, 1999). Disturbance terms or residual error terms "reflect the unexplained variance in the latent endogenous variable/s due to all unmeasured causes" (Garson, 2002, p.3). In this model there is only one disturbance term which is D1. Curved double-headed arrows connect variables that may be correlated with each other (Arbuckle & Wothke, 1999). Examples of these variables are TOPICAL KNOWLEDGE and CONFIDENCE.

Step 2: Model identification

Before a model can be tested, it must be ensured that the model does not have any identification problems (Hair et al., 1998; Tabachnick & Fidell, 2001). This is necessary as only identified models can allow for a unique estimation of every model parameter. If this requirement is not met, estimation of parameter estimates may not be successful (Kline, 1998). The model developed in this study did not have any problems with identification as it was over-identified with 63 degrees of freedom.

Step 3: Model estimation

The Maximum Likelihood Estimation method (MLE) was used for three reasons. The first reason is that the MLE "makes estimates based on maximizing the probability (likelihood) that the observed covariance are drawn from a population assumed to be the same as that reflected in the coefficient estimates. That is, the MLE picks estimates which have the greatest chance of reproducing the observed data" (Garson, 2002, p.3). Secondly, it is the most commonly used method and thirdly, it can be used with sample sizes of between 100 to 150 (Hair et al., 1998).

Step 4: Testing model fit

To test the data fit of the hypothesized model the data set in this study was subjected to the maximum likelihood estimation method using the AMOS (Analysis of Moment Structures) model-fitting program (Arbuckle & Wothke, 1999). Before the structural model was tested, the measurement models were first tested. Results of the measurement models are not reported here, but it must be mentioned that the results are varied. The measurement model related to test-takers' attributes showed acceptable model fit whereas the one related to test task facets did not show a good fit. Despite the poor fit of the second measurement model, we decided to proceed with the testing of the structural model following Loehlin's argument, "If one is in an exploratory mode anyway, there is clearly no mandate that all measurement problems must be resolved completely before any structural problems can be addressed" (Loehlin, 1998, p. 198).

Step 5: Model re-specification

In cases where fit indices do not indicate good model fit, models can be re-specified and subsequently re-tested. In this study, the model was not re-specified as the fit indices suggest acceptable model fit.

Hypotheses Tested in Structural Model

To examine the causal relationship between *Confidence, Anxiety, Topical Knowledge, Examiner, Language Ability* and *Test Performance* the following hypotheses were put forth:

Hypothesis 1
Confidence has no positive effect on Performance

Hypothesis 2
Anxiety has no negative effect on Performance

Hypothesis 3
Topical Knowledge has no positive effect on Performance

Hypothesis 4
Examiner has no negative effect on Performance

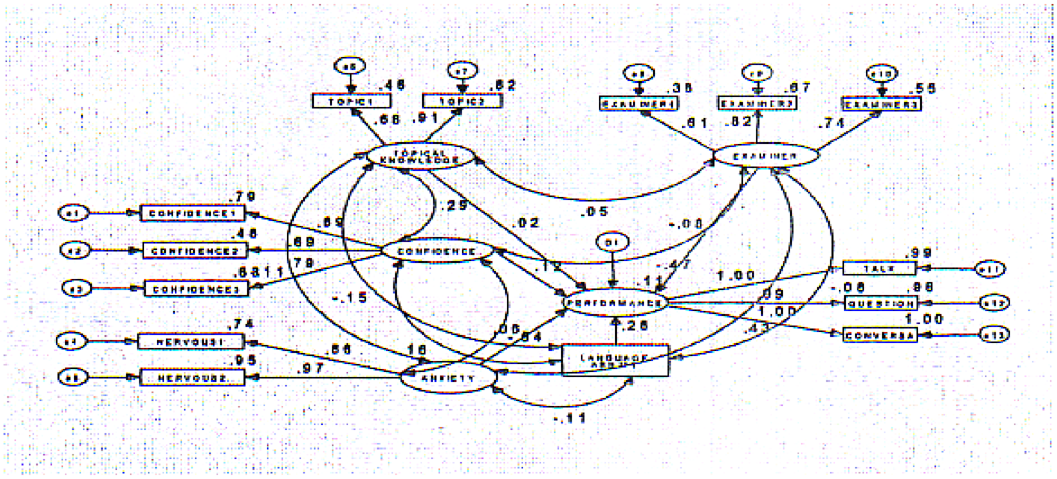
Hypothesis 5
Language Ability has no positive effect on Performance

