

Test-takers' performance appraisals, appraisal calibration, state-trait strategy use, and state-trait IELTS listening difficulty in a simulated IELTS Listening test

Authors: Aek Phakiti, The University of Sydney, Australia

Grant awarded: 2014

Keywords: "Appraisals, confidence, calibration, IELTS Listening test, cognitive and metacognitive strategies, state and trait, international students, structural equation modeling, Rasch Item Response Theory, quantitative method"

Abstract

This study investigates the nature of test-takers' appraisal confidence and its accuracy (calibration), reported trait and state strategy use and IELTS Listening difficulty levels in a simulated IELTS Listening test.

Appraisal calibration denotes a perfect relationship between appraisal confidence in test performance success and actual performance outcome. Calibration indicates an individual's monitoring accuracy. The study aims to examine four aspects theoretically related to IELTS Listening test scores: (1) test-takers' trait (i.e., generally perceived) and state (i.e., context-specific) cognitive and metacognitive strategy use for IELTS Listening tests; (2) test-takers' calibration of appraisal confidence for each test question (i.e., single-case confidence) and for entire test sections (i.e., relative-frequency confidence); (3) trait and state test difficulty perception in IELTS Listening tests; and (4) test difficulty and test-takers' ability as key factors affecting the above variables.

The study recruited 376 non-English speaking background (NESB) international students in Sydney, Australia. Quantitative data analysis techniques including Rasch Item Response Theory, Pearson-Product-Moment correlations, *t*-tests, analysis of variance (ANOVA), and structural equation modeling (SEM) were used.

It was found that test-takers were miscalibrated in their performance appraisals, exhibiting a tendency to be overconfident across the four test sections. Their appraisal calibration scores were found to be worst for very difficult questions. Gender and academic success variables were also examined as factors affecting test-takers' calibration. The SEM analysis conducted suggests that there are complex structural relationships among test-takers' appraisal confidence, calibration, trait and state cognitive and metacognitive strategy use, IELTS Listening difficulty, and IELTS Listening performance.

The study has advanced our knowledge of strategic processes, including appraisal calibration and strategy use that affect IELTS Listening test performance. The outcomes of the study can inform IELTS by providing empirical evidence of the reasons for test score variation among different success levels. Recommendations for future research are discussed.

Publishing details

Published by the IELTS Partners: British Council, Cambridge English Language Assessment and IDP: IELTS Australia © 2016. This online series succeeds *IELTS Research Reports Volumes 1–13*, published 1998–2012 in print and on CD. This publication is copyright. No commercial re-use. The research and opinions expressed are of individual researchers and do not represent the views of IELTS. The publishers do not accept responsibility for any of the claims made in the research. Web: www.ielts.org

AUTHOR BIODATA

AEK PHAKITI

Aek Phakiti is Associate Professor in TESOL at The University of Sydney. His research focuses on language testing and assessment, second language acquisition, and research methods in language learning. He is the author of *Strategic Competence and EFL Reading Test Performance* (Peter Lang, 2007), *Experimental Research Methods in Language Learning* (Bloomsbury, 2014), and, with Carsten Roever, *Quantitative Methods for Second Language Research: A Problem-solving Approach* (Routledge, forthcoming) and *Language Testing and Assessment* (Bloomsbury, forthcoming, 2018).

With Brian Paltridge, he has edited *Continuum Companion to Research Methods in Applied Linguistics* (2010, Continuum) and *Research Methods in Applied Linguistics: A Practical Resource* (2015, Bloomsbury). With Peter De Costa, Luke Plonsky and Sue Starfield, he is a co-editor of *The Palgrave Handbook of Applied Linguistics Research Methodology* (Palgrave, 2017). He is Associate Editor of *Language Assessment Quarterly* and *University of Sydney Papers in TESOL*.

In 2010, he was a recipient of the TOEFL Outstanding Young Scholar Award, and the University of Sydney Faculty of Education and Social Work Teaching Excellence Award. He is Vice President of ALTAANZ (Association for Language Testing and Assessment of Australia and New Zealand).

IELTS Research Program

The IELTS partners – British Council, Cambridge English Language Assessment and IDP: IELTS Australia – have a longstanding commitment to remain at the forefront of developments in English language testing. The steady evolution of IELTS is in parallel with advances in applied linguistics, language pedagogy, language assessment and technology. This ensures the ongoing validity, reliability, positive impact and practicality of the test. Adherence to these four qualities is supported by two streams of research: internal and external.

Internal research activities are managed by Cambridge English Language Assessment's Research and Validation unit. The Research and Validation unit brings together specialists in testing and assessment, statistical analysis and item-banking, applied linguistics, corpus linguistics, and language learning/pedagogy, and provides rigorous quality assurance for the IELTS test at every stage of development. External research is conducted by independent researchers via the joint research program, funded by IDP: IELTS Australia and British Council, and supported by Cambridge English Language Assessment.

Call for research proposals:

The annual call for research proposals is widely publicised in March, with applications due by 30 June each year. A Joint Research Committee, comprising representatives of the IELTS partners, agrees on research priorities and oversees the allocations of research grants for external research.

Reports are peer reviewed:

IELTS Research Reports submitted by external researchers are peer reviewed prior to publication.

All IELTS Research Reports available online:

This extensive body of research is available for download from www.ielts.org/researchers

INTRODUCTION FROM IELTS

This study by Aek Phakiti of the University of Sydney was conducted with support from the IELTS partners (British Council, IDP: IELTS Australia, and Cambridge English Language Assessment) as part of the IELTS joint-funded research program. Research funded by the British Council and IDP: IELTS Australia under this program complement those conducted or commissioned by Cambridge English Language Assessment, and together inform the ongoing validation and improvement of IELTS.

A significant body of research has been produced since the research program started in 1995, with over 110 empirical studies receiving grant funding. After a process of peer review and revision, many studies have been published in academic journals, IELTS-focused volumes in the *Studies in Language Testing* series (www.cambridgeenglish.org/silt), and in the *IELTS Research Reports*. Since 2012, in order to facilitate timely access, individual reports have been published on the IELTS website after completing the peer review and revision process.

In this study, Phakiti investigated the relationship between candidates' perceptions of their performance on the IELTS Listening test and their actual performance on the test. The study found that this group of candidates was overconfident about their abilities, more so on harder test questions, and males more so than females. While high-ability candidates were under-represented in the study sample, there was some evidence that these candidates may exhibit the opposite tendency of underestimating their ability.

This tendency of less skilled individuals overestimating themselves is known more popularly as the Dunning-Kruger effect. It has been observed across a number of areas from skill in driving to chess-playing ability to financial knowledge. Kruger and Dunning's (1999) original study also showed it to be true with regard to knowledge of English grammar, and now we know it is also true with regard to listening comprehension.

Kruger and Dunning argue that it is lack of skill itself that leaves people unable to recognise their poor performance. The current study adds to that explanation, indicating that it is also potentially moderated by other factors. It was shown, for example, that estimates based on a single test item

were less accurate compared to estimates based on a block of items. Another is the difference in estimates between men and women, indicating that gender, or some other factor on which the genders differ, affects such estimates.

The more important question is whether anything could be done about it. A number of the areas studied by Kruger, Dunning and their colleagues are ones where people are presumed to have received substantial feedback on, which would indicate that ability to estimate one's abilities is potentially not susceptible to feedback or training. More formal studies to show whether this is indeed the case would be quite useful.

In any event, we know from the studies that there is at least one solution to the problem of inaccurate self-evaluations, which is: to become better at the thing itself. The better one's language abilities, the less one overestimates one's abilities, and indeed potentially underestimates them. Thus, instead of trying to improve people's self-evaluations, which may well be impossible, we can work instead on improving people's language ability, which we know to be possible.

How will we know when we have solved the problem? Many years ago I was told: when you think you know everything, they give you a Bachelor's degree. When you know there are things you don't know, then they give you a Master's degree. And when you know that you don't know anything, that's when they give you a Ph.D. With this in mind, may all language learners get their Ph.Ds!

**Dr Gad S Lim, Principal Research Manager
Cambridge English Language Assessment**

References

Kruger, J & Dunning, D, 1999, 'Unskilled and unaware of it: How difficulties in recognising one's own incompetence lead to inflated self-assessments', *Journal of Personality and Social Psychology*, vol. 77, no. 6, 1121–1134.

CONTENTS

1	INTRODUCTION	8
1.1	Operationalised definitions of the key constructs.....	9
2	REVIEW OF THE LITERATURE	10
2.1	L2 listening processes	10
2.2	General research on test-taking strategies	12
2.3	Research on test-taking strategies in IELTS Listening tests.....	13
2.4	Research on individuals' appraisal calibration	15
2.4.1	Defining appraisal calibration	15
2.4.2	Metacognition and appraisal calibration	16
2.4.2.1	Metacognition	16
2.4.2.2	Appraisal calibration	20
2.4.2.3	The local mental model (LMM).....	21
2.4.2.4	The probabilistic mental model (PMM).....	23
2.4.2.5	Internal and external feedback	23
2.4.2.6	Two types of appraisal confidence	24
2.4.3	Empirical findings about individuals' appraisal calibration.....	24
2.4.4	Research on test-takers' appraisal calibration in language testing and assessment.....	25
2.4.5	Implications for the present study.....	26
2.4.5.1	Research problems	26
2.4.5.2	Research questions.....	27
3	RESEARCH QUESTIONS	27
3.1	Research context	27
3.2	Research design	27
3.3	Ethical considerations	28
3.4	Research settings	28
3.5	Participants	29
3.6	Research instruments.....	29
3.6.1	Trait and state cognitive and metacognitive strategy use and IELTS listening test difficulty questionnaires	29
3.6.2	The simulated IELTS Listening test.....	30
3.6.3	Single-case appraisal confidence and relative-frequency appraisal confidence scales.....	31
3.7	Data collection	31
3.7.1	Appraisal confidence rating practice.....	32
3.8	Data analysis	33
3.8.1	Item-level analysis	33
3.8.1.1	Analysis of the trait and state questionnaires.....	33
3.8.1.2	Analysis of the IELTS Listening test.....	36
3.8.1.3	Analysis of the single-case and relative-frequency questionnaire.....	39
3.8.2	Data analysis to address the research questions.....	41
3.8.2.1	Analysis of appraisal calibration.....	41
3.8.2.2	Appraisal calibration score	41
3.8.2.3	T-tests	42
3.8.2.4	Analysis of variance (ANOVA).....	42
3.8.2.5	Structural equation modeling (SEM).....	43
4	FINDINGS	46
4.1	What is the nature of test-takers' appraisal confidence and appraisal calibration in an IELTS Listening test?.....	46
4.1.1	The nature of test-takers' appraisal confidence and IELTS Listening test performance	46
4.1.2	Test-takers' appraisal calibration scores	47
4.1.3	Correlations between appraisal confidence and performance	50
4.1.4	Model of IELTS Listening test performance	51
4.1.5	Correlations between single-case appraisal confidence and relative-frequency appraisal confidence.....	52
4.1.6	Models of single-case and relative-frequency appraisal confidence	53
4.1.7	SEM correlations between appraisal confidence and IELTS Listening test performance	55
4.1.8	CFA of appraisal calibration	57
4.2	What is the nature of test-takers' appraisal calibration in easy, moderately difficult, difficult and very difficult IELTS Listening questions?	59

4.2.1	Appraisal confidence and performance based on test difficulty levels	59
4.2.2	Paired-samples <i>t</i> -tests between appraisal confidence and performance based on question difficulty levels	60
4.2.3	Correlations between appraisal confidence and performance based on IRT test difficulty levels	61
4.3	Do male and female test-takers differ in their appraisal confidence and calibration scores in an IELTS Listening test?	65
4.4	Do test-takers with different ability levels differ in their appraisal calibration scores?	71
4.4.1	ANOVA results on appraisal calibration scores among the six ability groups	72
4.5	What are the structural relationships among test-takers' appraisal confidence, calibration, trait and state cognitive and metacognitive strategy use, IELTS Listening test difficulty, and IELTS Listening performance?	79
4.5.1	Trait cognitive and metacognitive strategy use	79
4.5.2	State cognitive and metacognitive strategy use	80
4.5.3	The relationships between trait and state MSU and CSU	81
4.5.4	The relationships among trait and state MSU and CSU and appraisal confidence	83
4.5.5	Trait and state cognitive strategy use, appraisal confidence, and IELTS Listening test performance	86
4.5.6	Trait and state MSU and CSU and appraisal calibration	88
4.5.7	Trait and state cognitive strategy use, appraisal confidence, trait and state IELTS Listening test difficulty, and IELTS test performance	90
5	DISCUSSION	92
5.1	Discussion of the findings	93
5.1.1	Research question 1: The nature of test-takers' appraisal confidence and calibration in IELTS Listening test tasks	93
5.1.2	Research question 2: The nature of confidence and calibration in easy, moderately difficult, very difficult and extremely difficult questions	95
5.1.3	Research question 3: Gender differences in appraisal confidence and calibration scores	96
5.1.4	Research question 4: Test-takers with different success levels and their appraisal calibration scores	97
5.1.5	Research question 5: The structural relationships among test-takers' confidence, calibration, trait and state cognitive and metacognitive strategy use, IELTS listening test difficulty, and IELTS Listening performance	98
5.2	Limitations of the present study	99
6	CONCLUSIONS AND IMPLICATIONS	100
6.1	Implications for the IELTS Listening test	101
6.2	Implications for language teaching and IELTS test preparation	101
6.3	Recommendations for future research	102
6.4	Concluding remarks	104
	ACKNOWLEDGMENTS	104
	REFERENCES	105
	APPENDIX 1: RESEARCH INSTRUMENTS	112
A1.1	General instructions	112
A1.2	Background questionnaire	112
A1.3	Trait strategy use and IELTS listening difficulty questionnaire	113
A1.4	Practice IELTS Listening test questions with appraisal confidence rating	114
A1.5	The IELTS Listening test	115
A1.6	State strategy use and IELTS listening difficulty questionnaire	124
A1.7	Answer keys	125
A1.8	IELTS Listening tapescripts	127
A1.9	Example of feedback to students	133
	APPENDIX 2: IRT ANALYSIS	134
A2.1	Calculating fit statistics	134
A2.2	Item fit graph: Misfit order	134
A2.3	Item statistics: Measure order	135
A2.4	Person statistics: Measure order	136

List of tables

Table 1: Taxonomy of the trait and state cognitive and metacognitive strategy use and IELTS Listening test difficulty questionnaires.....	30
Table 2: Summary of the four sections of the IELTS Listening test.....	31
Table 3: Single-case appraisal confidence explanations.....	32
Table 4: Distributions for trait cognitive and metacognitive strategies and trait IELTS Listening difficulties.....	33
Table 5: Distributions for state cognitive and metacognitive strategies and state IELTS Listening difficulties.....	34
Table 6: Taxonomy of the trait and state cognitive and metacognitive strategy use and state and trait IELTS Listening test difficulty questionnaires.....	34
Table 7: Descriptive statistics for the trait and state cognitive and metacognitive strategies and state and trait IELTS Listening difficulties (N = 376).....	35
Table 8: Internal consistency estimates (Cronbach's alpha) (N = 376).....	35
Table 9: Summary of case estimates (N = 388).....	36
Table 10: Descriptive statistics of the IELTS test performance variables (N = 376).....	37
Table 11: Internal consistency estimates (Cronbach's alpha) for the IELTS Listening test (N = 376).....	37
Table 12: IELTS Listening question difficulties with Cronbach's alpha coefficients.....	39
Table 13: Distributions for single-case appraisal confidence of the 40 questions (N = 376).....	39
Table 14: Distributions of single-case appraisal confidence and relative-frequency appraisal confidence across the four IELTS sections (N = 376).....	40
Table 15: Internal consistency estimates (Cronbach's alpha) for the single-case appraisal confidence.....	40
Table 16: Common symbols used in SEM.....	43
Table 17: Summary of the key GOF criteria and acceptable fit levels and interpretations.....	44
Table 18: Descriptive statistics of the single-case and relative-frequency appraisal confidence and IELTS Listening test performance variables (N = 376).....	46
Table 19: The paired-sample t-test results between single-case and relative-frequency appraisal confidence.....	46
Table 20: The paired-sample t-test results between single-case and relative-frequency appraisal confidence.....	47
Table 21: Test-takers' calibration scores in the IELTS Listening test (N = 376).....	47
Table 22: The paired-sample t-test results (N= 376).....	49
Table 23: Pearson-Product-Moment correlations between appraisal confidence and IELTS Listening performance (N = 376).....	50
Table 24: Pearson-Product-Moment correlations between single-case and relative-frequency confidence.....	52
Table 25: Comparisons between SEM and Pearson-Product-Moment correlations (N = 376).....	57
Table 26: Descriptive statistics of test-takers' IELTS Listening scores and single-case appraisal confidence according to IRT test difficulty levels (N = 376).....	60
Table 27: The paired-sample t-test results between appraisal confidence and performance based on IRT test difficulty levels (N = 376).....	60
Table 28: Comparisons between SEM and Pearson-Product-Moment correlations based on test difficulty levels.....	64
Table 29: Descriptive statistics of appraisal confidence and IELTS Listening performance between male and female test-takers (N = 376).....	66
Table 30: Descriptive statistics of male and female test-takers' appraisal calibration scores.....	67
Table 31: Test of homogeneity of variances.....	67
Table 32: Result of the one-way ANOVA for IELTS Listening scores and single-case appraisal confidence.....	68
Table 33: Result of the one-way ANOVA for appraisal calibration scores.....	68
Table 34: Test of homogeneity of variances.....	72
Table 35: Descriptive statistics of test-takers' appraisal calibration scores.....	72
Table 36: The Scheffe post hoc test in Sections 1 and 3, moderately difficult questions and difficult questions among the six ability groups (N = 376).....	75
Table 37: Summary of two of the highest IRT ability test-takers' performance and appraisal confidence.....	76
Table 38: Summary of two of the lowest IRT ability test-takers' performance and confidence.....	77
Table 39: Pearson-Product-Moment correlations between appraisal calibration and IELTS Listening accuracy and appraisal confidence (N = 376).....	93
Table 40: Pearson-Product-Moment correlations between appraisal calibration and IELTS Listening accuracy and appraisal confidence based on difficulty levels (N = 376).....	96

List of figures

Figure 1: A multidimensional model of strategic competence (Phakiti 2007b, p. 152) 18

Figure 2: Human information processing (Phakiti 2007b, p. 157)..... 19

Figure 3: Cognitive processing and confidence level generation in solving a multiple-choice test task
(adapted from Gigerenzer et al. 1991 by Phakiti 2005, p. 30)..... 22

Figure 4: Flow chart of the data collection procedures 32

Figure 5: IRT item difficulty and person ability map (N = 388)..... 38

Figure 6: Calibration of performance appraisal diagram..... 41

Figure 7: A hypothesised one factor model of trait planning strategy use Time 1 (Phakiti, 2007b, N = 651) 44

Figure 8: A flow chart of SEM used in the present study 45

Figure 9: Test-takers' appraisal calibration diagram (single-case appraisal confidence) of the overall test 48

Figure 10: Test-takers' appraisal calibration diagram (single-case appraisal confidence) 48

Figure 11: Test-takers' appraisal calibration diagram (relative-frequency appraisal confidence)..... 49

Figure 12: Test-takers' appraisal calibration diagram (single-case appraisal confidence) of Section 4 51

Figure 13: The CFA model of IELTS Listening test performance 52

Figure 14: CFA of single-case appraisal confidence 53

Figure 15: CFAs of relative-frequency appraisal confidence 53

Figure 16: The SEM model of the relationship between single-case appraisal confidence and
relative-frequency appraisal confidence 54

Figure 17: The SEM model of the relationship between the latent single-case appraisal confidence
and the latent IELTS Listening test performance..... 55

Figure 18: The SEM model of the relationship between the latent relative-frequency appraisal confidence
and the latent IELTS Listening test performance..... 56

Figure 19: The CFAs of single-case appraisal calibration and relative-frequency appraisal calibration..... 57

Figure 20: SEM model of the relationship between latent single-case and relative-frequency appraisal
calibration (N = 376) 58

Figure 21: The second-order CFA of a latent calibration factor (N = 376)..... 59

Figure 22: Test-takers' appraisal calibration diagram based on easy questions (k = 7, N = 376)..... 61

Figure 23: Test-takers' appraisal calibration diagram based on moderately difficult questions (k = 11, N = 376). 61

Figure 24: Test-takers' appraisal calibration diagram based on difficult questions (k = 12, N = 376) 62

Figure 25: Test-takers' calibration diagram based on very difficult questions (k = 9, N = 376) 62

Figure 26: Test-takers' appraisal calibration diagram based on the four difficulty levels (N = 376) 63

Figure 27: The SEM model of the relationship between the latent single-case appraisal confidence
and the latent IELTS Listening test performance based on test difficulty levels..... 64

Figure 28: Male and female test-takers' appraisal calibration diagram in Section 1..... 69

Figure 29: Male and female test-takers' appraisal calibration diagram in Section 3..... 69

Figure 30: Male and female test-takers' appraisal calibration diagram in easy questions..... 70

Figure 31: Male and female test-takers' appraisal calibration diagram in moderately difficult questions 70

Figure 32: Distribution of test-takers based on IRT ability 71

Figure 33: Distribution of the six test-taker groups based on the IRT ability 71

Figure 34: A calibration diagram of Groups 1 and 6 on Section 1 of the IELTS Listening test..... 76

Figure 35: Appraisal calibration diagram of test-taker IRT logit 3.76 (Group 1) 77

Figure 36: Appraisal calibration diagram of test-taker IRT logit 3.24 (Group 1) 78

Figure 37: Appraisal calibration diagram of test-taker IRT logit -2.78 (Group 6) 78

Figure 38: Appraisal calibration diagram of test-taker IRT logit -2.49 (Group 6) 79

Figure 39: The SEM model of the relationship between trait MSU and trait CSU 80

Figure 40: The SEM model of the relationship between state MSU and state CSU..... 81

Figure 41: The SEM model of the relationship between trait and state MSU and CSU 82

Figure 42: The SEM model of the relationship of single-case appraisal confidence to trait and
state MSU and CSU (N =376) 84

Figure 43: The SEM model of the relationship of single-case and relative-frequency appraisal confidence
to trait and state MSU and CSU (N =376) 85

Figure 44: SEM model of trait and state cognitive strategy use, appraisal confidence, and
IELTS test performance 86

Figure 45: SEM model of trait and state cognitive strategy use and IELTS test performance 87

Figure 46: SEM model of trait and state cognitive strategy use and appraisal calibration 89

Figure 47: The SEM model of trait and state cognitive strategy use, appraisal confidence, trait and state
IELTS Listening test difficulty, and IELTS test performance..... 90

1 INTRODUCTION

It is a well-established practice for English-medium universities to consider non-English speaking background (NESB) international applicants' English language proficiency level as one of the most important admission criteria (second only to academic performance). The International English Language Testing System Academic (IELTS) is one of the most widely used academic language tests by receiving academic institutions in Australia. It is considered to provide trustworthy evidence of international applicants' English language proficiency, which is then used in the admissions decision-making process.