Table 2 presents the hypothesis of each construct or latent variable association of the model. The results indicate that test-taker attributes (confidence and anxiety) and test task facets (topical knowledge and examiner) have no significant direct effect on Performance. Language ability, on the other hand, indicates a significant effect on test performance. The amount of variance in Performance accounted for by all three variables is 11% (Figure 2). This means that about 89 percent of the variance in Performance is accounted for by other factors that were not included in the model.

Table 2
Hypothesis of Each Latent Variable Association

Construct Associations	Alpha Level	Parameter Estimates	p-value	Significant (Yes/No)
Confidence with Test Performance	0.05	0.11	0.563	No
Anxiety with Test Performance	0.05	0.06	0.665	No
Topical Knowledge with Test Performance	0.05	0.02	0.643	No
Examiner with Test Performance	0.05	-0.08	0.533	No
Language Ability (EPTCORE) with Test Performance	0.01	0.28	0.004	Yes

Figure 2
Standardized parameter estimates of the structural model



As mentioned earlier, apart from investigating the effects of test-taker affective reactions and test task characteristics on test performance, one other aim of this study is to examine the relationships between test task facets and test-taker affective reactions. The results of the analysis (Table 3) indicate that there is a positive and significant correlation between level of anxiety and attitude towards examiner/interlocutor (.43 significant at *p*d” 0.01), and between level of confidence and topical knowledge (.32 significant at *p*d” 0.05). A significant negative correlation between confidence level and attitude towards examiner/interlocutor (-.46 significant at *p*d” 0.01), and between level of confidence and test anxiety (-.63 significant at pd” 0.01) were also found.

Of another interest is the relationship between language ability and level of confidence and anxiety. The results of analysis show that there is a positive but nonsignificant correlation between confidence level and language ability, and as expected, a negative but nonsignificant relationship between anxiety level and language ability. Concerning the relationship between language ability and perceived difficulty related to examiner and topical knowledge, there is a negative but nonsignificant correlation. This, too, perhaps would yield a different outcome given a different data set.

Table 3
Correlation between Test-Taker Characteristics, Test Task Facets and Language Ability

	Estimate	*p*-value
Anxiety and Language Ability	-0.113	0.276
Confidence and Language Ability	0.158	0.148
Examiner and Language Ability	-0.080	0.483
Topical Knowledge and Language Ability	-0.149	0.178
Confidence and Anxiety	-0.642	0.000
Examiner and Topical Knowledge	0.053	0.665
Confidence and Topical Knowledge	0.288	0.019
Anxiety and Topical Knowledge	-0.106	0.348
Confidence and Examiner	-0.474	0.001
Anxiety and Examiner	0.434	0.002

Evaluating the Model Fit

To evaluate the model fit, another set of hypotheses was formulated:

H_0 = data fit the model
H_1 = data does not fit the model

Failure to reject the null hypothesis is desired in SEM as it would indicate that the overall model is able to predict the observed variance-covariance matrix in the dataset (Hair et al., 1998). In other words, failure to reject the null hypothesis would indicate that the overall model has a good fit. To evaluate the model fit, two types of measures were used: absolute fit measures and incremental fit measures. The absolute fit indices determine "the degree to which the overall model (structural and measurement) predicts the observed covariance or correlation matrix" (Hair et al., 1998, p. 654). The incremental fit indices on the other hand "compares the proposed model to some baseline model, most often referred to as the null model, which is a single construct model with all indicators perfectly measuring the construct" (Hair et al., 1998, p. 657). As SEM has "no single statistical test that best describes the strength of the model's prediction" (Hair et al., 1998, p. 653), these indices have to be used together in the evaluation of the overall model fit.

Table 4 presents the fit indices used for the evaluation of the overall model. The Chi-square measure is 65.500, with 63 degrees of freedom, $p > 0.05$. The *p*-value which is above 0.05 indicates an acceptable fit of the model (Hair et al., 1998). The fit indices provide further support for the hypothesized model. The normed chi-square value is within the accepted range ($3.0 > \chi^2/df > 1.0$) and the Root Mean Square Error of Approximation (RMSEA) is below 0.08. The incremental fit measures are all above 0.90 and the Goodness of Fit Index (GFI) is also above the recommended value; however, the GFI is slightly below 0.90 (0.864). In terms of parsimony, the model shows reasonable values.

Table 4
Goodness-of-fit indices of the Overall Model

	Observed	Expected
Absolute Fit Measures	65.500	
Chi square (χ^2) of estimated model	(df=63)	
Significance level (p <)	0.390	$p > 0.05$
Normed Chi-square (χ^2/df)	1.040	$3.0>\chi^2/df>1.0$
Root Mean Square Error of Approximation (RMSEA)	0.020	< 0.08
Goodness of Fit Index (GFI)	0.918	>0.90
Adjusted GFI	0.864	>0.90
Incremental Fit Measures		
Normed Fit Index (NFI)	0.951	>0.90
Incremental Fit Index (IFI)	0.998	>0.90
Tucker-Lewis Index .	0.997	>0.90
Comparative Fit index (CFI)	0.998	>0.90
Parsimony		
Parsimony ratio	0.692	
Parsimony-adjusted NFI	0.658	
Parsimony-adjusted CFI	0.691	

Discussion

In addressing the concerns in LT concerning sources of variability that may affect test performance, a number of studies have been conducted to look at learner factors. This study focuses on the possible overlap of learner variables, factors, or interfaces that may affect performance in a speaking test situation. Other than language ability, learner factors/sources of variability that may affect speaking performance investigated in this study were learner confidence, anxiety, reaction towards tasks and reaction towards examiners.

It was found that the only significant factor that affected speaking test performance (as far as the subjects of the study are concerned) was language ability. However, the results have to be treated cautiously considering the limitations of the study. The first limitation has to do with the truncated sample. Inferences drawn from the results of this study will be limited to populations that share similar characteristics to the sample used in this study as it only included the upper intermediate and advanced students and not students in the low proficiency groups. Secondly, inter-rater reliability between the 14 raters who rated the performance of the subjects could not be determined; hence, the associations between the constructs measured in this study will have to be interpreted with caution. The third limitation pertains to the dependency of the response to the questionnaire. It was found that there was a tendency amongst respondents to overuse the mid-point category (i.e., neutral).

In spite of its limitations, this study has been successful in a number of respects. The first concerns the associations between learner affective factors (namely, anxiety and confidence) and task characteristics (topic and examiners), and language ability. It was found that none of the factors has any significant relationship with examinees' language ability suggesting that examinees' level of test anxiety and confidence in the test situation were not related to their level of language ability.

Secondly, the study allows us to observe the nature of the associations between affective factors and the impact of their influence on test performance. The findings of this study has shown that a relationship exists between affect and test performance though it may not be statiscally significant for the present dataset. This is congruent with findings of past studies. For example, Frierson and Siegel (1984) and Madsen and Murray (1984) found a small but significant variance in test scores that is attributable to test anxiety (as cited in Fulcher, 1996).

Thirdly, this study has also shown that test task characteristics contribute substantially to test-related anxiety and level of confidence. Confidence and anxiety levels were found to be significantly associated with examiner factor. In particular, the more negative their attitude was towards the examiner the higher the tendency to feel anxious and less confident in the task situations. This supports the findings of previous studies which have found that aspects of the test and the testing situation can affect reactions to a test (e.g., Brown, 1993) and, in certain instances, performance.