Given the high-stakes nature of the use of IELTS (e.g., academic admission, immigration purposes), IELTS validation research is essential not only to provide a good understanding of the nature of language test performance through various test tasks, but also to improve the quality of the test and the interpretation of test-takers' scores. Test validation can also help ascertain whether decisions made on the basis of the test score (e.g., for admissions purposes) are theoretically and empirically sound or not.

While several researchers propose various intertwined criteria for evaluating test validity evidence, Chapelle, Enright and Jamieson's (2008, 2010) criteria are among the most comprehensive: (1) evaluation (e.g., evidence of targeted listening abilities); (2) generalisation (e.g., evidence of score consistency across different test tasks or questions); (3) explanation (e.g., listening scores reflect target language proficiency; usefulness of test scores, performance feedback); (4) extrapolation (e.g., evidence of the test's relations to other relevant, real-life conditions in both test and non-test contexts); and (5) utilisation (e.g., evidence of appropriate educational decision-making practices, fairness and consequences of test use). This study can provide the validity evidence related to evaluation, generalisation and explanation.

Although the major factor that explains a test score should be ability in the target language (the construct of interest), it has been well understood that there are factors other than the target language constructs that also contribute to a test score (Bachman 2000). For example, test-takers may perform differently when they take a multiple-choice test as compared to when they take a construct-response test (i.e., test-methods facets).

People who are motivated to do well in a test are likely to invest more effort and to self-regulate to complete a test than those who are not (i.e., individual characteristics). Bachman (2000) further suggests that understanding the effects of test tasks on test performance and how test-takers cognitively interact with given test tasks is the most pressing issue facing language testing. In particular, the conceptualisation of test difficulty should not be understood and interpreted merely from an analysis of test task characteristics and pre-determined difficulty levels set by the test developers, but rather test difficulty should be viewed as a function of complex interactions between a given test-taker and a given test task (Bachman 2000).

Examining the interaction between test-task characteristics and test-takers' characteristics is also relevant to Weir's (2005) socio-cognitive validity framework, which highlights the equal importance of both test-takers' mental processing and their use of language to perform a test task. Weir's validity framework considers various local types of validity before, during (i.e., cognitive and contextual validity) and after the test event (i.e., scoring, consequential and criterion-related validity). The present study provides validity evidence associated with the cognitive validity (i.e., how a test task represents or activates the cognitive processes involved in the listening); and the context validity (i.e., the extent to which a test task is associated with the target linguistic demands and settings; see also see Field 2009a; Shaw & Weir 2007) of a test task.

Second language (L2) ability is known to be highly complex and multidimensional (McNamara 1996) because it involves both internal factors (e.g., individual characteristics and language ability) and external factors (e.g., social contexts, test tasks, and setting). Such complexity and the multidimensionality of L2 ability make it difficult to validly assess it (e.g., Bachman & Palmer 1996, 2010; McNamara 1996). In the past three decades, we have seen numerous evolving theoretical models proposing the components of L2 ability (e.g., Bachman 1990; Bachman & Palmer 1996, 2010; Canale & Swain 1980; Hymes 1972). Of interest in the current study is the notion of 'the ability for use' (Hymes 1972), which has been conceptualised as 'strategic competence' in the communicative language ability (CLA) model in Bachman (1990) and Bachman and Palmer (1996, 2010).

According to Bachman and Palmer (2010), strategic competence is a cognitive mechanism that mediates the internal processes with the test task and setting.

In their revised models, Bachman and Palmer (2010) describe strategic competence as being composed of (1) goal setting, (2) appraisal (monitoring and evaluating), and (3) planning. According to Bachman and Palmer (2010), strategic competence manifests itself as a set of meta-cognitive strategies, which regulate cognitive strategies, linguistic processes and other psychological processes, such as world knowledge and affect (e.g., motivation and anxiety). Of particular interest to the present study is a revised strategic competence facet, namely *performance appraisals* (formerly related to assessing such as in 'assessing the situation'). Bachman and Palmer (2010) point out that "appraising the correctness or appropriateness of the response to an assessment task involves appraising the individual's response to the task with respect to the [individual's] perceived criteria for correctness or appropriateness" (p. 51).

The present study aims to examine four aspects that are theoretically related to test scores:

1. test-takers' trait (i.e., generally perceived) and state (i.e., context-specific) cognitive and metacognitive strategy use in IELTS Listening tests
2. test-takers' appraisal confidence and calibration for each test question (i.e., single-case confidence) and for the entire test section (i.e., relative-frequency confidence)
3. trait and state test difficulty perception in IELTS Listening tests
4. test difficulty and test-takers' ability as key factors affecting the above variables.

Inferential statistics such as Pearson-Product-Moment correlations, *t*-tests, analysis of variance (ANOVA), and structural equation modeling (SEM) are used to address the research aims.

1.1 Operationalised definitions of the key constructs

There are relevant constructs in the research literature and some researchers prefer to use different terms to describe similar constructs. To be consistent in the use of terms, this section introduces working definitions of the common key constructs mentioned in this study.

Appraisal calibration: A psychological construct of test-takers' ability to accurately determine the extent to which they are successful in answering a test question or completing a task

Appraisal confidence: A level of test-takers' confidence in the correctness of their answer to a test question or task. Appraisal confidence can be measured using a percentage scale.

Cognitive strategy use: The conscious and intentional processes of employing language knowledge, domain-general knowledge (e.g., world knowledge), domain-specific knowledge, and/or prior experiences related to listening comprehension that help listeners comprehend audio text and answer test questions or complete tasks. Cognitive strategies include memorising, comprehending, and retrieving information simultaneously from the working and long-term memories.

Listening difficulty: Test-takers' perceived feelings about cognitive difficulties arising from participating in a listening task and their judgments on the extent to which they perceive a level of difficulty being experienced.

Metacognitive strategy use: The conscious and intentional processes of controlling how cognitive strategies are used to address a listening test task. Metacognitive strategies include goal setting, planning, monitoring, and evaluating or appraising.

Performance appraisal: The monitoring function of control processing during language processing that identify whether test-takers perceive they have completed a test task successfully and to what extent they perceive they have been successful

State: A specific instance of performance, thoughts or feelings that occur currently or within a specific context or time. State can be observed during an event (e.g., via introspection) or after an event has been completed (e.g., retrospection). A state performance is a result of an interaction between an individual's information processing and the characteristics of a given task and context.

Strategic competence: The higher-order cognitive mechanism that takes control of thoughts or behaviours during test task completion. Strategic competence is made up of strategic knowledge and strategic regulation (see further below). Strategic competence underlies the effective use of metacognitive processes that regulate thoughts or cognitive processes.

Strategic competence is made up of both automatic metacognitive processing as well as conscious metacognitive processing. That is, if test-takers can monitor their performance unconsciously or effortlessly and their performance is also successful, they possess strategic competence. However, when they experience difficulties, they realise the need to be able to explicitly take control of their thoughts so as to help them complete a given task successfully.

Strategic knowledge: What learners know about their accumulated metacognitive strategy use, such as goal setting, planning, and appraising. Strategic knowledge, which tends to reside within the long-term memory, includes declarative knowledge (knowing what metacognitive strategies they possess), procedural knowledge (knowing how to use the metacognitive strategies they possess), and conditional knowledge (knowing when to use the metacognitive strategies they possess).

Strategic regulation: The metacognitive processes learners use to regulate their thoughts while addressing a given test task. Strategic regulation tends to take place within the working memory and may involve interaction among declarative, procedural and conditional knowledge.

Trait: A context-free pre-disposition of an individual regarding ability, knowledge, thoughts, or feelings that is enduring over time. A trait is more stable than a state (see above). For example, a person may be perceived by others as anxious. The degree to which that person is anxious in a specific context (state anxiety) may not be the same as the degree to which he/she is generally anxious (trait anxiety). During the course of a cognitive development or language acquisition, a trait is not necessarily a permanent state.

2 REVIEW OF THE LITERATURE

This section presents the theoretical frameworks underpinning the current study. It presents the relevant research literature in L2 listening, test-taking strategies, and appraisal calibration.

2.1 L2 listening processes

The construct of L2 ability is undeniably complex, as there are various modes of language use, such as reading, listening, speaking, writing, vocabulary and grammar. This study focuses on assessing listening and, in particular, the IELTS Listening section. This study focuses on just one skill because

each language skill is unique and complex (vanPatten 1994) and should be specifically and comprehensively researched (Schmidt 1995).

L2 listening is a multidimensional socio-cognitive process, which requires consideration not only from the neurological, linguistic, and psycholinguistic perspectives but also from the social-contextual perspectives in language use (see e.g., Buck 2001; Field 2008, 2013; Goh 2008; Vandergrift 2015; Vandergrift & Goh 2012). Assessing L2 listening is complex because of the need to not only consider models and theories of L2 listening, but also because of the required components of psychometric properties in the measurement of listening ability or assessment task performance. Additionally, the issues of ethics, fairness and the consequences of the use of test results need to be considered. The IELTS Listening test is one of the four modules used to assess academic English.

It has been well documented that listening comprehension is affected by several factors, which interact with one another (see Buck 2001; Field 2008, 2013; Vandergrift 2015; Vandergrift & Baker 2015). Two such factors are the listener and the context in which the test is taken. Listener factors include linguistic knowledge, topic knowledge, strategic competence or metacognition, working memory, motivation and anxiety. Contextual factors include speaker factors (e.g. accents), text characteristics (e.g., speech rate and density and modification of information), organisation of texts (e.g., step-by-step text or text with cross references), text types (e.g., transactional/non-reciprocal versus interactional/reciprocal), and task characteristics (e.g., true/false, multiple-choice, constructed-response questions).

According to Vandergrift and Goh (2012), L2 listening is not only an area of great weakness for many students, but also the area which receives the least structured support and systematic attention from teachers in the L2 classroom. There are several models of L2 listening (e.g. Field 2008, 2013; Goh 2008; Rost 2011; Vandergrift & Goh 2012) that are useful to help us understand the processes and factors influencing L2 listening comprehension and test performance.

According to Vandergrift and Goh (2012), in the perception phase, the listener needs to decode incoming speech phonetically. During the parsing phase, the listener parses the phonetics from memory and begins to activate potential words,

which depends on his/her level of language proficiency. The bottom-up processing takes place during the first two phases. It is a decoding process that segments the sound in the text into meaningful units. In the utilisation phase, the listener generates a conceptual framework that matches the sound stream by referring to the context and their prior knowledge. This phase is related to the allocation of meaning to the input being heard. During the utilisation phase, top-down processing (e.g., the application of context and prior knowledge to interpret the message) is required as prior knowledge is stored and retrieved from the long-term memory to comprehend the sound stream.

It should be noted that neither bottom-up nor top-down processing is adequate for successful listening comprehension. In the case of bottom-up processing, the listener cannot cope with ongoing audio text, which often results in a loss of comprehension, while in the case of the top-down processing, the listener does not necessarily have all the prior knowledge essential to make sense of the audio text. Hence, successful listening requires interaction between the two types of processing.

It is also important to examine the important roles of working and long-term memories during listening. The working memory is the platform where the information is processed in the parsing phase through a phonological loop. This memory has a limited capacity to keep information for a long time and is, therefore, the place where the listener needs to segment text meaning in association with the long-term memory. The long-term memory is the platform where the listener stores and retains various types of knowledge (e.g., declarative, procedural and conditional knowledge, world knowledge, and in particular linguistic knowledge).

Field (2013) also provides a cognitive processing model of listening that is somewhat similar to that of Vandergrift and Goh (2012). However, Field (2013), proposes five levels of processing, which include: (1) input decoding (e.g., transforming acoustic information into groups of syllables); (2) lexical search (e.g., word-level matches to what is heard); (3) parsing (e.g., relating lexical material to the co-text to identify or clarify lexical meaning and construct a syntactic pattern with reference to pragmatic, background and socio-linguistic knowledge); (4) meaning construction (e.g., employing world knowledge or making inferences); and (5) discourse construction (e.g., making an important decision or judgment about the

new information gathered in relation to what has already been collected).

Field (2013) describes the process by which the listener may form a hypothesis about what is being heard and then revise it on the basis of new evidence. The hypothesis forming process is regarded as a tentative process of listening during the decoding phase. During meaning construction, the listener needs to supply his/her own information including pragmatic, contextual, semantic, and inferential information. During the discourse construction phase, the listener needs to decide what is relevant, what to store for later use (i.e., selection), and what new information to add to the developing meaning representation (i.e., integration). The listener also needs to compare new information with that already collected to check for consistency or congruence (i.e., self-monitoring) and to consider the relative, hierarchical importance of new and old information in order to construct key points with supporting points (i.e., structure building). The part of monitoring for consistency processing is relevant to the investigation of calibration in the present study.

According to Field, lower-proficiency listeners are likely to spend their time dealing with the first three levels in Field's model (2013), whereas higher-proficiency listeners are able to handle more in the last two levels as they are able to deal with more complex linguistic features and cognitive load in their working memory. Field also notes the important role of strategic competence in L2 listening proficiency because it helps L2 listeners make sense of listening in a real world setting, allowing them to extend their "comprehension beyond what their knowledge and expertise might otherwise permit" (p. 108).

A challenging task for L2 listening researchers is to identify listening strategies that appear to constitute the characteristics of a successful L2 listener. Field points out that listening strategy use takes place not only in regard to "the use of contextual and co-textual 'top-down' information in order to solve local difficulties of comprehension" (p. 108), but also at various word levels, particularly when listeners are uncertain about the reliability of what has been understood, leading them to use the most likely word matches in spite of the context and co-text.

2.2 General research on test-taking strategies

In the past few decades, test-taking strategy research has benefited greatly from language learning strategy research which focuses on the importance of metacognition (i.e., knowledge about and regulation of one's thinking), which underpins strategy use in terms of conceptualisation, operationalisation and utilisation of strategy taxonomies (e.g., cognitive, metacognitive, affective, and social strategies). In language testing research, the ability to use effective and suitable strategies during the completion of test tasks is conceptualised to be related to strategic competence (see Phakiti 2007b; Purpura 1999). When students take a language test, they encounter test questions and tasks and are expected to produce language in response to the given test questions or tasks. Their test scores are used to determine not only how well they have done in the test, but also the level of their language ability or proficiency relative to some standard. Test-takers need to be concerned with how well they are doing in the test and hence to check their ongoing test performance.

Language testing researchers generally aim to examine the nature of the strategy types used to respond to test tasks (e.g., cognitive or metacognitive strategies), how they are related to one another and to language test performance. There is consensus that strategy use or strategic processing has a component of awareness or consciousness and takes place within the working memory realm (Cohen 2011; Phakiti 2008a).

According to Alexander, Graham and Harris (1998), strategies differ from skills and other common processes in the test-takers' levels of awareness and deliberation, rather than the nature of the processes per se. For example, when test-takers automatically check their test performance without being aware of such evaluative processing, it can be said that this processing is a common, unreflective process, rather than a monitoring strategy. However, when they tell themselves to check their test performance before submitting the test, it can be said that this type of monitoring is a strategy. In the latter case, test-takers can report the conscious level of their processing, whereas in the former, they might not realise they have engaged in such a process.

Much test-taking strategy research has focused on defining and measuring strategies via the use of both quantitative (e.g., Likert-type scale questionnaires; e.g., Bi 2014; Phakiti 2003b, 2008a; Purpura 1999; Song 2004, Zhang & Zhang 2013) and qualitative (e.g., interviews and think-aloud protocols; Cohen & Upton 2007; Phakiti, 2003b) methodologies in various language testing and assessment contexts (see also Cohen 2011, 2014). Furthermore, test-taking strategy research has benefited from several advancements in research methodology, including applications of sophisticated statistical analysis (e.g., structural equation modeling).

Purpura (1999) was the first to examine the relationship between generally perceived cognitive and metacognitive strategies and language test performance as assessed by UCLES's First Certificate in English Anchor Test. Purpura employed a structural equation modeling (SEM) approach with 1,382 test-takers. The study found that cognitive processing was a multi-dimensional construct including a set of comprehending, memory and retrieval strategies that operated to influence language performance. Metacognitive strategies were found to be unidimensional, consisting of a single set of assessment processes. Purpura tested for a hierarchical relationship among metacognitive processing, cognitive processing and language test performance. Purpura also found that high-ability test-takers employed some metacognitive processing more automatically than low-ability ones. These different patterns in turn had a significant impact on test-takers' language performance. It should be noted that Purpura defines strategies to be both conscious and unconscious processes and deliberately chooses to use *processing* instead of strategies.

Phakiti (2008a) examined the relationships between test-takers' strategic knowledge (i.e., trait strategies) and strategic regulation (i.e., state strategies) and high-stakes, EFL reading test performance on two occasions using a SEM approach. The terms *trait* and *state* are borrowed from anxiety research (Spielberger 1972), which highlights the importance of the two dual constructs of trait anxiety (a relatively stable attribute of a person to be anxious across settings and situations) and state anxiety (a transitory anxiety state in a specific context and/or time).

Research suggests that trait anxiety is stable over time, whereas state anxiety fluctuates across time and is manifested by trait anxiety (Phakiti 2007b). It should, however, be noted that the term *trait* does not imply an immutable disposition (Hertzog & Nesselroade 1987) because during cognitive development and language learning, or as one matures and learns, the trait can gradually change.

In Phakiti (2008a), 561 Thai university student test-takers were asked to answer a trait strategy use questionnaire prior to their midterm and final reading achievement tests and, immediately after completing each test, they were requested to answer a state strategy use questionnaire. Phakiti found a complex relationship among the variables as follows. First, trait metacognitive strategy use (MSU) directly and strongly affected trait cognitive strategy use (CSU) on both occasions (0.95 and 0.96, respectively). It was found that the relationships between trait MSU and CSU were stable over time. Second, trait CSU did not greatly affect state CSU (0.22 and 0.25, respectively). Third, trait MSU directly affected state MSU in a specific context (0.76 and 0.79, respectively), which in turn directly affected state CSU (0.76 and 0.75, respectively). Finally, state CSU directly affected a specific language test performance. This study provided strong evidence for the theoretical distinction between state and trait strategy use in that trait strategy use is more stable than state strategy use and that their relationship is highly complex when modelled over time.

Since the publication of Phakiti (2008a), new studies have examined the similar dimensions of metacognitive and cognitive strategy use in a variety of test contexts (e.g., Bi 2014; Zhang, Gao & Kunnan 2014; Zhang & Zhang 2013). Recent research has found that test-takers' reported strategy use is significantly related to test score variance (small to medium effect sizes; Bi 2014; Zhang, Gao & Kunnan 2014; Zhang & Zhang 2013).

The majority of strategic processing research in language testing and assessment has largely relied on the use of research instruments, such as Likert-type scale questionnaires, think-aloud or verbal protocol methods and stimulated-recall techniques (see e.g., Cohen 2011; Cohen & Upton 2007).

Although Likert-type scale questionnaires are fruitful to aid our understanding of the nature of

strategic processes and to capture some of test-takers' perceived performance appraisals during test taking, they cannot tell us exactly how test-takers judge the correctness of their test performance during their test taking. This is merely because questionnaires are given either at the beginning of the test (e.g., Purpura 1999; Song 2004) or at the end of the test (e.g., Bi 2014; Phakiti 2003b, 2008a).

One limitation of self-report methods, such as Likert-type scale questionnaires, is that they do not allow researchers to make robust inferences regarding test-takers' monitoring processes and monitoring accuracy due to variations in test tasks and the level of task difficulty across test sections. Think-aloud or verbal protocol techniques, while allowing researchers to explore such processes within an individual, face difficulty in their generalisability as they cannot be easily standardised, often yield a small sample size and are expensive to conduct.

In order to advance our understanding of strategic competence in language testing and assessment further, researchers should not merely rely on Likert-type scale questionnaires but should search for additional forms of quantitative measures of online monitoring processes to triangulate with questionnaires.

2.3 Research on test-taking strategies in IELTS Listening tests

As presented earlier, IELTS is a standardised English test, largely used for assessing international students' English language proficiency, although it is also used in other contexts such as for employment and immigration purposes. It is jointly developed by the British Council, the University of Cambridge Local Examination Syndicate (UCLES) and IDP Education Australia; see Aryadoust 2011, 2013).

There are four parts to the IELTS Listening test, comprising a conversation with transactional purposes, a prompted monologue with transactional purposes, a discussion dialogue in an academic context and a monologue in an academic context. Each part assesses different related skills.

Aryadoust (2013, p. 6) pointed out that the IELTS Listening test is a "while-listening performance test" because test-takers need to read test items before and as they hear audio texts and provide answers to test questions or tasks.

Field (2009) defines it as having a simultaneous listen-read-write format. Several people have critiqued this test type in terms of its potential negative washback effects, the presence of confounding variables (e.g. reading, writing, memory capacity) and difficulties in its validation (Aryadoust 2013).

The IELTS Listening module is the least researched of the IELTS test modules. Several IELTS validation studies have looked at the predictive validity of IELTS Listening results to academic performance, self-assessment or other measures of international students and have frequently found a weak positive or weak negative correlation (see Aryadoust 2011 for a review). Recent validation studies on the IELTS Listening test related to the present study (i.e., those studying cognitive processes) are subsequently discussed. For the purpose of this section, three studies that examined strategy use in IELTS Listening tests have been identified and are discussed as they have implications for the present study.

Field (2009) examined the cognitive validity of Part 4 (an academic lecture) of a retired IELTS Listening test using a stimulated recall method with 29 participants. Field compared two listening conditions: test and non-test conditions. Two audio texts were used (Texts A and B). Under test conditions, the participants listened to the text and answered the test questions. Under non-test conditions, they took notes and wrote a brief summary of the lecture. Fifteen participants heard Text A under test conditions and Text B under non-test conditions and 14 participants heard Text B under test conditions and Text A under non-test conditions. At the end of each test, participants were asked to report on the processes involved in completing the task under test and non-test conditions.

It was found that participants employed a variety of strategies under test conditions (e.g., using collocations to help locate their answers, using the ordering of test items). It was also found that under test conditions, their processing was superficial. Some participants reported that they focused more on lexical matching, rather than on the general meaning of the lecture. Field also found that nearly a third of the participants reported that note-taking under the non-test conditions was less demanding than under the test conditions, suggesting distinctive processes are required under each condition.

Nonetheless, some contradictory evidence about the nature of the cognitive demands of note-taking while performing the lecture-based listening task emerged. Test-takers found note-taking to be more demanding under non-test conditions in terms of constructing meaning representations, dealing with propositional density and topic complexity, and distinguishing important facts from peripheral information. Field identified the potential mismatches between the processes required by the IELTS lecture-based listening tasks under test conditions and those under non-test conditions, which had implications for the cognitive validity of this part of the IELTS Listening test.

Badger and Yan (2009) investigated the differences in the use of tactics and strategies between eight native/expert speakers of English (NESE) and 24 native speakers of Chinese (NC) in IELTS Listening tests. They utilised a think-aloud protocol to identify participants' cognitive and metacognitive strategy use. The researchers distinguished tactics from strategies. Strategies were defined as conscious steps taken by test-takers, whereas tactics were defined as the individualised processes test-takers used to carry out the strategies.