Additionally, the findings also suggest that the level of one's perceived difficulty of task due to topical knowledge and examiner also appeared to be related to confidence. Nonetheless, an assumption cannot be made about the causal relationship between confidence and one's perception of difficulty, that is, whether perceived difficulty related to topic and examiner were caused by level of confidence or vice versa. In addition, since the subject sample was representative of a population of students of the high language proficiency, inference about the association must again be treated with caution.

On the other hand, it was found that neither anxiety nor confidence affected performance in the speaking test. A possible explanation to this finding is again probably due to the sample of the study since they only constituted examinees in the high level of proficiency. Most likely, the results would not be the same had the study included examinees from the lower proficiency as test anxiety has been reported to affect those with low oral proficiency compared to those with high proficiency (Young, 1991). It is, therefore, suggested that a similar study should be done to include samples from a broader range of language proficiency and that the relationships between these two affective factors be viewed in conjunction with different individuals' language proficiency.

Conclusion

This study is rather exploratory in nature in that it explores the relationships and possible overlaps or interfaces between two different facets of variables that are commonly examined in numerous studies on second language performance assessment (namely, in speaking test situation): affective reactions and test task characteristics. Affective reactions included were examinees' anxiety and confidence in completing the speaking tasks while test task characteristics included were the examiner and topic related facets. In the attempt to see the direct effects of the affective factors and task characteristics on test performance, it was found that neither had any significant association with performance.

The use of SEM allows for further observation in describing the interactions between the task characteristics and learner affective reactions and their attributes in relation to speaking performance and language ability. Indeed, the interactions between the variables were observed but found to be rather 'complex.'

Limitations of the study were addressed and therefore two recommendations are put forward. There is a need to replicate this study in a more controlled situation where rater effects and context effects could be minimized if not eliminated. Replication of the current study with different groups of language learners in different learning contexts is also necessary to allow for generalizations in other L2 settings.

Secondly, there is a need to investigate other competing models. The SEM, despite its sophistication, can only confirm that a particular model fits the given dataset but it cannot disconfirm others (Tabachnick & Fidell, 2001; Hair et al., 1998). Also, we may not have included other more important aspects pertaining to task characteristics that might possibly influence test takersma:' affective reactions and test performance (for example, planning time and nature of response) for a more comprehensive view of the nature of interaction between these elements. After all, the variables taken into consideration in this particular study only accounted for 11% of the variance in performance.

References

Arbuckle, J.L., & Wothke, W. (1999). *Amos 4.0 User's Guide*. Chicago: Small Waters.

Bachman, L.F. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.

Bachman, L.F, & Cohen. (1998). Language testing- SLA interfaces: An update. In Bachman, L.F. and Cohen (Eds.), *Interfaces Between Second Language Acquisition and Language Testing Research*. Cambridge: Cambridge.

Brindley, G., & Slatyer, H. (2002). Exploring Task Difficulty in Listening Task Assessment. *Language Testing, 19 (4)* 369-394.

Brown, A. (1993). The role of test taker feedback in the test development process: test takers' reactions to a tape-mediated test of proficiency in spoken Japanese. Language Testing, 10 (3), 277-303.

Chalhoub-Deville, M. (1995). Deriving oral assessment scales across different tests and rater groups. *Language Testing, 12 (1)*, 16-30.

Clapham, C. (1998). The effect of language proficiency and background knowledge on EAP students' reading comprehension. In Anthony John Kunnan (Ed.), *Validation in Language Assessment.* Mahwah, New Jersey: Lawrence Earlbaum.

Elder C., Iwashita N., & McNamara, T. (2002). Estimating the difficulty of oral proficiency Tasks: What does the test-taker have to offer? *Language Testing, 19(4),* 347-368.