No statistical differences between the two groups in terms of the overall strategy use were found, but out of 13 identified strategies, two statistical differences in metacognitive strategies (i.e., directed attention and comprehension monitoring) were found. The NC group had higher scores on these two strategies than the NESE group. Out of 51 identified tactics, two cognitive tactics (i.e., fixation on spelling, and inferring information using world knowledge) and five metacognitive tactics (i.e., identifying a failure in concentration, identifying a problem with the amount of input, identifying a problem with the process of answering a question, confirming that comprehension has taken place and identifying partial understanding) were significantly different. Of the seven significant tactics, only one tactic (i.e., inferring information using world knowledge) was higher for the NESE group. It is important to note that the parameter estimates might not be stable given the sample sizes ($N = 8$ versus $N = 24$) used for inferential statistical comparisons. The researchers did not mention or provide evidence of whether the statistical assumptions for the independent *t*-tests were met.

Furthermore, the researchers did not articulate or seek to further understand why the high proficiency group reported significantly less use of strategies and tactics than the lower proficiency group. Was it because for high proficiency test-takers, such levels of strategic processing were automatic and that they did not realise that they engaged in it? Was it because lower proficiency test-takers had such low levels of language competence that conscious processing or tactics could not facilitate their test performance? Or was it because the coding of the think-aloud protocols was based more on frequency of use than on the qualities of strategies and tactics under examination?

Winke and Lim (2014) examined the effects of testwiseness and test-taking anxiety on IELTS Listening test performance among English language learners in the US, through an experimental research design that focused on the influence of two types of strategy instruction (total of four hours of instruction which spanned over two weeks). The first group (N = 21) received an explicit instruction on how to use test-taking strategies and skills, whereas the second group (N = 22) received an implicit instruction which focused on vocabulary. The control group (N = 20) did not get a practice test section or any of the strategy instruction, but had two conversational English classes on American culture. The researchers used the IELTS Listening pre-test and post-test (different/parallel tests) on a computer that recorded test-takers' eye movements, responses to a three-part questionnaire on listening strategies, test-taking strategies and test anxiety (for both pre- and post-tests), and stimulated-recall interviews at the end of the data collection.

Winke and Lim did not find a statistically significant difference in the post-test scores among the three groups, indicating that the instructions did not produce an effect on participants' test performance improvement. The researchers also found no differences in reported listening strategies, test-taking strategies and test anxiety among the three groups. The researchers did not find a relationship between testwiseness and test anxiety, but found a negative but weak correlation between test anxiety and the IELTS scores on both pre- and post-test occasions.

By recording eye fixation durations, the researchers compared eye movements of high-anxious test-takers (N = 12) and low-anxious test-takers (N=12) while they read the test instructions.

It was found that the low-anxious test-takers spent far less time reading the instructions. The researchers further found that test-taking anxiety was related to how much time test-takers spent on cloze questions. It is unsurprising that the researchers found no significant differences across the three groups, given the level of treatment conditions (two-hours session per week, for two weeks) and the limited sample sizes. While eye-tracking technology introduced an impressive research technique, it remains questionable to what extent it can be used to infer listening processes and whether it is an authentic way to reflect what test-takers actually do when they take a paper-and-pencil IELTS Listening test.

In summary, these three studies contributed significantly to an understanding of how test-takers employ test-taking strategies to deal with IELTS Listening test tasks. Nonetheless, all these studies had a small sample size, which makes it difficult to generalise their findings regarding test-takers' strategic processes in IELTS Listening tests.

While these studies touched on some aspects of performance appraisals during test taking, little is known about how well calibrated IELTS test-takers are in an IELTS listening test. As further discussed below, when individuals are unrealistic about their performance while taking a test, they are less likely to be engaged in the use of metacognitive strategies or to put effort into the completion of a given task. Hence, appraisal calibration research has a potential to provide valuable insights into the psychology of test-takers.

2.4 Research on individuals' appraisal calibration

2.4.1 Defining appraisal calibration

Calibration research has a long history in psychological research (see Hattie 2013; Stone 2000). Calibration denotes a perfect relationship between confidence in performance success and actual performance outcome. It is related to the accuracy of individuals' judgment of current task success (Alexander 2013; Dinsmore & Parkinson 2013).

For the purpose of the present study, the term appraisal calibration is used and defined as 'a psychological construct of test-takers who can accurately judge or estimate their achievement in test taking' (e.g., Bjorkman 1994; Glenberg, Sanocki, Epstein & Morris 1987; Schraw 2009; Stone 2000).

For example, when test-takers answer a listening test question, they should be able to estimate the likelihood of their answer being correct. Test-takers are calibrated when their confidence in test correctness matches their actual performance perfectly. Calibration can be investigated by asking test-takers to report on their level of appraisal confidence in test performance success immediately after they answer a test question or complete a task. This is done by test-takers using an appraisal confidence rating scale. Their appraisal confidence is then matched against their actual test performance. If there is no difference between test-takers' appraisal confidence and actual test performance, it can be said that their performance appraisal is realistic or calibrated.

For example, if there are 10 questions in a test and Jane answers eight of them correctly, her performance is 80%. If her average appraisal confidence across the 10 questions is also 80%, her appraisal calibration score is 0%. It can then be said that Jane is calibrated in her performance appraisal because there is no gap between her performance appraisal and actual test performance. The deviation from 0% can be seen as error of performance appraisal. The closer the value of the calibration score to 0%, the better calibrated a test-taker is. If her appraisal confidence is 50%, her appraisal calibration score will be -30%. This means that Jane underestimates her test performance. If her appraisal confidence is 95%, her appraisal calibration score will be +15%. This means that Jane overestimates her test performance.

According to Bachman and Palmer (2010, p.49), "appraising the degree to which the language use or assessment task has been successfully completed" is critical to success in test taking. Field (2013) highlights the importance of 'monitoring accuracy' during the discourse construction processes in listening.

It should be noted that first, calibration in the current study is different from the term 'calibration' generally used in the language testing and assessment literature, which deals with the calibration of test items in terms of their difficulty, rating scales and rater training, and which has implications for parallel test forms. Indeed, the idea is similar to calibrating a test or a rater as the aim is to align a test with the target constructs or a rater with the target assessment criteria.

Instead, calibration in this study is about test-takers knowing whether they can or cannot do well in a specific test task (i.e., whether they are realistic or unrealistic in their perceptions of their performance).

Second, the term calibration in the present study is not used as a new label for self-assessment as an alternative assessment in language testing or as learner autonomy in language learning research. Self-assessment research typically asks students to report on the extent to which they believe they know or can do things in the target language using a Likert-type scale. Researchers subsequently examine the correspondence of students' self-assessment to language proficiency test scores, or other relevant measures, thereby evaluating students' self-assessment validity (see Fulcher 2010; Oscarson 1997, 2014; Ross 1998).

Unlike traditional self-assessment research, the study of appraisal calibration resides within human information processing in which the role of metacognition and self-regulation operates in conjunction with working and long-term memories (Dinsmore & Parkinson 2013).

Appraisal calibration is, therefore, indicative of the concurrent validity of test-takers' performance appraisals at the time they are completing a cognitive task as judged by external criteria, such as the correctness of their answers.

2.4.2 Metacognition and appraisal calibration

2.4.2.1 Metacognition

Metacognition is one of the most researched topics in psychological research in educational and developmental psychology (Efklides 2008, 2011; Kliehman & Stankov 2007; Tobias & Everson 2009). Metacognition can be simply described as thinking about thinking, knowing about knowing or cognition of cognition (Flavell 1971). It is a person's ability to self-regulate by planning, monitoring and evaluating his/her learning.

Metacognition leads to metacognitive strategy use that helps manage other cognitive processes to complete tasks. There is consensus that metacognition is composed of two dual components (Alexander et al. 1991): (1) knowledge of cognition (i.e., the accumulated autobiographical information about one's own cognitions); and (2) the regulation of cognition (i.e., the ongoing monitoring and regulation of one's own cognitions; Nelson 1994).

On the one hand, knowledge of cognition (or metacognitive knowledge) is associated with a person's long-term knowledge about his/her own nature as an information processor, about the nature of a task and its demands, about how to achieve task demands under varying conditions, and how to employ a range of strategies to tackle task difficulties (Flavell 1992).

On the other hand, the regulation of cognition is related to a person's concurrent information processing system, which involves ongoing or occasional monitoring, self-assessment and planning which help warrant task completion (Nelson 1994; Nelson & Narens 1994). Metacognitive experiences are defined as the conscious realisation of one's own current, ongoing cognition. Nelson and Narens's (1990) model of metacognition has been influential in several theories and empirical studies in metacognition (e.g., Efklides 2008, 2011). In their model, there are two levels of mental processes: object and metacognitive levels. Two dominant processes, namely control and monitoring, play a critical role in directing information flows between the object and metacognitive levels. Such metacognitive processes result in modifications of processes at the object level. Highly routinized or practised metacognitive strategies, such as planning and reviewing can become automatised at the object level, suggesting that the same metacognitive processes can be explicit at the meta level (i.e., conscious metacognitive strategies), but implicit at the object level (i.e., automatic metacognitive processes).

Phakiti (2007b) proposes a multidimensional model that helps locate strategic competence in human information processing (see Figure 1). This model places strategic competence within the metacognitive/conscious realm of human information processing. Note that Phakiti's current view of strategic competence is that some of its processes may operate at an unconscious level (see Phakiti 2016). Strategic competence is placed along a continuum of strategic knowledge and strategic regulation. Strategic knowledge resides in the long-term memory, which can be *stable* over time, whereas strategic regulation operates within the working memory, which is *malleable*. Strategic regulation constitutes the set of strategies required by a given task. In particular, metacognitive strategies work in concert with cognitive and affective strategies.

Figure 1 suggests that individuals can be aware of their own information processing when the context or situation is unfamiliar, in disequilibrium,

cognitively demanding, high-stakes, risky and/or constrained in some way. This kind of situation requires them to be cautious about their performance and thinking. Contexts are viewed as *ephemeral*. An interaction between a context and individuals results in what is called state performance. It is important to note that the place of strategic and automatic processing is hypothesised in Figure 1. Strategic processing as driven by strategic regulation can be seen at the conscious level, whereas automatic processing can be seen at the unconscious level. At the bottom of this model (the unconscious level), a person may react to a situation with unreflective and habitual thoughts as a situation becomes more familiar, less demanding, less weighty and relatively safer. Strategic competence in this kind of context may be unconscious, subconscious or unreflective (see Phakiti 2016; Purpura 2014). At the top of this model (the conscious level), a higher-level appraisal generates the possibility of a new, perhaps different, interpretation of the situation as well as a broader range of strategic actions to cope with the current situation (Phakiti 2007b, p. 154).

Strategy use occurs at a conscious level and this is where the role of strategic competence is normally studied or inferred. Further applying his model (Figure 1) through the lens of Gagné, Yekovich and Yekovich's (1993) model of human-information processing, Phakiti (2007b) also presented an associated model that postulates what may be going on during information processing in greater detail. Phakiti included affective aspects of human information processing in this complementary model. Figure 2 presents Phakiti's (2007b) model for human information processing during language use or test taking.

According to Figure 2, working memory (WM) interacts with several components, such as long-term memory (LTM), affect, metacognitive monitoring, and metacognitive control. Metacognitive monitoring and control perform an executive function to mediate WM with LTM and affect (Phakiti 2007b). During information processing, metacognitive monitoring and control constantly (but not always as indicated by a dashed arrow) regulate information processing events. These mechanisms result in conscious mental actions, such as goal setting, planning how to achieve goals, assessing the current situation, monitoring goal attainment, and checking and evaluating current performance in WM. These two models have implications for further examination of the nature of individuals' performance appraisals and their calibration.

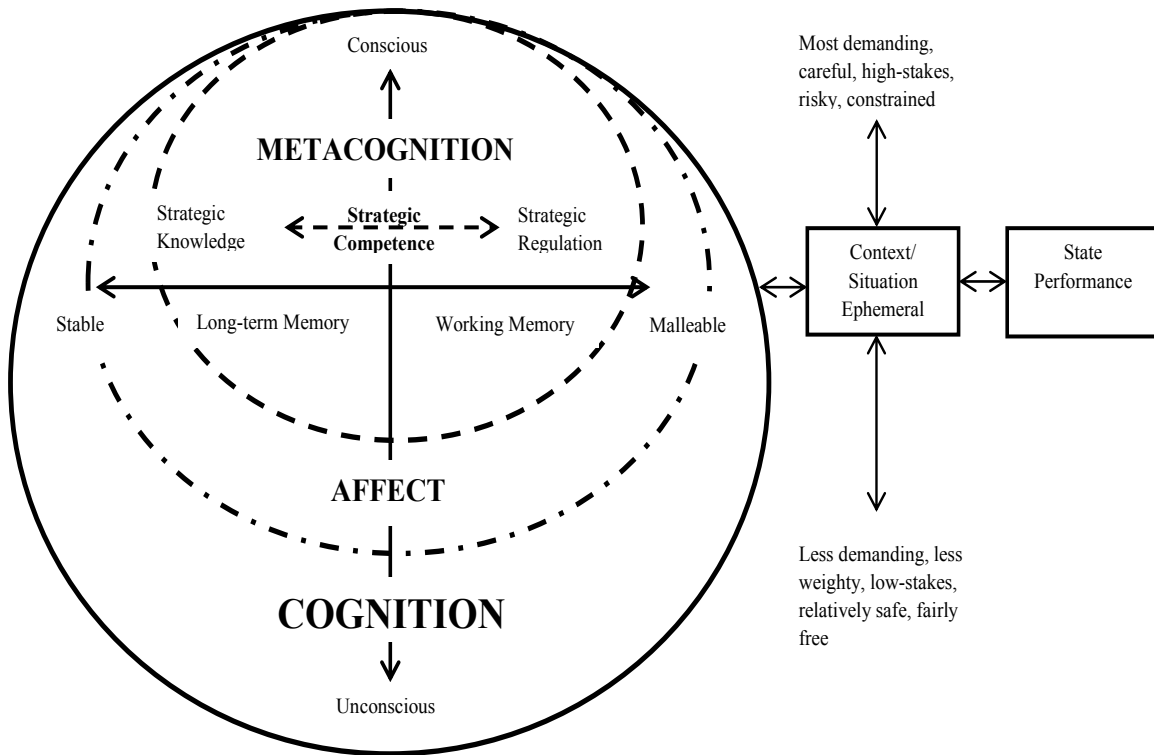


Figure 1: A multidimensional model of strategic competence (Phakiti 2007b, p. 152)

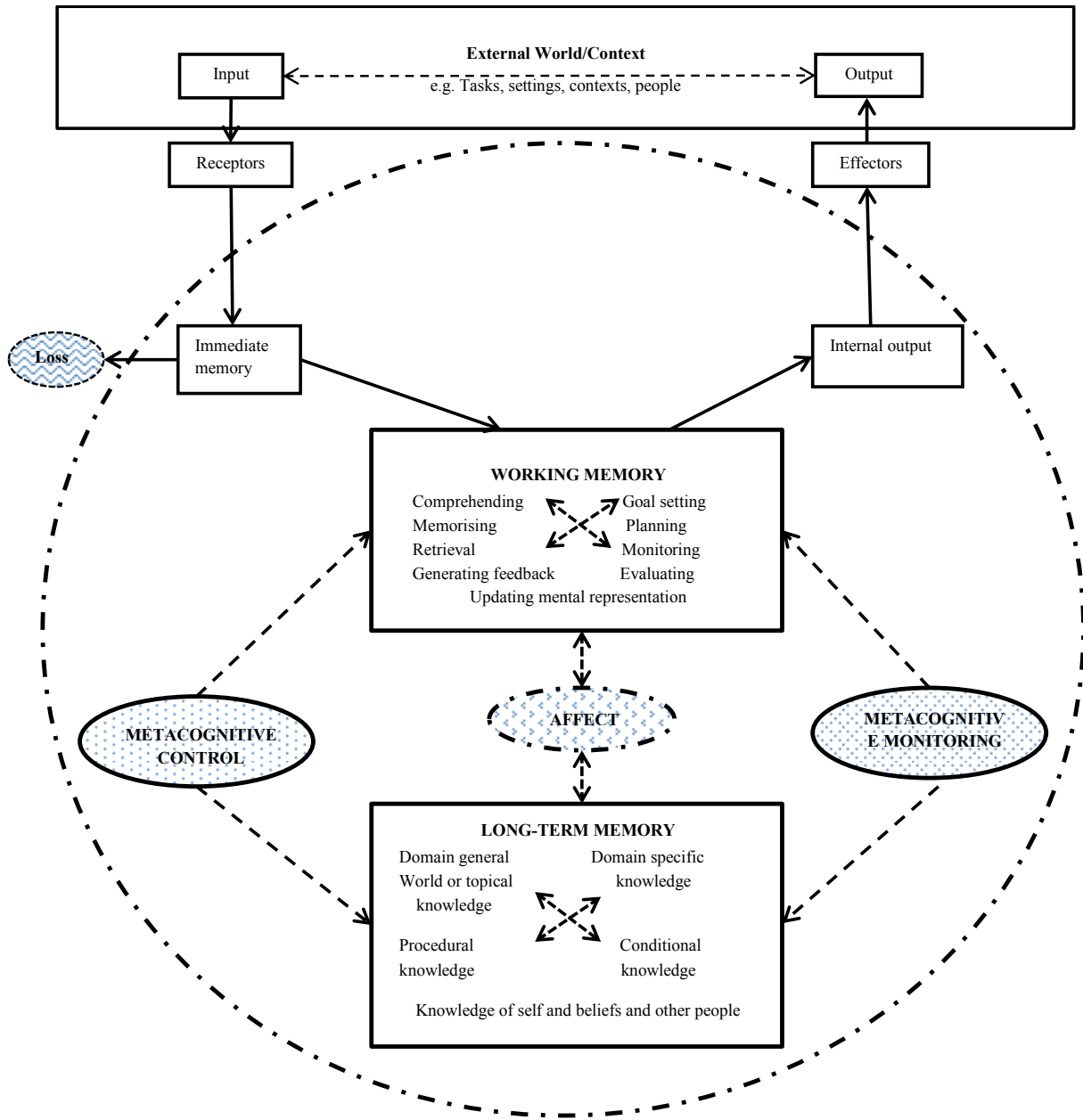


Figure 2: Human information processing (Phakiti 2007b, p. 157)

2.4.2.2 Appraisal calibration

Research into the quality of students' performance appraisals is known as calibration research. Performance appraisals can be ongoing control processes that form part of metacognitive control and monitoring (see Figure 2). To be consistent throughout the report, the term 'performance appraisals' is used instead of metacognitive judgments – the term often used in psychological research on calibration. Students' performance appraisals can be defined as individuals' engagement in evaluating their current performance (e.g., how am I doing in this test task?) and the accuracy of their performance appraisals (see e.g., Ackerman & Wolman 2007; Dinsmore & Parkinson 2013; Koriat 2011). Performance appraisals are related to self-monitoring and evaluation, which are part of individuals' metacognition and self-regulation. According to Hattie (2013), students differ in the amount of performance appraisals they engage in during their learning and many will not regulate their action when they do not have any knowledge about their current performance. Research in educational psychology has found that students' self-assessment and evaluative behaviour during study time is related to their learning success (Dunlosky, Serra, Matvey & Rawson 2005).

Successful students engage in judgment of learning, which is then transferred to what they do during test taking (e.g., asking whether they have answered each question correctly, or in sufficient detail). According to Crisp, Sweiry, Ahmed and Pollitt (2008), students should begin to develop an understanding of test expectations since students take numerous tests. When students study for a test, they will be strategic and purposeful in their study preparation. That is, while studying for a test, students will evaluate whether they have adequately learnt what will be tested (e.g., do I know enough to pass or perform highly in the test?). In a test situation, students' performance appraisals take place with respect to their personal criteria for correctness or appropriateness (see Schraw 2009). Given that students' performance appraisals are subjective in nature, it is important that they are as valid and accurate as possible (Labuhn, Zimmerman & Hasselhorn 2010).

Of course, people vary in terms of what and how they appraise their performance. Performance appraisals can be expressed in the forms of levels of happiness (not at all happy to very happy), satisfaction (very dissatisfied to very satisfied)

or perceived certainty or probability of success (not at all sure to very sure). They can be quantified as high, medium, or low. In calibration research, a probability as expressed through appraisal confidence ratings in percentages is usually adopted and treated as individuals' subjective feelings about the probability of their performance being successful (e.g., Bøjorkman, 1994; Gutierrez & Schaw 2014; Schraw, Kuch & Gutierrez 2013; Yates, Lee & Shinotsuka 1996). Humans have the ability to use ratio scales to estimate their confidence (e.g., Edwards 1967; Gigerenzer, Hoffrage & Kleinbölting 1991; Stone 2000).

In order to examine calibration, two variables are needed: appraisal confidence in the correctness of performance, and actual performance judged by external standards. This confidence in turn is treated as a probability, which is considered subjective in the sense that different individuals may have different probabilities about their degree of success for the same test question. This technique of assessing appraisal confidence is known as a micro-analytic calibration technique (Cleary, 2009, 2011).

In the current study, appraisal confidence scales were initially calculated on the basis of the number of given possible responses to a multiple-choice question. As a rule of thumb, the starting point (the lowest) on an appraisal confidence rating scale depends on the number of alternatives (k) given to a question (i.e., $100/k$) (Kleitman & Stankov 2001). So, if there are four options to a question, a confidence rating scale will include 25%, 50%, 75% or 100%. However, as pointed out by Phakiti (2005), there is a need to distinguish the chance of getting the answer correct (i.e., 25%, 50%, 75% or 100%) from the actual confidence in performance (0% to 100%). That is, when individuals know that they have 25% chance of getting the answer correct, it does not necessarily imply that their lowest appraisal confidence will be 25%. This recognition has led to an inclusion of 0% in the confidence scale, so that individuals will be calibrated when their performance is 0% and their confidence is also 0%.

An exclusion of 0% would lead to an unrealistic inflation of their appraisal confidence levels. In the current study, a 90% confidence option was also added because test-takers' appraisal confidence may be high, but not as high as 100%, or as low as 75%. The addition of 90% does not fit in the rule of thumb stated. It should be noted that there is no perfect appraisal confidence rating scale.

Findings in appraisal calibration can still be influenced by the artefact of any scales being used due to the chance factor (guessing factor). For example, if test-takers guess an answer to the question with zero appraisal confidence and if they are correct, they are considered miscalibrated and underconfident. Findings in appraisal calibration can also be complex when test-takers eliminate unlikely answers (intelligent guessing). If their answer is correct, they are likely to rate an appraisal confidence lower than 100% because of the presence of uncertainty of being correct, thereby being underconfident.

Figure 3 presents a local mental model (LMM) and a probability mental model (PMM) of how an individual rates their appraisal confidence in test correctness (in a multiple-choice context). Phakiti (2005) has revised and extended a model by Gigerenzer, Hoffrage and Kleinbölting (1991). Gigerenzer et al.'s (1991) theory of LMM and PMM was supported by several empirical studies (see Juslin 1994; Schneider 1995). Gigerenzer et al.'s (1991) model of realism has merit to advance our knowledge of appraisal calibration research in L2 testing and assessment contexts. It should be noted that Gigerenzer et al. (1991) did not include 0% on the confidence scale they used, nor an internal and external feedback loop as part of information processing in their model. Figure 3 is particularly useful to help researchers understand how appraisal confidence may be generated; it is briefly described below (see Phakiti 2005 for further detail). Although Figure 3 refers to multiple-choice test situations, it could be extended to construct-response test tasks in which test-takers generate their own possible cues or answers using available sources.

2.4.2.3 The local mental model (LMM)

When L2 test-takers are presented with a test task (e.g., listening text, a set of questions), they initially attempt to construct a local mental model (LMM) of the task. This attempt is related to memory searching and rudimentary logical operations (Gigerenzer et al. 2001). When test-takers can recall the exact relevant knowledge, they will have sufficient evidence for the answer and have an appraisal confidence level of 100%. According to Gigerenzer et al. (2001), an LMM can be successfully constructed in one of the following

three conditions: (1) the knowledge can be retrieved from memory for all the alternative responses; (2) the intervals do not overlap and can be retrieved mentally; and (3) elementary logical operations, such as the method of exclusion, can compensate for any missing knowledge. Performance appraisals within this LMM can be highly automatic (little conscious attention is involved) because information processing is fast and so individuals are not necessarily aware of their appraisal confidence level.

However, they can report their appraisal confidence level in the correctness of their answer if they are asked to do so. Considering a task completion in reference to LMM is useful to explain why sometimes people do not think about their appraisal confidence. Language tasks in the local mental model are usually easy tasks that do not require complex processes to complete. In the context of L2 use, when a person has mastered the target language necessary for use and is performing the task in a familiar environment (after extensive practice or in the areas of expertise), appraisal confidence can be tacit or implicit (Phakiti 2005; see also Figure 1).

Following Gigerenzer et al. (2001), Phakiti (2005) further specifies the parameters of an LMM as follows: first, the LMM needs to be viewed as local because in a four-option, multiple-choice test, a person needs to take the four given alternatives into account. Second, it is direct because it involves only the target variable and hence no probability cues are required to be generated. Third, it needs to be assumed that no complex inferences, besides those involving fundamental operations of deductive logic, take place. During this information processing, in some instances, it may take time for test-takers to retrieve information from memory. If this is the case, they may implicitly or subconsciously invoke internal and external feedback. Finally, if the search for information or a highly valid cue to answer the test question is successful, the confidence in the knowledge produced is certain or definite. According to Gigerenzer et al. (1991), within the LMM, memory can still fail and thus certain knowledge retrieved can be inaccurate. Consequently, inaccurate retrieval of information which affects appraisal confidence level can be a source of miscalibration (i.e., overconfidence) within an LMM.

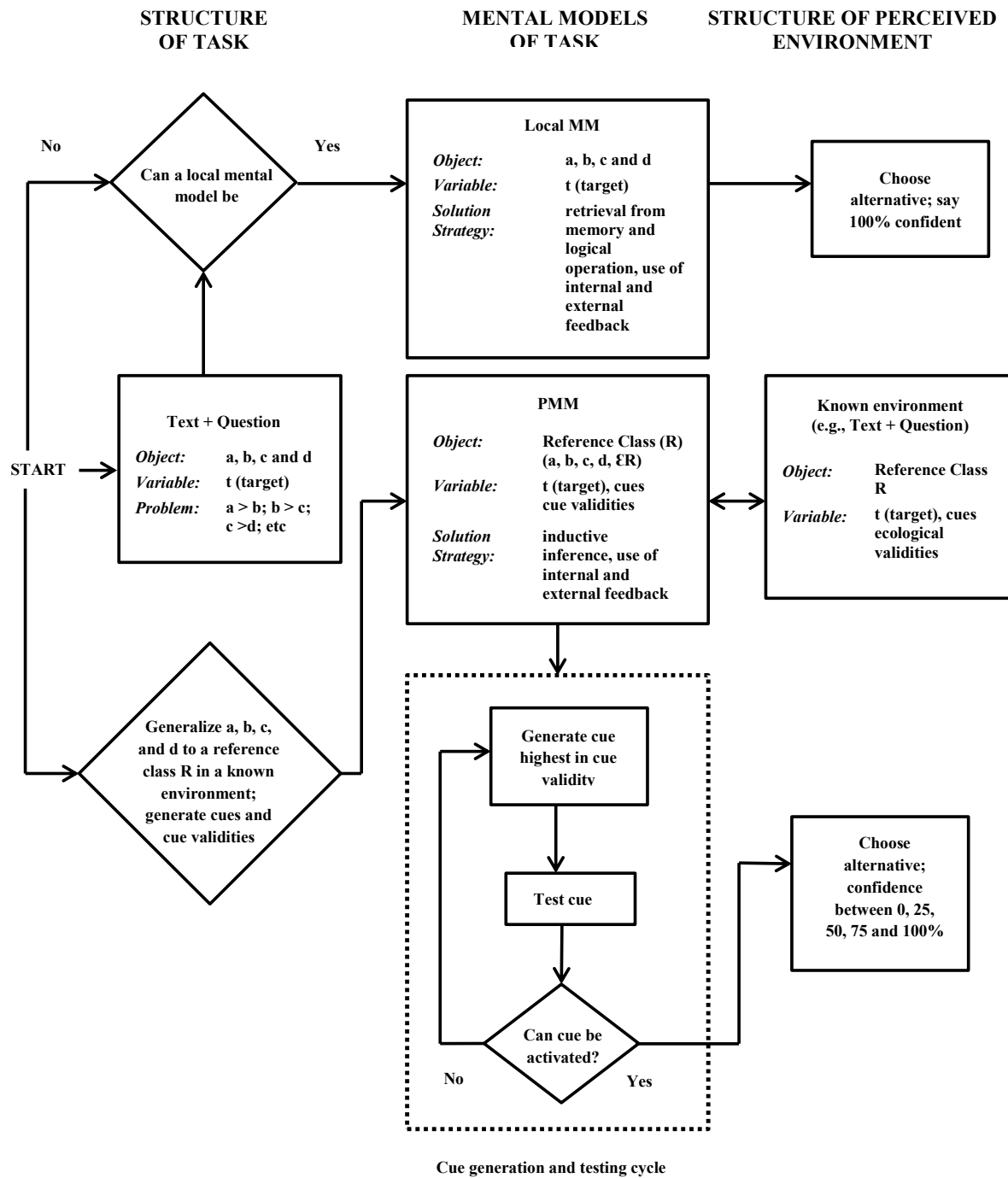


Figure 3: Cognitive processing and confidence level generation in solving a multiple-choice test task (adapted from Gigerenzer et al. 1991 by Phakiti 2005, p. 30)

On the basis of the LMM, if an L2 test-taker has the adequate linguistic knowledge required by the given task, he/she will construct their mental model at the local cognitive level and their appraisal confidence will be generated locally. In Figure 3, appraisal confidence generation requires two cognitive stages. In the LMM, confidence generation involves: (1) searching knowledge or understanding as the task is attempted; (2) deciding the most appropriate answer; (3) evaluating the validity of the chosen answer; and (4) rating confidence in the given response. In the PMM, the test-taker needs to: (1) retrieve from memory a subset of the available cues (e.g., the frequencies with which a given combination of cues predicts the right and wrong answers); (2) aggregate or prioritise the cue validities which eventually result in an internal feeling of knowing; and (3) express this internal feeling of knowing in terms of numerical probability through a confidence rating.

To illustrate how a person may construct a mental model in a listening test, the following example is used (Cullen, French & Jakeman 2014, p. 17).

A test-taker hears: We had a good response to our survey and we found that, while 80% of our students drink coffee, only 15% drink tea, with the rest preferring water.

Question: The survey found that the majority of students drink

A. water B. coffee C. tea

0%	25%	50%	75%	90%	100%
----	-----	-----	-----	-----	------

In this example, if the test-takers have all the required linguistic knowledge for this question, they are most likely to choose B and provide a 100% confidence rating. However, if they heard 18% instead of 80%, they would likely choose A with a 100% confidence. If they were unsure about what was heard or do not fully understand the given choices (see Figure 3, if “no”), then they might construct a probabilistic mental model.

2.4.2.4 The probabilistic mental model (PMM)

Performance appraisals as translated into appraisal confidence ratings within the PPM can differ from those in the local mental model because they are generated on a probability basis. Gigerenzer et al.’s (1991) PMM theory uses the following terms to explain the appraisal confidence generation phenomenon: (1) a reference class of objects that

are mentioned in the test questions and tasks; and (2) a target variable that represents a category of interest within a given test task or situation. To continue to complete a test task, a person uses a PMM as the basis for a process of inductive inferences by employing a network of other variables. Inductive inferences include: using all possible contextual information mentioned in the test items as logical evidence to make sense of, or to interpret, meanings as a method of answering the test question; guessing word meaning; predicting outcomes; supplying missing information; determining the author’s tone, and so on. Such a network of inferences represents the probability cue that is then used to discriminate between given alternatives (in the case of a multiple-choice test).

During PMM processing, each of the probability cues for the answers has a different level of *validity* for the target variable (i.e., the desired answer or performance). That is, individuals define the probability representation associated with the most likely answer on the basis of their own perceived cognitive feedback and the contextual resources available to them. It is only when a highly valid cue matches the ecological validity that corresponds to the correct answer that their PMM operations lead to the correct answer. If their appraisal confidence is high, they are likely to be calibrated. If they think that the generated probability cue is not ecologically valid, and if their appraisal confidence is also low, they are likely to be calibrated as well. On the basis of this theory, test task completion based on invalid cues and invalid confidence can be a source of miscalibration within the PMM.

Testing and evaluating cues requires monitoring and evaluating processes that result in cognitive feedback into their information processing loop. If the number of problems or test questions is large and there is an element of time pressure, and if the activation rate of cues is low, then the cue generation and testing cycle ends soon after the first cue activated has been found (Gigerenzer et al. 1991). For realistic learners, if no cue can be activated within this attempt, it can be assumed that a test-taker’s answer is constructed randomly and a 0% (or 25 %) appraisal confidence should be provided.

2.4.2.5 Internal and external feedback

Feedback can dramatically influence appraisal confidence (Butler & Winne 1995; Stone 2000). One primary role of feedback in relation to a person’s appraisal calibration is to improve the quality of performance and accuracy of appraisal confidence ratings.

Internal feedback (i.e., internally self-generated feedback within an individual during task engagement) includes: (1) subjective judgments of success in the given task in regards to the desired goals; (2) judgments of a relative productivity of various cognitive processes (e.g., strategies along with expected rates of progress); and (3) positive or negative feelings associated with knowledge and performance outcomes. External feedback includes: (1) outcome feedback (e.g., an indication of right or wrong answers on the basis of the best information cue obtained); and (2) cognitive feedback (e.g., valid reasons for good or bad performance).

In a real test situation, however, test-takers may not have direct access to cognitive feedback from an outsider. However, through social interactions prior to the test, they may have received some form of cognitive feedback, meaning that they might be able to generate their own cognitive feedback during test completion. Cognitive feedback can be expected to have a significant impact on performance appraisals during cognitive engagements in a test, whereas outcome feedback tends to impact appraisal confidence in overall achievement. When external feedback enhances internal feedback, individuals engage in better self-monitoring, self-testing, and performance appraisals. Without sufficient feedback, individuals can fail to adjust their information processing because when the test task difficulty increases, test-takers may be overconfident in their performance.

It can be argued that in high-stakes situations, high validity of appraisal confidence is pivotal because good appraisal confidence is by itself feedback for further use during test completion. If test-takers realise that their appraisal confidence is low, they may be aware of the need to call for strategies that may help to improve their test performance.

It can be argued that in both LMM and PMM, internal feedback plays a crucial role in producing successful performance and accurate appraisal confidence levels as the result of performance appraisals. According to Phakiti (2005), the use of feedback within the LMM and the PMM can, however, differ significantly. Internal feedback in the LMM can be implicit and more automatic than that within the PMM because individuals do not need to test the many generated cues and their hierarchical validities before making a decision on the correct answer.

2.4.2.6 Two types of appraisal confidence

Performance appraisals can be measured through two types of confidence: *single-case* appraisal confidence of each test item, and *relative-frequency* appraisal confidence of an overall test performance. According to Gigerenzer et al. (1991), these two categories rely on different cognitive processes and should not be correlated. On the one hand, single-case appraisal confidence is largely determined by test-takers' perceived knowledge about the answer to a question and the available choices at the time of completing that question. During human information processing, single-case appraisal confidence is related to specific monitoring processes that provide internal feedback for the specific cognitive processing required to deal with the given tasks at hand. On the other hand, relative-frequency appraisal confidence is associated with the overall number of questions test-takers have completed and thought of being correctly answered.

Kleitman and Stankov (2001) pointed out that relative-frequency appraisal confidence is influenced by contextual factors pertaining to the entire test (e.g. test instructions, the characteristics of test item questions, and time constraints). Relative-frequency appraisal confidence is related to the overall internal feedback that allows test-takers to self-reflect on how well they have performed in the whole test.

It is, however, empirically unclear by Gigerenzer et al. (1991) why single-case appraisal confidence should not be related to relative-frequency appraisal confidence, especially when relative-frequency appraisal confidence is reported after single-case appraisal confidence.

2.4.3 Empirical findings about individuals' appraisal calibration

In calibration research, individuals are asked to complete a series of test questions and provide their appraisal confidence levels in the correctness of their answers using appraisal confidence scales. Generally speaking, research has found that the relationship between students' appraisal confidence in test correctness and actual test correctness is weak to moderate (e.g., r values were around 0.20 in Epstein, Glenberg & Bradley, 1984; Glenberg & Epstein 1985; less than 0.35 in Maki & Serra 1992; and as large as 0.69 in Weaver & Bryant 1995).

A literature review by Hattie (2013) suggests that people are usually found to be overconfident in their performance. Appraisal calibration researchers have examined individual and environmental factors that may influence the nature of test-takers calibration or miscalibration (see e.g., Schraw et al. 2013). For example, previous research found that individual factors, such as the lack of required knowledge, motivation, and inaccurate activation of prior knowledge, influenced the accuracy of students' appraisals (e.g., Björkman 1994; Juslin, Winman & Persson 1995; van Loon, de Bruin, Van Gog & van Merriënboer 2013). Contextual factors, such as the context, task characteristics (e.g., linguistic and task complexity and allowed time), and measurement instruments have also been found to influence individuals' appraisals (e.g., Johnson & Bruce 2001; Kleitman & Stankov 2001).

In addition to individual factors which affect individuals' appraisal calibration, calibration research has usually found an intriguing phenomenon that people demonstrate when they encounter different levels of test task difficulty. This phenomenon is known as the *hard-easy effect* by which people tend to be overconfident in difficult questions (when they should be underconfident), but to be underconfident in easy questions (when they should be realistic or overconfident; see Kleitman & Stankov 2001; Stone 2000). In the context of language testing and assessment, it is worthwhile examining whether the hard-easy effect exists because language tests typically demonstrate a range of test item difficulty levels. The hard-easy effect has been viewed as an external factor that explains individuals' poor appraisal calibration.

It is important to note that what is seen to be difficult by a language tester may not necessarily be considered difficult by a specific group of test-takers (Bachman 2000). Test difficulty levels are arguably relative to a specific group of test-takers. In order to examine the effect of test difficulty levels on test-takers' appraisal calibration, Rasch Item Response Theory (IRT) analysis can be used since it identifies test item difficulty in relation to a specific group of test-takers. The majority of previous calibration studies in psychology did not employ Rasch IRT to identify a level of test item difficulty in a test and use this information to evaluate test-takers' calibration (see Cummings 2006).

2.4.4 Research on test-takers' appraisal calibration in language testing and assessment

In the context of language testing and assessment, it is important to know whether test-takers' language ability levels play a crucial role in influencing the accuracy of their performance appraisals. Since test-taking strategy research typically finds differences between high-ability and low-ability students in the quantity and quality of strategy use (see Cohen 2011), it can be hypothesised that high-achieving test-takers are better calibrated than low-achieving ones. Some appraisal calibration research which examined the calibration of experts (e.g., lawyers, doctors) in their occupational areas often found that experts are typically well-calibrated and are often underconfident in their appraisal confidence judgments. For example, meteorologists can in most cases accurately predict the weather (e.g., Murphy & Brown 1984; Murphy & Winkler 1984). It has been theorised that the reason for experts being calibrated in their predictions is that they are able to use appropriate information from the environment as the basis of their decision making in a particular case (see Kleitman & Stankov 2001).

In language testing and assessment research, an investigation into test-takers' calibration is relatively new but has begun to make its way into the mainstream language testing and assessment literature. Phakiti (2005) used both single-case confidence (i.e., confidence for each test item) and relative-frequency confidence (i.e., confidence for the overall test) to match 295 Thai test-takers' English placement test scores. Item Response Theory (IRT) was employed to investigate the test reliability including item and person fit indices, to identify question difficulty levels (e.g., easy, moderately easy and difficult test items) and to classify test-takers into ability levels. It was found that: (1) test-takers were not well calibrated, exhibiting a tendency to be overconfident; (2) high-ability test-takers were better calibrated than low-ability test-takers and tended to be underconfident in their test performance, whereas low-ability test-takers tended to be overconfident; (3) female test-takers were found to have better appraisal calibration scores than their male counterparts; and (4) test-takers tended to be overconfident in difficult questions and underconfident in easy questions.

In a small-scale study, Phakiti (2007a) reported on one of the three data sets of 22 test-takers who took an EFL multiple-choice reading test (similar to a paper-based TOEFL test). It was found that: (1) test-takers were generally calibrated, but (2) they were overconfident in difficult questions and underconfident in easy questions (the presence of the hard-easy effect).

Stankov and Lee (2008) investigated the nature of confidence judgments among 824 native speakers of English who took two reading and listening sections of the Test of English as a Foreign Language Internet-based (TOEFL iBT). The researchers focused on multiple-choice questions. It was found that test-takers were overconfident in both the reading and listening sections. Pearson-Product-Moment correlation coefficients between appraisal confidence and performance were 0.61 (reading 1), 0.52 (reading 2), 0.45 (listening 1) and 0.48 (listening 2). The researchers also found that appraisal confidence was predictive of other measures, such as ability, personality, and metacognition.

Using the same data set, Stankov, Lee and Paek (2009) examined the relationships among appraisal calibration scores (noted as realism scores) and other academic performance. They found that test-takers' appraisal calibration scores, which indicated overconfidence, had a negative correlation with other academic measures (-0.21 with high school GPA; -0.43 with SAT (Scholastic Aptitude Test) total and -0.46 with ACT (American College Test)). It should be noted that the starting points of the confidence scales in these two studies were 20%, which might indicate that the finding of overconfidence was partially an artefact of the confidence measures employed. That is, test-takers were already 20% overconfident from the start.

Using a survey method, Stankov, Lee, Luo and Hogan (2012) used a 10-point confidence rating scale from 0%, 10%, ..., 90% to 100% to examine the relationships between confidence and accuracy in grammar, vocabulary and reading tests among 1,940 participants. It was found that the correlations between confidence and accuracy were 0.48 (English Grammar), 0.56 (English vocabulary), and 0.49 (English reading comprehension).

Though a structural equation modeling (SEM) approach, Phakiti (2016) examined the relationships between test-takers' single-case appraisal confidence, their reported strategy use and test performance among 294 Thai test-takers.

Test-takers were asked to rate their single-case appraisal confidence for each test question and answer the strategy use questionnaire at the end of the test. It was found that test-takers' appraisal calibration scores, as well as their single-case appraisal confidence scores, were moderately related to reported metacognitive strategy use. The influence of the use of cognitive strategies on test performance was found to be at its lowest when single-case appraisal confidence was simultaneously added in the SEM model. The finding might imply that inaccurate performance appraisals might result in poor use of cognitive and metacognitive strategies, which then could not enhance test performance. Phakiti points out that poor appraisal calibration could potentially limit the role of cognitive and metacognitive strategies in enhancing test performance.

It should be noted that to extend the work of Phakiti (2016), which only focused on state strategy use, the present study focuses on both trait and state strategy use.

2.4.5 Implications for the present study

2.4.5.1 Research problems

The review of the literature suggests that numerous test-taking strategy studies have largely aimed to clarify the role of strategic competence in language tests by examining reported strategies through the use of Likert-type scale questionnaires administered before or after a language test. If validation research on the cognitive validity of the IELTS Listening task is to progress further, empirical research must look into the key cognitive processes underlying test performance beyond the mere use of self-report, context-free strategy questionnaires or a small-scale introspective study to validate the test tasks.

It is important that the validity of test-takers' performance appraisals are examined because appraisal calibration may further explain test performance differences among test-takers. However, in a listening test, for example, performance appraisals cannot be properly examined using a questionnaire technique, which may be given before or after test-takers have completed the entire test section or test. Performance appraisals take place concurrently during listening and task completion and hence need to be measured promptly. To date, little is known about L2 test-takers' appraisal confidence judgment and calibration. A review of research on individuals' appraisal calibration in educational

and cognitive psychology has both theoretical and methodological applications for how an evaluation of language test-takers' performance appraisals can be studied.

Empirical evidence across calibration research in various academic disciplines shows that the ability to be calibrated (i.e., to be realistic) is essential in the face of test or task difficulty. Phakiti (2016) argues that the accuracy of test-takers' appraisal confidence is critical not only to strategic processes but also to test success. Few have theoretically or empirically investigated this appraisal calibration construct in language testing and assessment research. Furthermore, little is known about how performance appraisals and appraisal calibration are related to test-takers' reported strategy use and perceived difficulty in a test.

Furthermore, Phakiti (2007b, 2008a) and Bi (2014) have demonstrated the importance of trait strategy use (as related to strategic knowledge) and state strategy use (as related to strategic regulation) when explaining test performance variances. Perceived IELTS listening difficulty may also influence confidence in performance during test taking. It should be noted that perceived difficulty is not the same as test anxiety, although the two constructs may be related. Winke and Lim (2014), who examined the effects of testwiseness and test-taking anxiety on IELTS Listening test performance, found no significant relationship between the two constructs, but found that anxiety negatively correlated with IELTS Listening test performance. The current study aims to examine the relationship between test-takers' trait and state IELTS listening difficulty and their appraisal confidence, appraisal calibration and IELTS Listening test performance.

2.4.5.2 Research questions

Five research questions are asked:

1. What is the nature of test-takers' appraisal confidence and calibration in an IELTS Listening test?
2. What is the nature of test-takers' appraisal confidence and appraisal calibration in easy, moderately difficult, very difficult and extremely difficult IELTS Listening questions?
3. Do male and female test-takers differ in their appraisal confidence and appraisal calibration scores in an IELTS Listening test?
4. Do test-takers with different success levels differ in their appraisal calibration scores?