Fulcher, G. (1996). Testing tasks: issues in task design and the group oral. *Language Testing, 13 (1),* 23-49.

Fulcher, G. (1997). Testing tasks: issues in task design and the group oral. *Language Testing, 13 (1),* 23-49.

Gardner, R.C., & MacIntyre, P.D. (1993). On the measurement of affective variables in second language learning. *Language Learning, 43(2),* 157-194

Garson, G.D. (2002). *PA 765: Structural Equation Modeling.* Retrieved November 14, 2002, from http://www2.chass.ncsu.edu/garson/pa765/structur.htm

Hair, J.F., Anderson, R.E., Tatham, R.L., & Black, W.C., (1998). *Multivariate Data Analysis* (5th Edition). New Jersey: Prentice Hall.

Hill, K. (1998). The Effect of Test-Taker Characteristics on Reactions to and Performance on an Oral English Proficiency Test. In Kunnan, A.J (Ed.), *Validation in Language Assessment.* New Jersey: Erlbaum.

Kline, R.B. (1998). *Principles and Practice of Structural Equation Modeling*, New York: Guildford Press.

Kunnan, A.J. (1998). An introduction to structural equation modeling for language assessment research, *Language Testing*, Vol. 15, 295-332.

Lazaraton, A. (1996). Interlocutor support in oral proficiency interviews: the case of CASE, *Language Testing,13,*151-72.

Linacre, J.M. (1989). *Many-faceted Rasch Measurement.* Chicago IL: MESA Press.

McNamara, T. (1996). *Measuring Second Language Performance.* London: Longman O'Loughlin, K.J. (2001). *The Equivalence of Direct and Semi-direct Speaking Tests*. Cambridge: Cambridge University Press.

O' Sullivan, B. (2002). The impact of gender in oral proficiency testing. *Language Testing,19 (2),* 169-192.

Purpura, J.E. (1999). *Learner strategy use and performance on language tests:A structural equation modeling approach*. Cambridge: Cambridge University Press.

Purpura, J. (2002). Validating questionnaires to examine personal factors in L2 test performance: A confirmatory approach. Retrieved November 10, 2002, from http://dbm.hct.ac.ae/public/events/01_02/ctelt2002/jamespurpura.html

Saville, N. (2000). Research Notes Issue 1 (March). Retrieved January 23, 2003, from. http://www.cambridge_efl.org/rs_notes/0001/rs_notes1_4.cfm.

Skehan, P. (1998a). *A Cognitive Approach to Language Learning*. Oxford: Oxford University Press.

Skehan, P. (1998b). Processing perspectives to second language development, instruction, performance and assessment. London: Thames Valley Working Papers in Applied Linguistics, 4, 70-88.

Spurling, S., & Ilyin, D. (1985). The impact of learner variables on language test performance. *TESOL Quarterly, 19 (2),*28.

Tabachnick, B.G., & Fidell, L.S. (2001). *Using Multivariate Statistics.* Needham Heights, MA: Allyn and Bacon.

Tremblay, P.F., & Gardner, R.C. (1995). Expanding the motivation construct in language learning. *The Modern Language Journal*, *79*, 505-518.

Upshur, J.A., & Turner, C.E. (1999). Systematic effects in the rating of second-language speaking ability: Yest method and learner discourse. *Language Testing, 16 (1),* 8-111.

Van Lier, L. (1989). Reeling, writhing, drawling, stretching, and fainting in coils: Oral Proficiency interviews as conversations. *TESOL Quarterly 23,* 489-508.

Young, D. (1991). The relationship between anxiety and Foreign language oral proficiency ratings. *Language Anxiety: From Theory and Research to Classroom Implications* (pp.57-63). In Horwitz and Young (Eds), Englewood Cliff, NJ: Prentice-Hall.

Zeidner, M. (1988). Sociocultural differences in examinees' attitude towards scholastics ability exams. *Journal of Educational Measurement*, *25,* 67-75.