5. What are the structural relationships among test-takers' appraisal confidence, appraisal calibration, trait and state cognitive and metacognitive strategy use, trait and state IELTS Listening test difficulty, and IELTS Listening performance?

3 RESEARCH QUESTIONS

3.1 Research context

The present study focuses on calibration and strategy use by NESB international students in Australia while completing a simulated IELTS Listening test. International students are defined as those who do not hold citizenship or a permanent residence visa in the country where they are studying. They generally need to obtain a student visa prior to commencing their study. Furthermore, international students pay full fees, or partial / no fees if they are the recipients of a scholarship. According to Ramachandran (2011), they come to a host educational institute for a set time period.

According to Andrade (2010), the number of international students who travel abroad to study has increased significantly in the past few decades (see <https://internationaleducation.gov.au/research/international-student-data/pages/default.aspx>). In 2010, approximately 4.1 million international students were enrolled in higher education programs outside their home countries (Organization for Economic Cooperation and Development (OECD) 2012) and the number is projected to reach 8 million by 2020 (Forest & Altbach 2006). Andrade (2006, 2010) pointed out that many English-speaking countries, including the US, the UK, Australia, and Canada, have adopted policies to increase their international enrolment numbers as a national priority.

The present study focuses on NESB international students in Australia who study academic English as part of their preparation for university admission. This group of students is of interest because they are typical representatives of IELTS test-takers.

3.2 Research design

In order to address the research questions, the present study was designed to be a quantitative study. Quantitative research requires measurement of research constructs of interest. Because the key focus of the study is on test-takers' concurrent appraisal confidence in test performance, a standard IELTS Listening test procedure as carried out by the official test administration guidelines cannot be followed.

To achieve the research aim, a quasi-experimental design is adopted. In this design, all participants received the same IELTS test-taking conditions. The participants were given equal amounts of time to: (1) work on test material to help them become familiar with confidence rating during test taking; (2) complete each section and transfer their answers to the answer sheet; and (3) answer the listening strategy use questionnaires (discussed further below).

There is no control group in this quasi-experimental design because the present study does not aim to compare test-takers' confidence or performance under different test conditions or in the case of different manipulations of identified independent variables. However, statistical comparisons of appraisal confidence, appraisal calibration, and IELTS Listening test performance were performed in relation to test-takers' success in the given listening test and pre-identified proficiency levels, test difficulty levels and genders (males versus females).

3.3 Ethical considerations

As the study involved human participants, ethics approval from the University of Sydney was sought prior to the data collection (Project No. = 2014/846). The data collection procedure strictly followed the ethical protocols as approved by the University Ethics Committee. Participation in this research project was voluntary and participants' personal information was kept confidential.

All participants were unidentifiable and it was agreed that pseudonyms would be used if any participants were ever referred to. The data were used solely for this research purpose.

Prior to data collection, an official meeting was held with the participating language institutes (discussed below) regarding the research procedures that would be followed. Only when these were agreed to be acceptable did the authorities allow the researcher to conduct the study at their institutes.

The participants in this research were informed of the purpose of the research and of the procedures that would be followed. They were also informed of what the project aimed to achieve. All participants took part in the study on a voluntary basis. They were provided with the participant information statement and were required to sign a participant consent form. They were informed that they had the right to withdraw themselves or their associated data at any time.

The researcher respected the participants' choice to decline to participate in or to withdraw from the research at any time. Incentives to participate in the study included an opportunity to win a prize (an iPad), and individualised feedback on test performance and calibration scores.

3.4 Research settings

For the purpose of confidentiality and anonymity, the participating institutes cannot be named. The study took place in two institutes (across three different programs) in Sydney, Australia. The first institute (Institute A) is part of a major university that contributes to the university-wide effort to provide academic and language support to international students. Institute A provides preparatory English language courses and pathway programs to undergraduate and postgraduate degrees. Such pathway programs are typically designed to enhance students' academic language skills prior to their commencing a degree at the university. The student population consists of students largely from Asian countries, especially China (approximately 50%) and other non-Asian countries including European countries (e.g., Germany, Switzerland, and Spain) and South American countries (e.g., Argentina, Brazil, and Chile).

The second institute (Institute B) provides a range of courses including foundation studies programs, courses for high school students and general English language courses. The student population predominantly consists of students from Asian countries: approximately 90% come from China, while Saudi Arabian and Kuwaiti students make up about 5%, with the remaining coming from a variety of other Asian countries.

Students from two separate schools in Institute B took part in the study. The first school offers a foundation studies program that leads to both first and second semester entry to undergraduate courses at some universities in Australia and New Zealand. The completion of a foundation studies program provides an opportunity for NESB international students whose existing qualifications are insufficient to meet the entry requirements of an undergraduate degree to meet those requirements. The second school provides an international English language program and has up to 400 students at any one time. Students study at this school for a pre-determined length of time, from 4 weeks to 52 weeks. Students' ages at entry range from 16 to 25 years, with a median age of 20 years.

Students who successfully complete a language course at this school are able progress to a university foundation program at Institute B.

3.5 Participants

Originally 400 NESB international students from the two institutes agreed to participate in the study. All participants had experience taking the official IELTS test prior to the present study. However, data from only 376 test-takers were used in the study as 24 participants were excluded from the data sets because of (1) incomplete questionnaires, incomplete test question responses, and/or missing appraisal confidence ratings ($N = 11$), (2) misfitting test-takers as identified from Rasch Item Response Theory (IRT) analysis ($N = 10$) and (3) multivariate outliers as identified through the EQS 6.2 program ($N = 2$).

It should be noted that 225 test-takers (60%) of the total 376 reported their previous overall IELTS scores (mean = 5.67; SD = 0.75) and IELTS Listening score (mean = 5.71; SD = 0.94).

Overall, participants comprised: 141 people (37.5%) from Institute A; 142 (37.8%) from Institute B (first school); and 93 (24.7%) from Institute B (second school).

There were 138 males (37%) and 238 females (63%) in the study. The test-takers were between the ages of 18 and 45 ($M = 20.74$; $SD = 4.84$).

The participants included 218 students from China (58%), 21 students from Columbia (5.6%), 21 students from Saudi Arabia (5.6%), 18 students from Japan (4.8%) and 17 students from South Korea (4.5%). The remaining students were from Hong Kong, Taiwan, France, Venezuela, Chile, Italy, Spain, Turkey, Switzerland, Argentina, Germany, Mexico, Ecuador, Iraq, Indonesia, Vietnam, Qatar, Lao, Bangladesh, Sri Lanka and Myanmar (together these made up 22.3% of participants).

Students' first languages included Chinese (57.4%), Spanish (12.8%), Arabic (6.4%), Japanese (5.1%), Cantonese (4.5%) and Korean (4.3%). The other remaining first languages were French, Italian, Portuguese, Turkish, German, Bahasa Indonesian, Vietnamese, Persian, Laotian, Bengali and Burmese (18.8%).

3.6 Research instruments

The research instruments for the study were:

- (1) a trait cognitive and metacognitive strategy use and listening test difficulty questionnaire;
- (2) a state cognitive and metacognitive strategy use and listening test difficulty questionnaire;
- (3) a simulated IELTS Listening test; and
- (4) single-case appraisal confidence and relative-frequency appraisal confidence scales. Each of these instruments is discussed as follows.

3.6.1 Trait and state cognitive and metacognitive strategy use and IELTS listening test difficulty questionnaires

Trait and state listening strategy use in this study is based on the theory of human information processing (Gagné, Yekovich & Yekovich 1993) and Phakiti (2007b), which views human information processing as having a *structural component* of sensory receptors, and working and long-term memory arrays, and a *functional component* of information processing that describes the operations of comprehending, remembering, retrieving and controlling processes at different stages. In this study, only cognitive and metacognitive listening strategies were examined. These strategies were selected from the literature (e.g., Bachman & Palmer 2010; Badger & Yan 2009; Cohen 2011; Field 2009, 2013; Phakiti 2007b; Vandergrift & Goh 2012).

Both cognitive and metacognitive strategies are contextualised in this study to be related to IELTS Listening tests. Additionally, trait and state IELTS Listening difficulty items were included in the two questionnaires in order to examine how test-takers perceive the level of IELTS Listening difficulty generally (trait) and specifically (state) after completion of the given test.

Appendix 1 provides details of the research instruments used in the study. The trait strategy use questionnaire is written using the Simple Present as it asks test-takers about their generally perceived strategy use and IELTS Listening difficulty. The state strategy use questionnaire is written using the Simple Past as it asks test-takers about their strategic thinking and perceived difficulty while they were taking the test. An example of a trait planning strategy in an IELTS Listening test situation is 'I make sure I clarify what the test tasks require me to do,' and an example of a state planning strategy is 'I made sure I clarified what the test tasks required me to do'.

The state and trait strategy use questionnaires were developed and validated prior to actual use in the present study (by means of trials with 20 NESB international students not involved in the main study). Initially 35 questionnaire items were included in the pilot questionnaire. The pilot indicated that the number of items needed to be reduced because of the excessively high number of activities participants were required to do (discussed in the data collection procedure). The number of questionnaires items was reduced to 27 items for the main study. As the study aims to examine the correlation between test-takers' monitoring and evaluating strategies and their appraisal calibration, the questionnaire includes more monitoring and evaluating strategy items than cognitive strategy items.

Table 1 presents the strategy composites in the trait and state strategy questionnaires. Both questionnaires have the same taxonomy and items. The strategy questionnaires ask the participants to mark their awareness of strategy use on a 6-point Likert-type scale: 1 (Never), 2 (Rarely), 3 (Sometimes), 4 (Often), 5 (Usually), and 6 (Always). The strategy use scale defines a continuum of increasing levels of frequency/ intensity, i.e. low scores indicate a low level of awareness of strategy use and high scores indicate a high level of awareness of strategy use during test completion.

The listening difficulty questionnaires ask the participants to rate on a 5-point Likert-type scale: 1 (Not at all true), 2 (Not true), 3 (Neither), 4 (True), and 5 (Absolutely true). Low scores indicate a low level of perceived IELTS Listening difficulty and high scores indicate a high level of perceived IELTS Listening difficulty.

Scale	Subscale	No. of items	Items
1. Cognitive strategies	Comprehending	3	5, 6, 7
	Memory	2	8, 10
	Retrieval	3	9, 11, 12
2. Metacognitive strategies	Planning	4	1, 2, 3, 4
	Monitoring	5	13, 14, 15, 16, 17
	Evaluating	5	18, 19, 20, 21, 22
3. IELTS Listening difficulties	IELTS Listening difficulties	5	23, 24, 25, 26, 27
	Total	27	

Table 1: Taxonomy of the trait and state cognitive and metacognitive strategy use and IELTS Listening test difficulty questionnaires

3.6.2 The simulated IELTS Listening test

The IELTS Listening test is composed of four major sections, with a total of 40 questions. Sections 1 and 2 are related to social contexts, whereas Sections 3 and 4 are related to academic contexts which reflect authentic situations. Test-takers hear the audio text only once for each section. One test section may consist of more than one test format, which may result in different cognitive processes and mental model constructions being used. The test formats are mainly short-answer questions (2-3 words), information transfer tasks, and multiple-choice questions. Each question is worth one mark. An IELTS Listening test can be considered a complex and demanding cognitive activity which requires test-takers to listen to audio text, while at the same time reading printed tasks (choices, tasks). Test-takers construct or choose an appropriate answer as they listen. Table 2 summarises the four sections of the IELTS Listening test.

Originally an IELTS Listening test (IELTS 9, published by the University of Cambridge Local Examination Syndicate) was chosen for this research project. However, from the pilot study in which all the research instruments were tried out, it was discovered that some participants had recently practised that particular test. Their memory of the test was reported to affect the way they completed the test and rated their appraisal confidence. It was then decided to use sections from several different tests (IELTS 6, 7 and 9), so that no single student would remember the entire test.

A confidence rating scale was provided for the answer to each question. Participants were asked to rate their appraisal confidence in their answer immediately after they had answered each question. In order to make sure that participants understood how to rate their confidence, an explanation of the confidence scale and a practice test with confidence ratings were provided prior to the actual test (see the data collection procedure below). Participants were encouraged to engage in the listening texts and to complete the given test tasks, rather than trying to rely on what they might have remembered if they had done a particular section previously. It should be noted that there was no impact of the test results on participants' grades in the course in which they were enrolled. Later, participants were informed of their test scores as well as their calibration scores and feedback on their appraisal calibration (see Appendix 1: Example of feedback to students).

Section	Focus	Skills and purpose	Topic	Type of task
1	Conversational/ Transactional (2 speakers)	<ul style="list-style-type: none"> • Basic social/survival skills • Study-related language use • Ability to follow and respond to instructions • Ability to retrieve and extract explicitly stated information 	General topic (e.g., transportation, a product, restaurant, accommodation)	<ul style="list-style-type: none"> • Information transfer: Within the word limit, complete/label notes, a summary, a table, a diagram or chart, a map or a plan. • Multiple-choice questions: Choose an alternative from a multiple-choice question. • Short-answer questions: Provide a short answer (within the word limit) to a question. • Matching tasks: Match listed statements to possible answers. • Classification tasks: Classify the information provided in the question.
2	Monologue/ Transactional (1 speaker)		General topic (e.g., touring, holiday plan, camping, transportation)	
3	Conversation/ Academic (2+ speakers)		Academic/ topic (e.g., workplace, place, ecology)	
4	Monologue/ Academic (1 speaker)		Academic topic (e.g., history, theory, philosophy)	

Table 2: Summary of the four sections of the IELTS Listening test

3.6.3 Single-case appraisal confidence and relative-frequency appraisal confidence scales

In the present study, appraisal confidence refers to the degree to which test-takers can report on their performance appraisal in a specific test question. Appraisal confidence here is different from general self-confidence or self-efficacy (as discussed in the literature review). On the basis of previous research on calibration (see 2.4.2.2: Performance appraisals and calibration), the confidence rating scales were selected to be 0%, 25%, 50%, 75%, 90%, and 100%. These rating scales have been found to be sensitive and reliable in capturing test-takers' appraisal confidence.

In order to assess single-case confidence, appraisal confidence rating scales are provided underneath or next to each test question. Participants were asked to answer each question, after each of which they were required to provide an appraisal confidence rating in their response. At the end of each test section, participants were instructed to transfer their answers and appraisal confidence ratings to the provided answer sheet. In this way, their appraisal confidence in their test answer would be considered more current than if they had been asked to transfer them after they had completed all the four sections. They were also asked to rate their overall appraisal confidence for each of the four listening sections immediately following completion of that section.

The time allowed for participants to transfer their answers and confidence ratings for each section was three minutes. At the beginning of each section, participants were reminded to rate their appraisal confidence in their answer to a question as soon as they had answered it.

3.7 Data collection

Figure 4 summarises the data collection procedures. The overall data collection period was approximately 80 minutes. It took approximately one month to collect the data from 400 participants.

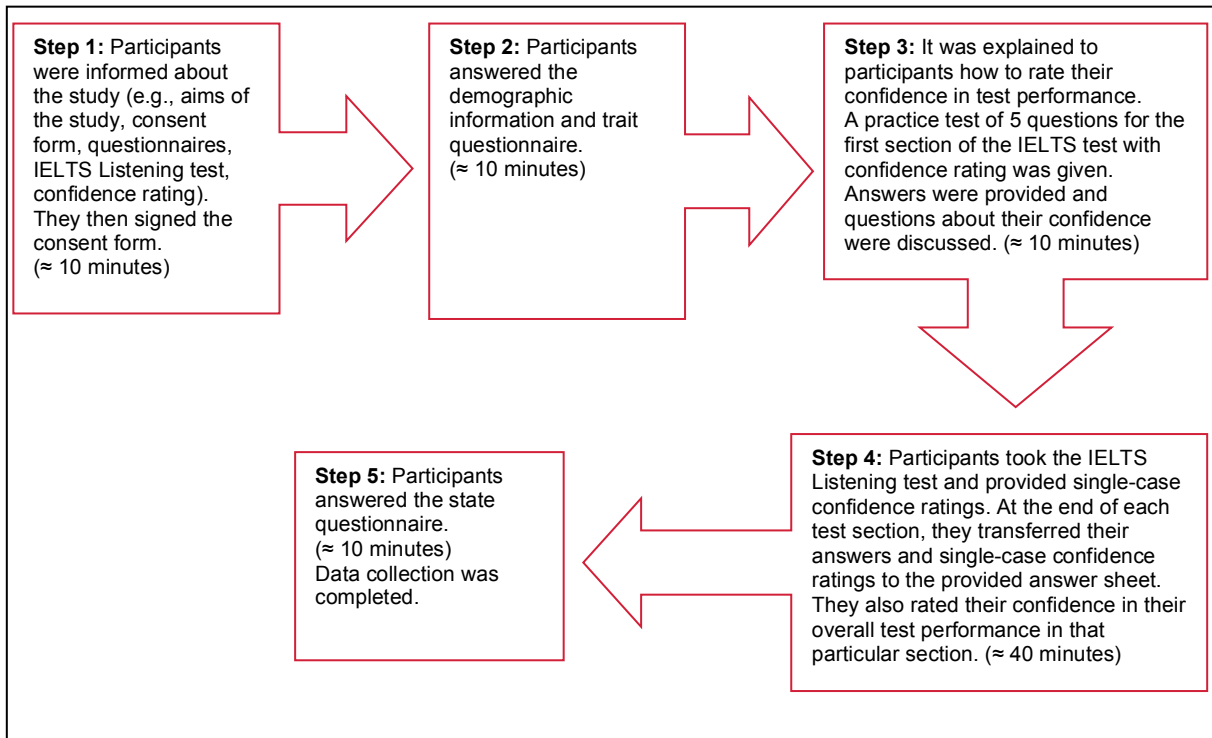


Figure 4: Flow chart of the data collection procedures

3.7.1 Appraisal confidence rating practice treatment

Step 3 of the data collection procedures included an explanation of the single-case appraisal confidence rating scale to make sure that participants understood its meaning. It should be noted that in Step 4, because test-takers were asked to rate their appraisal confidence while answering IELTS Listening test questions, their test performance might not reflect their true academic listening abilities as the rating of appraisal confidence could disrupt their listening processes. However, it should be noted that first, the appraisal confidence scales were embedded in each test question as well as on the answer sheet. Second, the Pearson-Product-Moment correlations of participants' reported official IELTS band scores and their IELTS Listening band scores to the IELTS scores in this study were 0.75 and 0.71, respectively (N = 225).¹ These correlation coefficients indicate that there is a strong relationship between test-takers' IELTS Listening performance and their previous IELTS scores. Table 3 summarises the explanations of the single-case appraisal confidence scales.

Confidence	Meaning	Explanation
0%	Extremely low confidence	You are not at all sure about your answer or you are not at all happy with your answer.
25%	Low confidence	You are not sure about your answer, but it may have a 25% chance of being correct.
50%	Medium confidence	Your answer has a 50–50% chance of being correct or incorrect.
75%	Quite high confidence	You are quite sure that your answer is correct, but there is a 25% chance it is incorrect.
90%	Very high confidence	You are quite sure that your answer is correct, but there is a 10% chance that it is incorrect.
100%	Absolute confidence	You are perfectly sure that your answer is correct.

Table 3: Single-case appraisal confidence explanations

¹ The correlation between participants' official overall IELTS score and official IELTS Listening score was 0.87 ($R^2 = 0.76$; N = 225).

Following the explanation of the single-case appraisal confidence rating scale, participants were given a practice test, which was based on the first five questions of Section 1. This practice test was based on IELTS Test 7 (Test 1) (see Appendix 1). Participants were given the answer keys for this practice test and asked whether their appraisal confidence rating for each question was realistic, overconfident or underconfident.

3.8 Data analysis

Various levels of data analysis were performed. At the item level, the questionnaire data (trait and state), IELTS Listening test data and appraisal confidence data were examined descriptively. Reliability analyses were then carried out: Rasch Item Response Theory (IRT) analysis for the listening test and Cronbach's alpha analysis for the questionnaire and appraisal confidence data. At the more global analytical level, the data were analysed to answer the research questions.

3.8.1 Item-level analysis

3.8.1.1 Analysis of the trait and state questionnaires

Prior to their use to address the research questions, descriptive statistics of the trait and state questionnaires were examined, followed by a reliability analysis (Cronbach's alpha). The aim of reliability analysis was to make sure that the underlying constructs of cognitive and metacognitive strategies and IELTS Listening difficulty were consistently captured. Exploratory factor analysis was not performed because the questionnaires were developed based on substantive theory, previous research and a pilot study. Rather, a confirmatory factor analytic approach was adopted (discussed further below). In this study, multiple observed variables were used to define a latent variable for the SEM analysis (discussed below) because by using multiple observed variables, the non-random measurement error can be estimated and evaluated. For a SEM analysis to be rigorous, data distribution and internal consistency estimates need to be evaluated. This information can be used to provide some confidence that certain assumptions, such as univariate normality, are not violated in the data set. Tables 4 and 5 report on the descriptive statistics of the trait and state strategy use and IELTS listening difficulty questionnaires. The skewness and kurtosis statistics indicate that all questionnaire items were normally distributed (i.e., the values were within ± 1).

Item	Minimum	Maximum	Mean	SD	Skewness	Kurtosis
Item 1	1.00	6.00	4.45	1.61	-0.88	-0.34
Item 2	1.00	6.00	4.64	1.32	-0.92	0.20
Item 3	1.00	6.00	4.62	1.29	-0.76	-0.07
Item 4	1.00	6.00	4.56	1.16	-0.54	-0.14
Item 5	1.00	6.00	4.28	1.29	-0.40	-0.55
Item 6	1.00	6.00	4.75	1.10	-0.64	-0.15
Item 7	1.00	6.00	4.73	1.12	-0.69	-0.05
Item 8	1.00	6.00	4.44	1.25	-0.50	-0.45
Item 9	1.00	6.00	4.81	1.16	-0.76	-0.16
Item 10	1.00	6.00	4.63	1.15	-0.56	-0.28
Item 11	1.00	6.00	4.29	1.17	-0.41	-0.13
Item 12	1.00	6.00	4.63	1.16	-0.45	-0.61
Item 13	1.00	6.00	4.35	1.26	-0.53	-0.25
Item 14	1.00	6.00	4.92	1.09	-0.88	0.33
Item 15	1.00	6.00	4.42	1.25	-0.56	-0.23
Item 16	1.00	6.00	4.83	1.18	-0.99	0.61
Item 17	1.00	6.00	4.55	1.10	-0.59	0.15
Item 18	1.00	6.00	4.23	1.14	-0.21	-0.48
Item 19	1.00	6.00	4.29	1.12	-0.30	-0.26
Item 20	1.00	6.00	4.67	1.26	-0.64	-0.49
Item 21	1.00	6.00	4.55	1.16	-0.61	-0.06
Item 22	1.00	6.00	3.99	1.36	-0.24	-0.72
Item 23	1.00	5.00	2.98	0.99	-0.10	-0.41
Item 24	1.00	5.00	2.94	1.16	0.08	-0.84
Item 25	1.00	5.00	3.03	1.08	0.00	-0.76
Item 26	1.00	5.00	2.91	1.13	0.22	-0.74
Item 27	1.00	5.00	2.74	1.16	0.26	-0.76

Table 4: Distributions for trait cognitive and metacognitive strategies and trait IELTS Listening difficulties (N = 376)

Item	Minimum	Maximum	Mean	SD	Skewness	Kurtosis
Item 1	1.00	6.00	4.59	1.40	-0.76	-0.33
Item 2	1.00	6.00	4.50	1.27	-0.45	-0.80
Item 3	1.00	6.00	4.38	1.29	-0.40	-0.57
Item 4	1.00	6.00	4.52	1.16	-0.47	-0.47
Item 5	1.00	6.00	4.23	1.26	-0.36	-0.41
Item 6	1.00	6.00	4.52	1.18	-0.43	-0.54
Item 7	1.00	6.00	4.47	1.17	-0.27	-0.89
Item 8	1.00	6.00	4.28	1.32	-0.45	-0.49
Item 9	1.00	6.00	4.34	1.30	-0.47	-0.53
Item 10	1.00	6.00	4.26	1.28	-0.36	-0.55
Item 11	1.00	6.00	4.08	1.27	-0.36	-0.46
Item 12	1.00	6.00	4.32	1.27	-0.47	-0.28
Item 13	1.00	6.00	4.28	1.35	-0.53	-0.37
Item 14	1.00	6.00	4.60	1.23	-0.61	-0.20
Item 15	1.00	6.00	4.23	1.28	-0.39	-0.40
Item 16	1.00	6.00	4.48	1.29	-0.69	-0.13
Item 17	1.00	6.00	4.41	1.26	-0.45	-0.47
Item 18	1.00	6.00	4.28	1.18	-0.29	-0.59
Item 19	1.00	6.00	4.27	1.22	-0.31	-0.58
Item 20	1.00	6.00	4.45	1.23	-0.37	-0.76
Item 21	1.00	6.00	4.28	1.31	-0.41	-0.56
Item 22	1.00	6.00	3.81	1.43	-0.16	-0.76
Item 23	1.00	5.00	3.31	1.13	-0.28	-0.70
Item 24	1.00	5.00	3.28	1.22	-0.22	-0.89
Item 25	1.00	5.00	3.26	1.09	-0.19	-0.68
Item 26	1.00	5.00	2.99	1.17	-0.01	-0.80
Item 27	1.00	5.00	3.07	1.18	-0.04	-0.89

Table 5: Distributions for state cognitive and metacognitive strategies and state IELTS Listening difficulties (N = 376)

Table 6 reports on the Cronbach's alpha coefficients for each subscale of the trait and state questionnaire. The Cronbach's alpha ranged from 0.60 (trait memory strategy) to 0.88 (IELTS Listening difficulties). The reliability coefficients were generally high and the low coefficients were within an acceptable range (0.60 or above, Pallant 2010) and suitable for the research purpose. However, the reliability coefficient of 0.70 or above is preferred. The state questionnaire was found to have higher reliability coefficients than the trait questionnaire.

Scale	Subscale	No. of items	Items	Trait Alpha	State Alpha
1. Cognitive strategies	Comprehending	3	5, 6, 7	0.67	0.78
	Memory	2	8, 10	0.60	0.68
	Retrieval	3	9, 11, 12	0.66	0.70
2. Metacognitive strategies	Planning	4	1, 2, 3, 4	0.82	0.86
	Monitoring	5	13, 14, 15, 16, 17	0.75	0.85
	Evaluating	5	18, 19, 20, 21, 22	0.74	0.82
3. IELTS Listening difficulties	IELTS Listening difficulties	5	23, 24, 25, 26, 27	0.83	0.88
			Overall	0.85	0.89

Table 6: Taxonomy of the trait and state cognitive and metacognitive strategy use and state and trait IELTS Listening test difficulty questionnaires

In order to answer the research questions related to the trait and state strategy questionnaires, a composite of each strategy category and IELTS Listening difficulties was generated. Based on Table 6, the scores from the designated strategy items were aggregated and divided by the number of items in the relevant set. For example, scores for Items 5, 6 and 7 were combined and divided by 3 to form the comprehending strategy composite or variable. It should be noted that at an item level, a questionnaire item was ordinal, but at a subscale level, questionnaire data were continuous as they were aggregated from different items. In structural equation modeling (SEM) terms, this method of aggregation is known as *item parcelling* (Little, Cunningham, Shahar & Widaman 2002). Item parcelling is desirable for statistical analysis because an observed variable to be used for inferential statistics is then made up of multiple observed items. Table 7 presents the summary descriptive statistics for the 14 composites from the trait and state questionnaires.

Item	Minimum	Maximum	Mean	SD	Skewness	Kurtosis
TCOM	2.00	6.00	4.59	0.91	-0.24	-0.56
TMEM	1.00	6.00	4.54	1.01	-0.38	-0.37
TRET	1.33	6.00	4.58	0.90	-0.39	-0.06
TPLAN	1.25	6.00	4.56	1.09	-0.63	-0.30
TMON	1.80	6.00	4.61	0.83	-0.43	-0.22
TEVA	1.60	6.00	4.34	0.84	-0.10	-0.19
TDIF	1.00	5.00	2.92	0.85	0.11	-0.44
SCOM	1.33	6.00	4.41	1.00	-0.15	-0.63
SMEM	1.00	6.00	4.27	1.14	-0.19	-0.74
SRET	1.33	6.00	4.25	1.00	-0.19	-0.42
SPLAN	1.00	6.00	4.50	1.07	-0.41	-0.54

T = trait S = state COM = comprehending MEM = memory
RET = retrieval PLAN = planning MON = monitoring EVA = evaluating

Table 7: Descriptive statistics for the trait and state cognitive and metacognitive strategies and state and trait IELTS Listening difficulties (N =376)

Item	Minimum	Maximum	Mean	SD	Skewness	Kurtosis
SMON	1.00	6.00	4.40	1.01	-0.43	-0.06
SEVA	1.80	6.00	4.21	0.98	-0.04	-0.63
SDIF	1.00	5.00	3.18	0.95	-0.15	-0.50

Table 7: Descriptive statistics for the trait and state cognitive and metacognitive strategies and state and trait IELTS Listening difficulties (N =376) (continued)

Table 8 presents the composites of the six state and trait variables with internal consistency estimates (Cronbach's alpha).

Composite	No. of items	Items used	Internal consistency
Trait cognitive strategy use	3	TCOM, TMEM, TRET	0.80
Trait metacognitive strategy use	3	TPLAN, TMON, TEVA	0.82
Trait IELTS Listening difficulty	1	TDIF	-
State cognitive strategy use	3	SCOM, SMEM, SRET	0.86
State metacognitive strategy use	3	SPLAN, SMON, SEVA	0.89
State IELTS Listening difficulty	1	SDIF	-
Total	14		0.87

Table 8: Internal consistency estimates (Cronbach's alpha) (N = 376)

3.8.1.2 Analysis of the IELTS Listening test

According to McNamara (1996) and Bond and Fox (2007), Rasch Item Response Theory (IRT) is a powerful measurement theory that can estimate both test-takers' ability levels and characteristics of the test items. The Rasch IRT model proposes a simple mathematical relationship between test-takers' ability and test difficulty. It then expresses this relationship as the probability of a certain response. Rasch IRT can help establish the model validity of the IELTS Listening test. Both Rasch IRT and structural equation modeling (SEM) provide statistical mechanisms for assessing how well the estimated model parameters fit the observed sample data (i.e., a model fitting data well makes good predictions about patterns in observed behaviours, whereas a model fitting poorly makes less accurate predictions). According to Reise and Widaman (1999), IRT is more advanced than SEM in predicting how well a model fits at the level of the individual (person-fit analysis). For SEM, parameter estimates at the level of the individual remain limited because a person's score on latent variables is not of any utility in SEM model testing and evaluation. This is because SEM mainly uses variance-covariance matrices.

In the current study, the IELTS Listening test data were Rasch IRT analysed (one parameter) for internal consistency, item difficulty, person ability and discrimination analysis. The Winsteps program was used. Rasch analysis not only reports on the internal consistency of a test, but also allows researchers to investigate the extent to which a particular test item is functioning to assess test-takers' ability, as well as the extent to which a particular test-taker is suitable for the given test. Fit statistics are produced in Rasch IRT. Misfitting items and test-takers can be identified through, for example, infit and outfit mean square statistics. In this study, severely misfitting test-takers and test questions were excluded from further analysis.

The IRT procedure in the present study can be summarised as follows:

- Once completed, the answer sheets were marked based on the answer keys. The marking scheme for the test was strictly followed. For example, if test-takers were asked to write an answer using a certain number of words (e.g., with the instruction NO MORE THAN TWO WORDS), they were penalised if they exceeded that number. The test was scored dichotomously (right = 1 or wrong = 0). Scoring was double-checked by a research assistant.

- Test-takers' score for each question was then entered into SPSS. Winsteps was used to impute the data from the SPSS file. It should be noted that test-takers' answers to the multiple-choice questions (A, B or C) were not used in Rasch IRT analysis. Only dichotomous scores were used.

Appendix 2 shows the Rasch IRT analysis outputs including the convergence table, map of item fit, item statistics, test-taker statistics, and distractor analysis (focusing on the discrimination analysis). The IRT analysis result of the IELTS Listening test indicated a reliability of 0.87, which was quite high, and reasonable. Table 9 presents the summary of case estimates. The person ability estimate mean of -0.01 suggests that the test was relatively difficult (further discussed in Figure 5). The mean was close to 0, indicating a well-matched test. The standard deviation of 1.10 person estimates indicates a good distribution of person ability. It should be noted that if the multiple-choice answers (e.g. questions 11-13, 31-36) had been analysed together with the dichotomous data (1 or 0), the reliability would have been higher.

Statistics	Value
Mean	-0.01
SD	1.10
SD (adjusted)	1.03
Reliability of estimate	0.87

Table 9: Summary of case estimates (N = 388)

Misfitting statistics for both test items and test-takers derived from the IRT analyses were used to determine whether some test questions or test-takers were misfitting. Misfitting is a statistical term used in Rasch IRT which is expressed as *mean square* or *t* statistics. These statistics enable researchers to investigate the coherence of a test-taker's responses as part of a set of responses from a larger group of test-takers. Misfitting test-takers are those whose abilities are not measured appropriately by this particular test. In other words, the direction of misfit is of the test to the test-taker, not the test-taker to the test. In this study, misfitting test-takers and questions were excluded from the data set. It was found that 10 test-takers were misfitting, so they were excluded from the data set.

Question 9 was identified as a misfitting test item and was subsequently removed from further statistical analyses. The item discrimination analysis indicated that the test items functioned well (*Point-Biserial* > 0.25 for the correct answers). Three questions (Questions 9 ($r = 0.05$), 13 ($r = 0.19$) and 32 ($r = 0.12$)) had the Point-Biserial statistic below 0.25.

Table 10 presents descriptive statistics of the four IELTS Listening test sections. Both raw scores and percentage scores are included. All variable skewness and kurtosis statistics were within ± 1.00 , which was suggestive of univariately normal distributions (skewness and kurtosis statistics were close to zero).

Section	No. of items	Minimum	Maximum	Mean	SD	Skewness	Kurtosis
1 (raw)	9	0.00	9.00	5.69	2.00	-0.42	-0.29
(%)		0.00	100.00	63.18	22.33	-0.42	-0.29
2 (raw)	10	0.00	10.00	4.57	2.28	0.14	-0.72
(%)		0.00	100.00	45.66	22.77	0.14	-0.72
3 (raw)	10	0.00	10.00	5.82	2.74	-0.27	-0.94
(%)		0.00	100.00	58.19	27.37	-0.27	-0.94
4 (raw)	10	0.00	10.00	3.22	2.06	0.78	-0.24
(%)		0.00	100.00	32.23	20.59	0.78	-0.24

Table 10: Descriptive statistics of the IELTS test performance variables (N = 376)

Table 11 presents the Cronbach's alpha coefficients for each IELTS Listening section.

IELTS Listening test	No. of items	Items used	Cronbach's alpha
Section 1	9	1, 2, 3, 4, 5, 6, 7, 8, 10	0.64
Section 2	10	11, 12, 13, 14, 15, 16, 17, 18, 19, 20	0.68
Section 3	10	21, 22, 23, 24, 25, 26, 27, 28, 29, 30	0.81
Section 4	10	31, 32, 33, 34, 35, 36, 37, 38, 39, 40	0.60
Total	39		0.87

Table 11: Internal consistency estimates (Cronbach's alpha) for the IELTS Listening test (N = 376)

Figure 5 presents the IRT item difficulty and student ability map. It shows a continuum of item difficulty and person ability. On the left are the units of measurement on the scale (called logits), ranging in this case from -3 to +4 (a 7-unit range). The average item difficulty is set at 0 as per convention. The ability of individual students is plotted on the scale (represented as '#' and '.'; each '#' represents 2 students and each '.' represents 1 student.). On the right are the item numbers for the test questions. The higher on the scale an item appears, the greater its level of difficulty. Similarly, the higher on the scale a test-taker appears, the greater the level of ability of that test-taker.

The item difficulty and person ability map in Figure 5 indicates a reasonably good match between the abilities of the test-takers and the test items. It shows a continuum of difficulty and ability. In Figure 5, item difficulty levels were identified and labelled for statistical analysis that examined the nature of test-takers' confidence and calibration in different test difficulty levels (discussed in Section 4: Findings and Section 5: Discussion).

Table 12 summarises the test questions at different difficulty levels, together with the IRT logit score spreads, the Cronbach's alpha coefficients of the test and confidence items. The IRT Logit score for each test question can be found in Appendix 2 (A2.3).

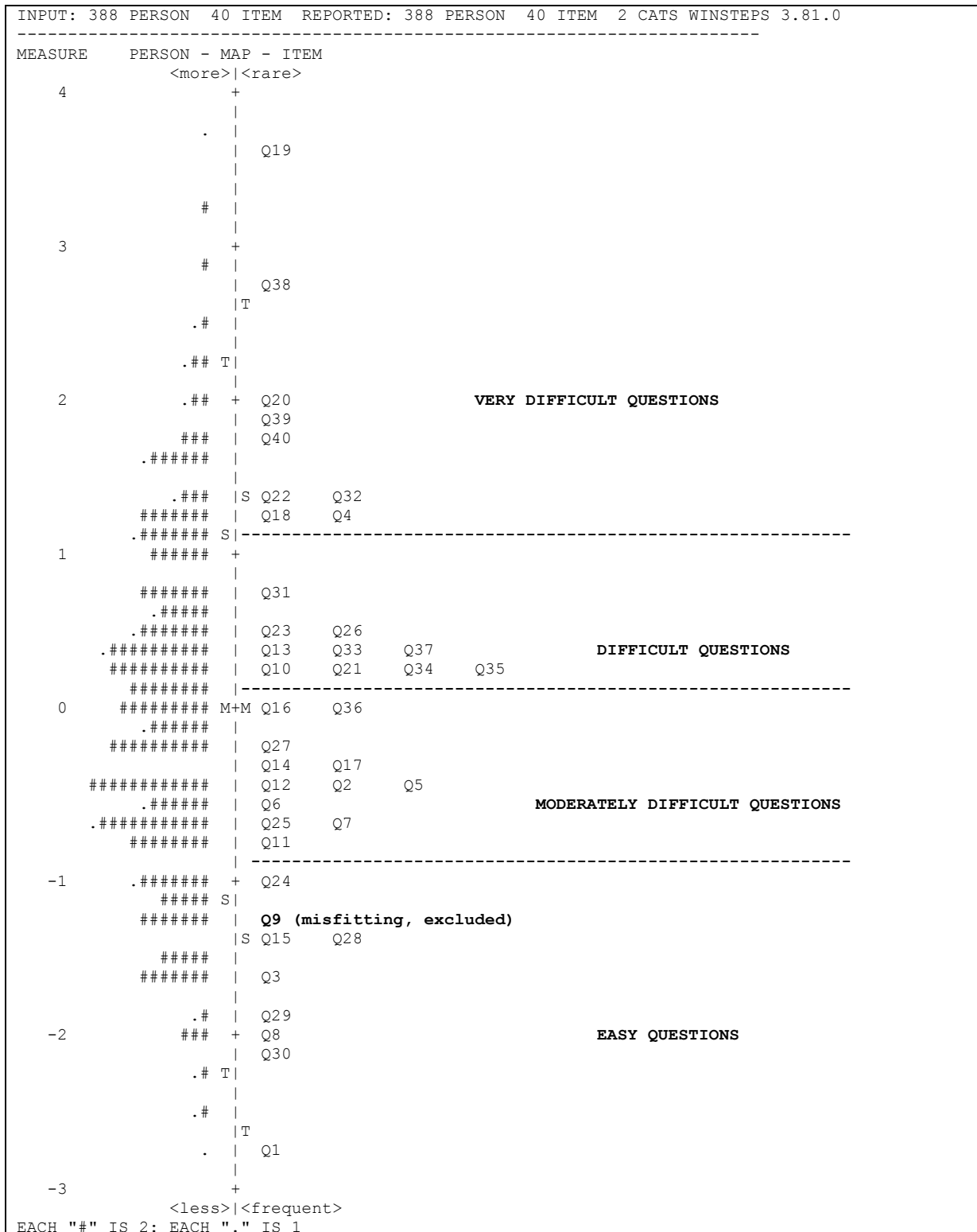


Figure 5: IRT item difficulty and person ability map (N = 388)

Difficulty level	No. of items	IRT Logit score spread	Items	Cronbach's alpha (test)	Cronbach's alpha (confidence)
Easy	7	-2.69 – -1.36	1, 3, 8, 15, 28, 29, 30	0.54	0.75
Moderately difficult	11	-1.02 – -0.31	2, 5, 6, 7, 11, 12, 14, 17, 24, 25, 27	0.73	0.86
Very difficult	12	-0.01 – 0.72	10, 13, 16, 21, 23, 26, 31, 33, 34, 35, 36, 37	0.66	0.88
Extremely difficult	9	1.28 – 3.64	4, 18, 19, 20, 22, 32, 38, 39, 40	0.66	0.84
Subtotal	39			0.87	0.95

Table 12: IELTS Listening question difficulties with Cronbach's alpha coefficients

3.8.1.3 Analysis of the single-case and relative-frequency questionnaire

Table 13 presents the descriptive statistics of the single-case confidence. It was found that some single-case appraisal confidence items had high values of skewness and kurtosis statistics, suggesting that the data were not normally distributed (i.e., univariately Kurtotic). However, in the case of appraisal confidence rating, this is not surprising because if a question is assessed to be easy for most test-takers, then those test-takers would report their confidence to be very high. This means that the scores would be largely distributed toward the upper end of the appraisal confidence continuum.

Item	Minimum	Maximum	Mean	SD	Skewness	Kurtosis
Q1CON	0.00	100.00	95.62	13.52	-4.01	18.32
Q2CON	0.00	100.00	82.22	27.87	-1.72	1.92
Q3CON	0.00	100.00	91.84	17.29	-2.96	9.73
Q4CON	0.00	100.00	81.29	25.24	-1.56	1.71
Q5CON	0.00	100.00	71.53	30.69	-0.97	-0.14
Q6CON	0.00	100.00	70.97	34.86	-1.00	-0.43
Q7CON	0.00	100.00	94.03	14.64	-3.57	14.60
Q8CON	0.00	100.00	91.91	16.97	-2.96	9.72
Q9CON	0.00	100.00	92.91	16.42	-3.15	11.65
Q10CON	0.00	100.00	65.94	39.12	-0.77	-1.07
Q11CON	0.00	100.00	65.47	31.07	-0.60	-0.82
Q12CON	0.00	100.00	66.10	30.63	-0.60	-0.74
Q13CON	0.00	100.00	57.77	29.82	-0.25	-0.89
Q14CON	0.00	100.00	62.46	35.07	-0.49	-1.14
Q15CON	0.00	100.00	65.03	35.26	-0.60	-1.05
Q16CON	0.00	100.00	55.89	32.46	-0.24	-1.12
Q17CON	0.00	100.00	52.34	32.64	-0.06	-1.17
Q18CON	0.00	100.00	34.27	38.26	0.56	-1.30
Q19CON	0.00	100.00	25.80	32.93	0.93	-0.56
Q20CON	0.00	100.00	36.94	40.73	0.46	-1.50
Q21CON	0.00	100.00	62.17	38.28	-0.48	-1.36
Q22CON	0.00	100.00	48.56	41.44	-0.01	-1.69
Q23CON	0.00	100.00	71.50	34.24	-0.99	-0.42
Q24CON	0.00	100.00	72.66	36.21	-1.10	-0.34
Q25CON	0.00	100.00	75.19	35.70	-1.26	0.04
Q26CON	0.00	100.00	58.25	40.78	-0.40	-1.52
Q27CON	0.00	100.00	61.41	41.35	-0.54	-1.44
Q28CON	0.00	100.00	74.99	37.04	-1.26	-0.08
Q29CON	0.00	100.00	80.85	31.81	-1.64	1.29
Q30CON	0.00	100.00	87.28	23.18	-2.24	4.54
Q31CON	0.00	100.00	53.47	32.30	-0.14	-1.16
Q32CON	0.00	100.00	51.32	31.07	-0.12	-1.07
Q33CON	0.00	100.00	46.93	29.44	0.03	-0.98
Q34CON	0.00	100.00	41.31	29.28	0.21	-0.90

Table 13: Distributions for single-case appraisal confidence of the 40 questions (N = 376) (continued over)

Item	Minimum	Maximum	Mean	SD	Skewness	Kurtosis
Q35CON	0.00	100.00	41.93	30.53	0.24	-1.00
Q36CON	0.00	100.00	46.15	32.93	0.17	-1.16
Q37CON	0.00	100.00	41.97	42.14	0.31	-1.63
Q38CON	0.00	100.00	29.32	36.96	0.85	-0.88
Q39CON	0.00	100.00	29.91	37.04	0.82	-0.91
Q40CON	0.00	100.00	49.12	40.32	-0.00	-1.63

Table 13: Distributions for single-case appraisal confidence of the 40 questions (N = 376) (continued)

It should be noted that the study examines the relationship between appraisal confidence and performance for each of the four listening test sections as a whole, rather than at the level of a single test question. It should, therefore, be noted that appraisal confidence data were continuous at a test section level, since appraisal confidence for each questions were aggregated, based on a series of test questions.

Table 14 presents the descriptive statistics of the single-case appraisal confidence and relative-frequency appraisal confidence for each test section.

Item	No. of items	Minimum	Maximum	Mean	SD	Skewness	Kurtosis
SC-con1	9	5.56	100.00	82.82	15.08	-1.19	1.72
SC-con2	10	0.00	100.00	52.21	25.60	-0.20	-0.85
SC-con3	10	0.00	100.00	69.29	26.42	-0.73	-0.49
SC-con4	10	0.00	100.00	43.14	25.32	-0.18	-0.78
RF-con1	1	0.00	100.00	79.15	18.77	-1.24	1.23
RF-con2	1	0.00	100.00	50.37	26.40	-0.17	-0.82
RF-con3	1	0.00	100.00	69.65	27.38	-0.86	-0.30
RF-con4	1	0.00	100.00	41.62	25.85	-0.26	-0.76

SC-con = Single-case confidence RF-con = Relative-frequency confidence 1-4 = Sections 1 to 4

Table 14: Distributions of single-case appraisal confidence and relative-frequency appraisal confidence across the four IELTS sections (N = 376)

Although single-case appraisal confidence ratings are subjective in nature since test-takers use their own criteria to appraise their performance, it is important to know whether the appraisal confidence scales measured their appraisal confidence consistently. Table 15 presents the Cronbach's Alpha coefficients for the single-case appraisal confidence. The Cronbach's Alpha coefficients were very high. The reliability of the single-case appraisal confidence in Section 1 was the lowest (0.75).

Single-case confidence	No. of items	Items used	Cronbach's alpha
Section 1	9	1, 2, 3, 4, 5, 6, 7, 8, 10	0.75
Section 2	10	11, 12, 13, 14, 15, 16, 17, 18, 19, 20	0.92
Section 3	10	21, 22, 23, 24, 25, 26, 27, 28, 29, 30	0.90
Section 4	10	31, 32, 33, 34, 35, 36, 37, 38, 39, 40	0.91
Total	39		0.95

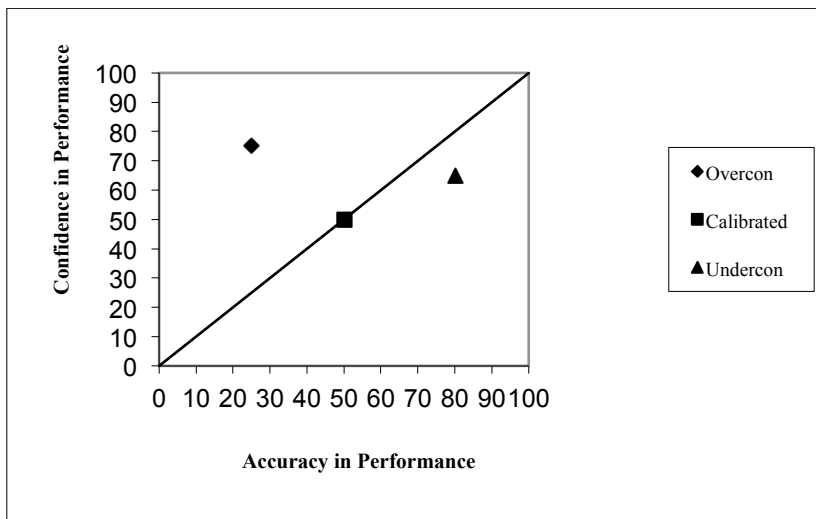
Table 15: Internal consistency estimates (Cronbach's alpha) for the single-case appraisal confidence (N = 376)

3.8.2 Data analysis to address the research questions

This section presents the data analysis that was conducted to answer the research questions in the study. This analysis includes: (1) analysis of test-takers' appraisal calibration; (2) Pearson-Product Moment Correlations; (3) *t*-tests; (4) analysis of variance (ANOVA); (5) confirmatory factor analysis (CFA); and (6) structural equation modeling (SEM). The IBM Statistical Package for Social Sciences (SPSS) version 22 for PC was used for all standard statistics. Microsoft Excel was used to produce calibration graphs. The EQS program version 6.2 (Bentler, 1985-2016) was used for confirmatory factor analysis and structural equation modeling.

3.8.2.1 Analysis of appraisal calibration

Figure 6 presents a calibration diagram that illustrates appraisal confidence ratings as a measure of the accuracy of a test-taker's performance appraisals. If a person's appraisal confidence is on the 45° line (the unity line), the person is said to be calibrated (appraisal confidence and performance match perfectly). If a person's appraisal confidence is above this line, the person is overconfident, and if it is below the line, the person is underconfident. Both overconfidence and underconfidence are known as miscalibration.



Overcon = Overconfident Undercon = Underconfident

Figure 6: Calibration of performance appraisal diagram

3.8.2.2 Appraisal calibration score

There are various ways to measure test-takers' calibration (see Schraw, Kuch & Gutierrez 2013). The simplest method is a subtraction method, which is a measure of absolute accuracy. It measures the degree to which each appraisal confidence is congruent with a given test question performance. That is, scores closer to zero indicate a high accuracy rate of appraisal calibration. Another common measure of appraisal calibration is correlational analysis between appraisal confidence and its associated test score. A correlational analysis is considered a measure of relative accuracy, which provides information about appraisal confidence relative to test performance. This measure cannot tell researchers whether and the extent to which students are overconfident or underconfident in their performance appraisals. Hence, both methods for examining students' appraisal calibration (i.e., an appraisal calibration score and a correlation coefficient) are normally used.

Subtraction method: Appraisal calibration can be computed by subtracting the appraisal confidence rating from the actual performance in a percentage term.

$$C = c - p$$

where C = Calibration; c = confidence expressed as a percentage; p = performance or test score expressed as a percentage.

It can be argued that, since an assessment of test-takers' performance on the basis of only one test question or one task type cannot capture test-takers' ability, language testers need to use a number of questions and a variety of tasks to infer test-takers' ability. Therefore, to avoid drawing erroneous conclusions about test-takers' performance, a test typically has a number of questions and a variety of tasks. This principle also applies to the analysis of test-takers' appraisal calibration. That is, appraisal confidence in test performance also needs to be collected over a series of test questions. Both test scores and appraisal confidence scales are separately aggregated to form an average test score and an average appraisal confidence score. Test scores need to be converted into percentages, so that test-takers' appraisal calibration scores can be computed (see Lin & Zabrucky, 1998).

As discussed earlier, a test-taker is considered calibrated when C equals zero, which suggests no discrepancy between confidence and performance. When C is larger than zero (+), he/she is overconfident and when C is negative (-), he/she is underconfident. Poor appraisal calibration is detected when the test-taker's calibration score is different from zero. An appraisal calibration score higher than 10% is non-negligible (Klietman & Stankov 2001; Stankov & Lee 2008).

In the present study, the following cut-off criteria are used to judge test-takers' calibration (see Figure 6): realistic (within $\pm 5\%$); just overconfident ($6 < 10\%$); generally overconfident (11–24%); extremely overconfident ($> 25\%$); just underconfident ($-6 < -10\%$); generally underconfident (-11– -24%); and extremely underconfident ($> -25\%$).

- **Correlation method:** Pearson-Product-Moment correlational analysis of students' appraisal confidence in performance and actual test performance is another method for estimating test-takers' relative calibration (Nelson, 1984). A high correlation coefficient indicates that students are calibrated, whereas a low correlation coefficient indicates that they are miscalibrated. When data are normally distributed, Pearson-Product Moment correlations can be used.

Correlation is a statistical analysis method that is often used by researchers to examine whether a linear relationship between two variables exists. A relationship (r) has a magnitude between 0 (no relationship at all) and 1 (perfect relationship).

A positive sign indicates that two variables increase or decrease in tandem, whereas a negative sign suggests that as one variable increases, the other decreases, and vice versa. It is important to note that the value of r needs to be multiplied with itself (i.e., squared) in order to see how much two variables of interest overlap. For example, if r is 0.50, it means that two variables share 25% of their content (i.e., 0.50×0.50). This is known as the shared variance (R^2) between two variables and can be treated as an effect size.

Apart from Pearson-Product-Moment correlations, the present study also examines the correlation between appraisal confidence and performance through SEM analysis (discussed below). Unlike SEM, a standard correlation analysis treats observed data as free of non-random errors of measurement (i.e., systematic errors implicit to the measurements). As a result, correlation can over- or underestimate the parameter of the relationship. Although researchers may have examined reliability estimates of their measurements and found them to be acceptable, the reliability estimates are not integrated in the raw scores before the correlational analysis is performed.

3.8.2.3 *T*-tests

Two types of *t*-tests are used in this study. A paired-samples *t*-test is used to examine whether two mean scores from the same group of participants differ significantly. For example, in this study, test-takers' listening scores are compared with single-case appraisal confidence scores. An independent-samples *t*-test is used to determine whether the mean scores between two groups of test-takers are significantly different. For example, test and appraisal confidence scores of male and female test-takers can be compared.

3.8.2.4 Analysis of variance (ANOVA)

An analysis of variance (ANOVA) is a parametric test for determining whether there is a statistically significant difference between the scores obtained by two or more groups. It can be used to infer the effects of one independent variable on other dependent variables. When ANOVA is used, the possibility of a Type I error is lower than when a *t*-test is used. Unlike a *t*-test, ANOVA separates the variance that is attributable to *between-group* differences from the variance that is attributable to *within-group* differences.

Thus, when there are more than two groups to compare, a one-way ANOVA is more robust in making a statistical inference as it performs a single analysis with a p -value of 0.05. The statistical assumptions of ANOVA include sample normality and homogeneity of variance.

3.8.2.5 Structural equation modeling (SEM)

SEM is the term used to describe multivariate statistical models for evaluating the validity of a theory or hypothesis through empirical data. In language testing and assessment research, it can help researchers elegantly and rigorously validate and/or develop a theory using empirical data. It provides researchers with a comprehensive method for testing theories and examining data fit. SEM can be applied for research purposes, such as to test substantive theory (hypothesis testing), organise concepts about data analysis into scientific models, include flexible provisions for models with latent variables, and determine direct or indirect (mediation) independence of one variable from another (see Ockey, 2014; Winke, 2014). Table 16 provides a summary of common symbols used in an SEM model.



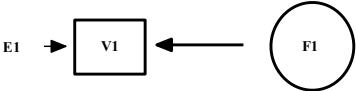
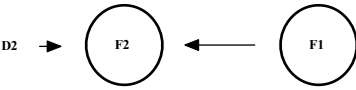


Symbols	Explanation
	<ul style="list-style-type: none"> Latent variable or factor (<i>Circles or ellipses</i>)
	<ul style="list-style-type: none"> Observed variable or indicator (<i>Boxes</i>)
	<ul style="list-style-type: none"> A causal relationship from a latent variable to an effect or a dependent variable (<i>Single-headed arrows or unidirectional arrow</i>) E1: Measurement error with the observed variable; used to represent error on an observed variable. Some SEM programs may use 'e' instead of E.
	<ul style="list-style-type: none"> An example of a basic structural model A single-headed arrow indicates a path coefficient for regression of one independent latent variable (exogenous latent variable) onto a dependent latent variable (<i>endogenous</i> latent variable) D2: Residual error (disturbance/error) in prediction of the dependent latent factor. D2 is associated with the prediction error of Factor 2 by Factor 1.
	<ul style="list-style-type: none"> A linear or non-directional relationship between two observed variables (<i>Double-headed arrows</i>)
	<ul style="list-style-type: none"> A linear relationship between two latent factors (<i>Double-headed arrows</i>)

Table 16: Common symbols used in SEM

The current study follows seven key steps that SEM researchers go through in the development of substantive SEM applications (see Bentler 2006; Bollen 1989; Byrne 2006; Kline 2011; Schumacker & Lomax 2010). These steps include (1) model specification, (2) model identification, (3) data collection and preparation, (4) model estimation, (5) model fit assessment, (6) model re-specification and modification, and (7) model interpretation and report. These steps are essential to guaranteeing the *statistical validity* (i.e., the accuracy) of a hypothesised SEM model. The maximum likelihood (ML) method was used in the current study. A SEM model in the current study is accompanied by standardised parameter estimates. Table 17 summarises and explains the key goodness-of-fit criteria for assessing SEM model fit.

GOF Criteria	Acceptable level	Interpretation
Chi-square (χ^2)	Table χ^2 value	Compare obtained χ^2 value with table value given <i>df</i> (degree of freedom). $\chi^2/df \leq 3$
Probability value of χ^2 (p)	$p > 0.001$	$p < 0.001$ indicates that the event occurs less than one time in a thousand. A <i>nonsignificant</i> χ^2 test implies that the data fit the model (unlike other standard statistics).
Root-mean-squared residual (RMR)	Indicates the closeness of Σ to S matrix. 0 indicates perfect model fit.	Researchers define the value level. A well-fitting model has a value of 0.05 or lower.
Root-mean-squared error of approximation (RMSEA)	< 0.05 , but not > 0.10	A value lower than 0.05 indicates a very good model fit. 90% confidence intervals should be used.
Akaike Information Criterion (AIC)	Small values indicate well fitting, parsimonious models.	This compares the value of the hypothesised model AIC with that of the independence AIC (i.e. null model).
Norm fit index (NFI)	0 (no fit) to 1 (perfect fit)	Value close to 0.95 reflects an excellent fit.
Non-norm fit index (NNFI)	0 (no fit) to 1 (perfect fit)	Value close to 0.95 reflects an excellent fit.
Comparative Fit Index (CFI) and Bollen (IFI)	0 (no fit) to 1 (perfect fit)	Value close to 0.95 reflects an excellent fit.
LISREL Goodness-of-fit index (GFI)	0 (no fit) to 1 (perfect fit)	Value close to 0.95 reflects an excellent good fit.

Table 17: Summary of the key GOF criteria and acceptable fit levels and interpretations

There are two components of an SEM model: measurement models and structural models. An explanation is useful for readers who are not familiar with SEM. Figure 7 illustrates an example of a measurement model.

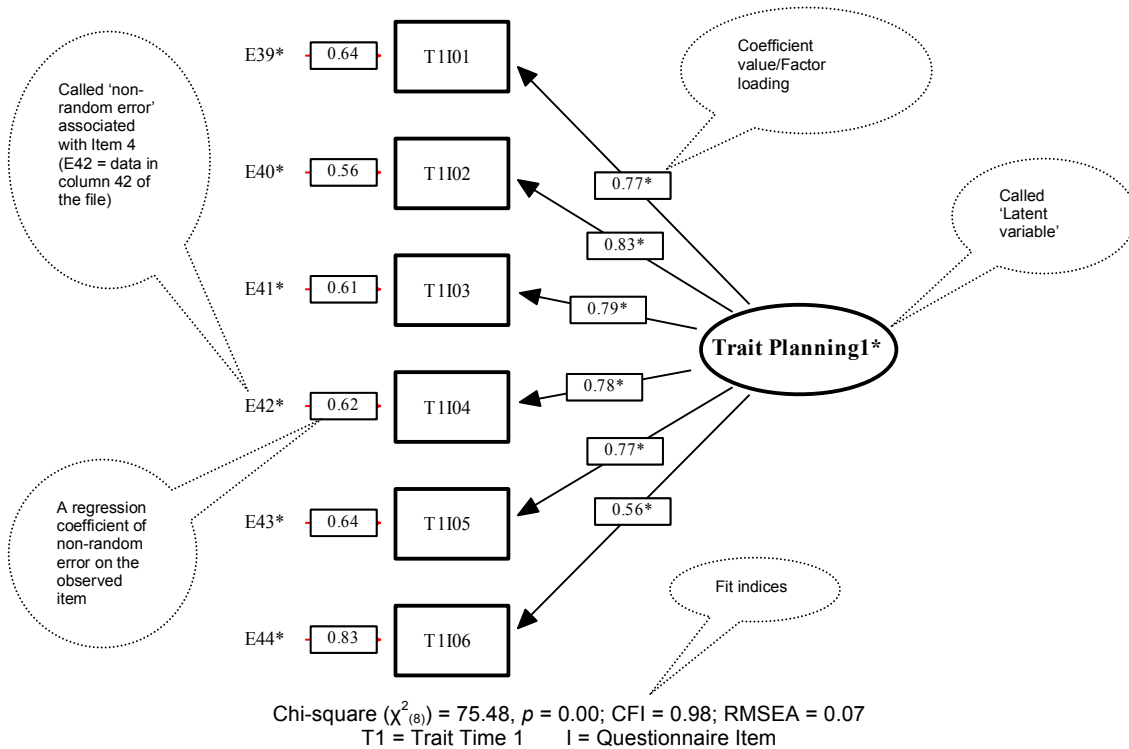


Figure 7: A hypothesised one factor model of trait planning strategy use Time 1 (Phakiti, 2007b, N = 651)

Measurement models for SEM are typically generated by means of confirmatory factor analysis (CFA). A CFA model represents a *measurement model* (representing a latent variable) in a SEM approach. The relationship between a factor and observed measures is defined in terms of *regression weights* that link factors to measures. A regression weight is commonly referred to as a *factor loading* which has a value between 0 and 1. CFA is used when researchers can draw upon theory and aim to confirm a hypothesis that a link between the observed and latent variables (i.e., a higher-order factor) exists. CFA can be used to test a connection with other CFAs in SEM analysis – known as a structural model.

Structural models are used to determine the relationships among latent variables. A hypothesised relationship (direct or indirect) should be informed by the theory or hypothesis to be tested, although some researchers may wish to explore possible relationships as suggested by the data set. Technically, in SEM, a latent variable that is used to predict another latent variable is called an ‘exogenous’ (independent) variable. The latent variable being predicted is called the ‘endogenous’ (dependent) variable. The error variances associated with observed variables are labelled as E and error variances associated with endogenous variables are referred to as disturbance (D) (see Table 16). A non-random error in SEM (e.g., E in Table 16) is computed as follows: $E = \sqrt{(1-r^2)}$, where $\sqrt{}$ = square root, r = factor loading. For example, in Figure 7 for the observed variable T1101, the non-random error associated with the factor loading of 0.77 can be computed as: $E = \sqrt{(1-0.77^2)} = \sqrt{(1-0.5929)} = \sqrt{0.4071} = 0.64$, so $r^2 + E^2 = 1$ (i.e., $0.5929 + 0.4071$). The calculation of D is the same as that for E, but can be complex if the number of independent latent variables affecting a dependent latent variable is high. In SEM, parameter estimates (i.e., relationships among latent variables) are of primary interest to researchers.

Unlike standard statistics, which assume data to be error-free, SEM separates the effects of error variances associated with observed variables (E) or endogenous latent variables (D) during parameter estimates. Parameter estimates in SEM are arguably more accurate than those generated by other standard statistics which do not take error variances into account in parameter estimates (this will be examined further in Sections 4 and 5). In SEM, the regression coefficient of an exogenous latent variable (independent factor) on an endogenous latent variable (dependent factor) is represented by *gamma* (γ), whereas that of an endogenous latent variable on another endogenous variable is represented by *beta* (β). Figure 8 summarises the SEM procedures in the present study.

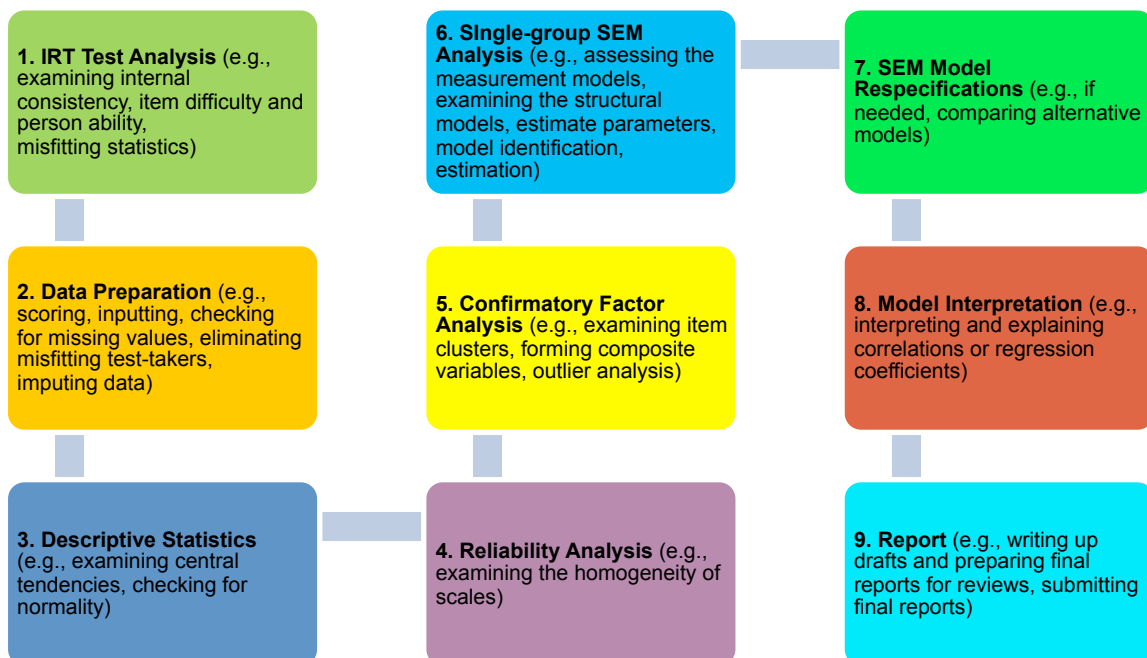


Figure 8: A flow chart of SEM used in the present study

4 FINDINGS

The following sections present the results of the four research questions.

4.1 What is the nature of test-takers' appraisal confidence and appraisal calibration in an IELTS Listening test?

The appraisal components of strategic competence (i.e., monitoring and evaluation) during IELTS Listening test-taking are examined in Research Question 1. These components were measured by asking test-takers to rate their appraisal confidence in the correctness of their test answers (single-case appraisal confidence and relative-frequency appraisal confidence).

4.1.1 The nature of test-takers' appraisal confidence and IELTS Listening test performance

Table 18 presents the descriptive statistics of test-takers' IELTS Listening test scores, single-case appraisal confidence scores and relative-frequency appraisal confidence scores. In Table 18, it can be seen that test-takers tended to perform better in Sections 1 (63%, SD = 22.33) and 3 (58%, SD = 27.37), which involved conversational and transactional test tasks, than in Sections 2 (46%, SD = 23) and 4 (32%, SD = 21), which involved monologue test tasks. Paired-sample *t*-tests were performed to determine whether test-takers' test performance differed significantly across the four test sections.

Table 19 presents the *t*-test results.

Section	No. of items	Mean	SD
IELTS1	9	63.18	22.33
SCCON in IELTS1	9	82.82	15.08
RFCON in IELTS1	1	79.15	18.77
IELTS2	10	45.66	22.77
SCCON in IELTS2	10	52.21	25.60
RFCON in IELTS2	1	50.37	26.40
IELTS3	10	58.19	27.37
SCCON in IELTS3	10	69.29	26.42
RFCON in IELTS3	1	69.65	27.38
IELTS4	10	32.23	20.59
SCCON in IELTS4	10	43.14	25.32
RFCON in IELTS4	1	41.62	25.85
Overall IELTS	39	49.82	18.66
Overall SCCON	39	61.86	19.21
Overall RFCON	4	60.20	19.27

SCCON = Single-case Appraisal Confidence RFCON = Relative-frequency Appraisal Confidence
1-4 = Test Section

Table 18: Descriptive statistics of the single-case and relative-frequency appraisal confidence and IELTS Listening test performance variables (N = 376)

		Paired Differences					<i>t</i>	Sig. (2-tailed)	Cohen's <i>d</i>
		Mean	SD	SEM	95% Confidence Interval of the Difference				
					Lower	Upper			
Pair 1	IELTS1 - IELTS2	17.51**	22.77	1.17	15.21	19.82	14.92	0.00	0.77
Pair 2	IELTS1 - IELTS3	4.99**	25.36	1.31	2.42	7.56	3.81	0.00	0.20
Pair 3	IELTS1 - IELTS4	30.95**	23.30	1.20	28.58	33.31	25.75	0.00	1.33
Pair 4	IELTS2 - IELTS3	-12.53**	22.47	1.19	-14.80	-10.25	-10.81	0.00	-0.46
Pair 5	IELTS2 - IELTS4	13.44**	21.27	1.10	11.27	15.59	12.24	0.00	0.63
Pair 6	IELTS3 - IELTS4	25.96**	23.02	1.19	23.62	28.29	21.87	0.00	1.17

** = $p < 0.01$ (2-tailed) 1-4 = Number of Test Section

Table 19: The paired-sample *t*-test results between single-case and relative-frequency appraisal confidence (N= 376)

There were statistically significant differences in the IELTS Listening test scores across the four sections, suggesting that Section 1 was the easiest and Section 4 was the most difficult. The effect sizes (Cohen's *d*) ranges from 0.20 (Pair 2; small effect size) to 1.33 (Pair 3; large effect size). The value of Cohen's *d* indicates the percentage of non-overlap of the data associated with two groups of participants. A Cohen's *d* value of 0 suggests that the score distributions between the two groups entirely overlap (i.e., there is no difference in the distributions). A Cohen's *d* value of 0.8 is located in the 79th percentile, indicating a non-overlap of 47.4% in the two compared distributions. A Cohen's *d* of 1.0 indicates that the score distributions between the two groups exhibit a 1 standard deviation difference. The effect size of 0.20 suggests that test performances in Sections 1 and 3 were significantly different to a small degree. On the other hand, the effect size of 1.33 indicates that differences in test scores between Sections 1 and 4 were large. The majority of the pairs exhibited a medium to large effect size (0.63-1.33; *D*-values of 0.2, 0.5, and 0.8 indicate small, medium, and large effect sizes respectively; Cohen 1988).

In Table 18, single-case and relative-frequency appraisal confidence judgments appear to be larger than their associated test scores. However, they did not seem to differ largely from each other. In order to find out whether test-takers' single-case appraisal calibration and relative-frequency appraisal calibration scores significantly differed from each other, paired-sample *t*-tests were performed. Table 20 presents the *t*-test results. It was found that, except for Pair 3, the two types of appraisal confidence were significantly different ($p < 0.05$), suggesting that single-case appraisal confidence judgments were larger than relative-frequency appraisal confidence judgments. However, the effect sizes of the significant differences were relatively small ($-0.03 < d < 0.31$).

		Paired Differences					<i>t</i>	Sig. (2-tailed)	Cohen's <i>d</i>
		Mean	SD	SEM	95% Confidence Interval of the Difference				
					Lower	Upper			
Pair 1	SCCON1 – RFCON1	3.67**	12.38	0.64	2.41	4.92	5.75	.000	0.31
Pair 2	SCCON2 – RFCON2	1.84**	12.12	0.63	0.60	3.06	2.93	.004	0.15
Pair 3	SCCON3 – RFCON3	-0.37	12.24	0.63	-1.61	0.87	-0.58	.560	-0.03
Pair 4	SCCON4 – RFCON4	1.52*	14.78	0.76	0.02	3.02	2.00	.047	0.10

SCCON = Single-case Appraisal Confidence RFCON = Relative-frequency Appraisal Confidence
 1-4 = Test Section * = $p < 0.05$ ** = $p < 0.01$

Table 20: The paired-sample *t*-test results between single-case and relative-frequency confidence (N= 376)

4.1.2 Test-takers' appraisal calibration scores

Table 21 presents test-takers' appraisal calibration scores across the four sections. Their appraisal calibration scores were calculated for both single-case and relative-frequency appraisal confidence judgments.

Section	Mean	SD
CALSC1	19.64	18.88
CALRF1	15.97	19.82
CALSC2	6.54	21.49
CALRF2	4.71	23.58
CALSC3	11.09	17.44
CALRF3	11.46	18.89
CALSC4	10.01	22.50
CALRF4	9.39	24.19
Overall CALSC	12.05	13.96
Overall CALRF	10.38	14.68

CALSC = Appraisal Calibration (Single-case) CALRF= Appraisal Calibration (Relative-frequency)
 1-4 = Number of Test Section

Table 21: Test-takers' calibration scores in the IELTS Listening test (N = 376)

According to Table 21, test-takers were found to be generally overconfident. Their appraisal calibration scores ranged from 6.5% (Section 2) to 19.6% (Section 1). Based on Table 21, test-takers had the best appraisal calibration scores in Section 2 because they were closer to zero. On average, test-takers were approximately 12% overconfident in their test performance. Figure 9 is an appraisal calibration diagram based on single-case appraisal confidence for the entire test. It reveals that a large majority of the test-takers were overconfident. The number of test-takers who did not perform well in the IELTS Listening test (i.e., those who scored between 10% and 50%) had the highest level of appraisal confidence in their success.

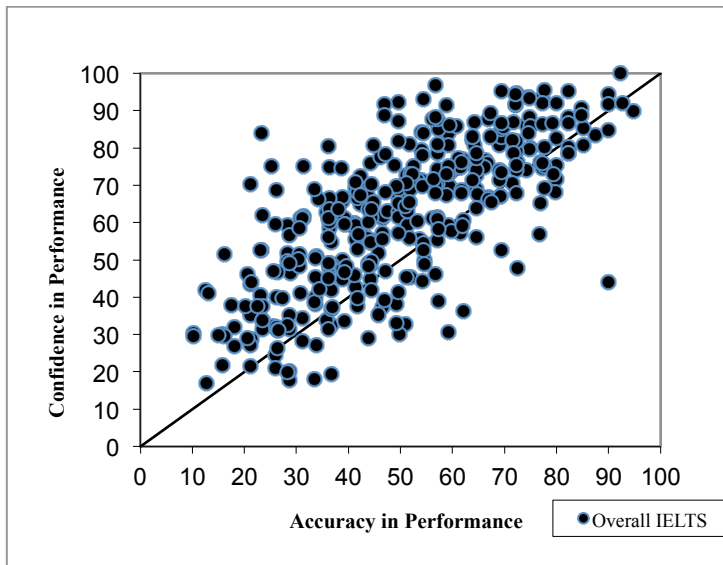


Figure 9: Test-takers' appraisal calibration diagram (single-case appraisal confidence) of the overall test (k = 39; N = 376)

Figures 10 and 11 present test-takers' appraisal calibration diagrams based on single-case appraisal confidence and relative-frequency appraisal confidence for each test section.

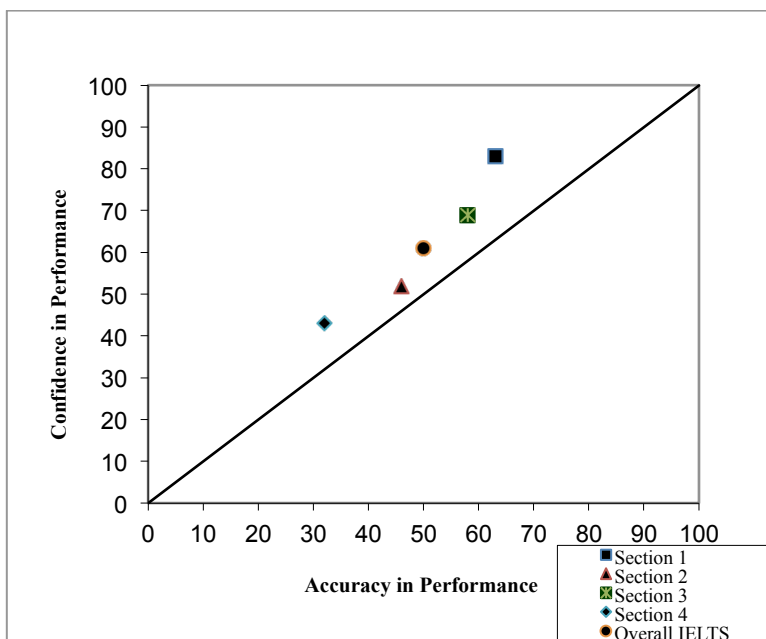


Figure 10: Test-takers' appraisal calibration diagram (single-case appraisal confidence) (N = 376)

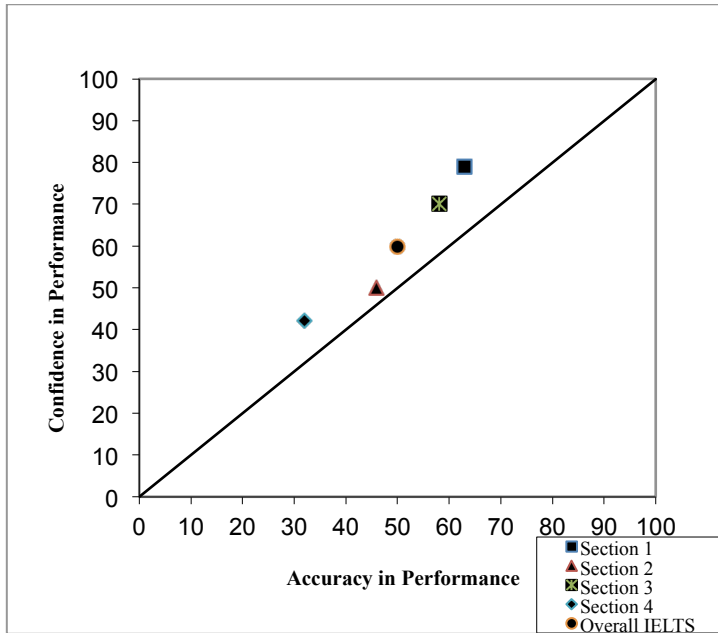


Figure 11: Test-takers' appraisal calibration diagram (relative-frequency appraisal confidence) (N = 376)

Paired-sample *t*-tests were performed to determine whether there was a statistically significant difference between single-case/relative-frequency appraisal confidence and IELTS Listening test performance in each test section. Table 22 presents the paired-sample *t*-test results. It was found that test-takers' appraisal confidence (single-case and relative-frequency) was significantly greater than their test performance in all the IELTS Listening sections ($p < 0.05$). The effect sizes (Cohen's *d*) ranged from 0.20 (Pair 8, small effect size) to 1.11 (Pair 1, large effect size). The majority of the pairs exhibited a medium to large effect size (*D*-values of 0.2, 0.5, and 0.8 indicate small, medium, and large effect sizes respectively; Cohen 1988). The paired-samples *t*-test results therefore suggest a significant mismatch between appraisal confidence and performance.

		Paired Differences					<i>t</i>	Sig. (2-tailed)	Cohen's <i>d</i>
		Mean	SD	SEM	95% Confidence Interval				
					Lower	Upper			
Pair 1	SCCON1 – IELTS1	19.64**	18.88	0.97	17.72	21.55	20.17	0.00	1.11
Pair 2	SCCON 2 – IELTS2	6.54**	21.49	1.11	4.36	8.72	5.90	0.00	0.31
Pair 3	SCCON3 – IELTS3	11.09**	17.45	0.90	9.33	12.86	12.33	0.00	0.64
Pair 4	SCCON4 – IELTS4	10.91**	22.50	1.16	8.63	13.19	9.40	0.00	0.49
Pair 5	Overall SCCON – Overall IELTS	12.05**	13.96	0.72	10.63	13.46	16.73	0.00	0.86
Pair 6	Overall RFCON – Overall IELTS	10.38**	14.68	0.76	8.89	11.87	13.71	0.00	0.71
Pair 7	RFCON1– IELTS1	15.97**	19.82	1.02	13.96	17.98	15.62	0.00	0.82
Pair 8	RFCON2 – IELTS2	4.71**	23.58	1.22	2.32	7.10	3.87	0.00	0.20
Pair 9	RFCON3 – IELTS3	11.46**	18.89	0.97	9.55	13.38	11.77	0.00	0.61
Pair 10	RFCON4 – IELTS4	9.39**	24.19	1.25	6.94	11.84	7.53	0.00	0.40

SCCON = Single-case Appraisal Confidence RF = Relative-frequency Appraisal Confidence
 ** $p < 0.01$ 1-4 = Number of Test Section

Table 22: The paired-sample *t*-test results (N= 376)

In summary, test-takers were not found to be well-calibrated in their performance appraisals of IELTS Listening test performance. They tended to be generally overconfident in their test performance (up to nearly 20%). They were found to approach good calibration in Section 2 of the IELTS Listening test.

4.1.3 Correlations between appraisal confidence and performance

Although the above findings indicate that test-takers were miscalibrated and tended to be overconfident in their performance across sections, there were some indications of the relationships between appraisal confidence judgments and IELTS Listening performance. Pearson-Product-Moment correlations were performed to investigate whether the relationships statistically existed and, if so, how much they were related. The assumptions for Pearson-Product-Moment correlations were met. Table 23 presents Pearson-Product-Moment correlations between test-takers' appraisal confidence and their IELTS Listening test performance.

IELTS Listening	SCCON	RFCON
Section 1	0.55** ($R^2 = 0.30$)	0.55** ($R^2 = 0.30$)
Section 2	0.61** ($R^2 = 0.37$)	0.55** ($R^2 = 0.30$)
Section 3	0.79** ($R^2 = 0.62$)	0.76** ($R^2 = 0.58$)
Section 4	0.54** ($R^2 = 0.29$)	0.48** ($R^2 = 0.23$)
Overall	0.73** ($R^2 = 0.53$)	0.70** ($R^2 = 0.49$)

SCCON = Single-case Appraisal Confidence RF = Relative-frequency Appraisal Confidence
 ** $p < 0.01$

Table 23: Pearson-Product-Moment correlations between appraisal confidence and IELTS Listening performance (N = 376)

According to Table 23, there were statistically significant correlations among single-case/relative-frequency appraisal confidence and IELTS Listening test performance ($p < 0.01$). The correlation coefficients for the single-case appraisal confidence ranged from 0.54 ($R^2 = 0.29$) for Section 4 to 0.79 ($R^2 = 0.62$) for Section 3. The coefficient of 0.54 indicates that test-takers' single-case appraisal confidence shared 29% of their test performance variance, whereas the coefficient of 0.79 suggests that test-takers' single-case appraisal confidence could predict about 62% of their test performance. The latter coefficient indicates that test-takers' performance appraisals were quite strong.

On the basis of the test mean score for Section 3 (58%, see Table 18), it might be said that the test tasks had a moderate difficulty level, which might facilitate test-takers' ability to more realistically appraise their test performance. On the other hand, since Section 4 was the most difficult section of all (with a mean score of 32%), test-takers might have experienced some difficulty in appraising their performance success. Being unaware that this test section was the most difficult section, they might have failed to adjust their mental processes to cope with the demands of the tasks. This presumption seems to be reflected in the appraisal calibration diagram for Section 4 (see Figure 12), in which most test-takers consistently exhibited a tendency to be highly overconfident in their performance.

It was observed that, as reported in Table 22, test-takers were found to have a better appraisal calibration score (based on single-case appraisal confidence) in Section 2 than in Section 3 (i.e., 6.5% versus 11%). However, the Pearson-Product-Moment correlation coefficient was found to be larger for Section 3 than for Section 2 (i.e. 0.79 versus 0.61). The two measures of appraisal calibration (subtraction formula and correlation) appear to produce contradictory results. That is, on average, test-takers were nearly calibrated in Section 2, but the linear relationship between appraisal confidence and performance was stronger in Section 3. It is, therefore, important to discuss what each measure of appraisal calibration can tell us.

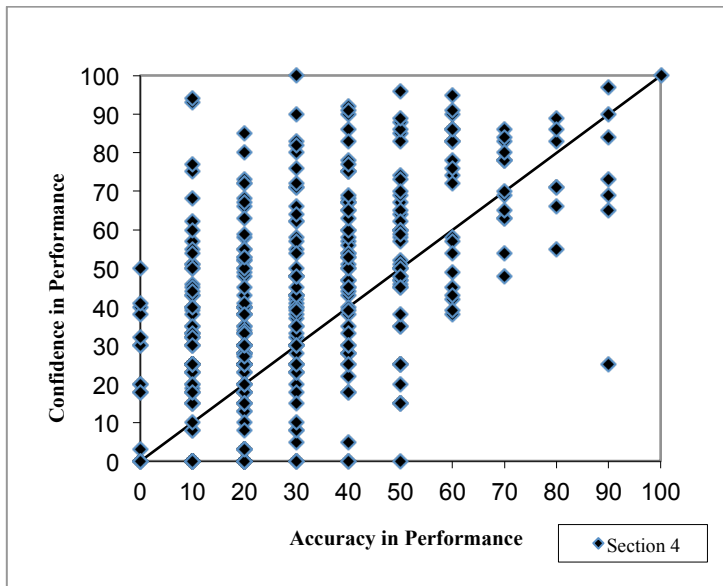


Figure 12: Test-takers' appraisal calibration diagram (single-case appraisal confidence) of Section 4 (K = 10; N = 376)

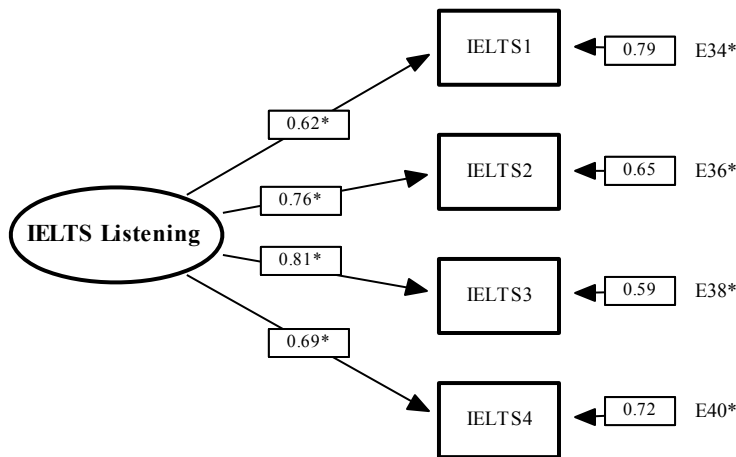
When test-takers are 25% over- or under-confident, their appraisal confidence deviates largely from their performance. Being overconfident or underconfident at this 25% or above level might affect test-takers' cognitive processes in test task engagement. That is, when they do not know that they are *not* performing a test task well, they cannot use strategies or problem-solving skills to address the given test task. On the other hand, if they do not know that they have already performed well, they may spend too much time attempting to complete the task over and over again or they might feel dissatisfied with their performance. It should be noted that the appraisal calibration scores above were calculated through the use of mean scores of all test-takers and errors of the mean scores are not considered in the calculation.

Unlike the appraisal calibration scores, the correlational analysis used all individual scores to estimate a relationship. A correlation coefficient estimates the shared variance between two variables in percentage terms. It can indicate whether a relationship is positive or negative. On the contrary, the positive or negative value of appraisal calibration scores indicates the kind of bias present (below or above 0). Pearson-Product-Moment correlation, like the calibration scores, treats all data as error-free, so the parameter estimate of the relationship may not be fully accurate. To have a better understanding of test-takers' appraisal calibration, different approaches to data analysis (e.g., comparing calibration at ability levels or question difficulty levels) should be employed.

The findings based on the Pearson-Product-Moment correlations in Table 23 suggest high variability of test-takers' ability to self-evaluate and estimate their test performance. While the effect sizes (i.e., R^2) are considered high for all sections according to Cohen's (1988) effect size interpretation (r values of 0.10, 0.30, and 0.50 are considered small, medium, and large effect sizes, respectively), much of the unexplained variance from both types of confidence judgments was relatively large. That is, the Pearson-Product-Moment findings indicate that test-takers could not accurately estimate their performance success. The lack of ability to appraise their performance implies that their strategic competence was inefficient and might not function properly to help them complete the given test tasks.

4.1.4 Model of IELTS Listening test performance

A CFA was performed to examine the latent factor of IELTS Listening test performance. The standard fit indices suggested very good model fit. Figure 13 presents a CFA model of IELTS Listening test performance, which is in the standardised solution. All observed variables (Vs), latent variables (Fs) and non-random errors (Es) are rescaled to have a variance of 1.0. The values in the standardised solution are the same as correlation coefficients.



Chi-square ($\chi^2_{(1)}$) = 1.19 $p = 0.28$ CFI = 1.00 RMSEA = 0.02 (90% CI = 0.00-0.06) 1-4 = Number of test section

Figure 13: The CFA model of IELTS Listening test performance (N = 376)

The factor loadings ranged from 0.62 (Section 1) to 0.81 (Section 3). To understand how well the observed variables measure a latent variable, the total common factor variance (h^2) can be computed. h^2 is the sum of squared factor loadings. That figure is then divided by the number of variables. $1 - h^2$ is then interpreted as the amount of unexplained variance or the extent to which the latent variable is not determined by the observed variables. In Figure 13, h^2 was 0.52 (i.e., $0.62^2 + 0.76^2 + 0.81^2 + 0.69^2 \div 4 = 2.09 \div 4$). Based on the h^2 value, the observed variables determined 52% of the latent variable, leaving 48% unexplained test performance variance. The total common factor variance was relatively large. This CFA model will be used as a measurement model of IELTS Listening test performance in subsequent SEM analysis. It should be noted that each factor loading can be treated as a correlation coefficient, so the relationships among different sections of the IELTS Listening test can be computed by multiplying two factor values (e.g., $0.62 \times 0.76 = 0.47$ for the correlation coefficient between Sections 1 and 2).

4.1.5 Correlations between single-case appraisal confidence and relative-frequency appraisal confidence

Gigerenzer et al.'s (1991) probabilistic mental model (see Figure 3) distinguishes single-case appraisal confidence from relative-frequency appraisal confidence. This model suggests that both types of confidence judgments *should not be correlated* to each other because they are based on different kinds of reference classes and kinds of cognitive processes (see the review of the literature). The researchers argue that single-case appraisal confidence relies on cues related to a specific task (e.g., types of questions, responses and the time allowed to complete a task), whereas relative-frequency appraisal confidence is based on cues that are specific to overall test contexts (e.g., test instructions, the characteristics of test tasks and the numbers of questions). Gigerenzer et al. (1991) further argue that different reference classes imply different cue validities and they should not be related. To examine whether these two types of appraisal confidence were related, Pearson-Product-Moment correlation analysis was performed. Table 24 reports on the Pearson-Product-Moment correlation coefficients between single-case and relative-frequency confidence. The correlation coefficients between single-case and relative-frequency appraisal confidence were found to be strong (0.75 for Pair 1 to 0.90 for Pair 3). The present findings do not support Gigerenzer et al.'s (1991) probabilistic mental model that specifies that single-case and relative-frequency appraisal confidence types are unrelated.

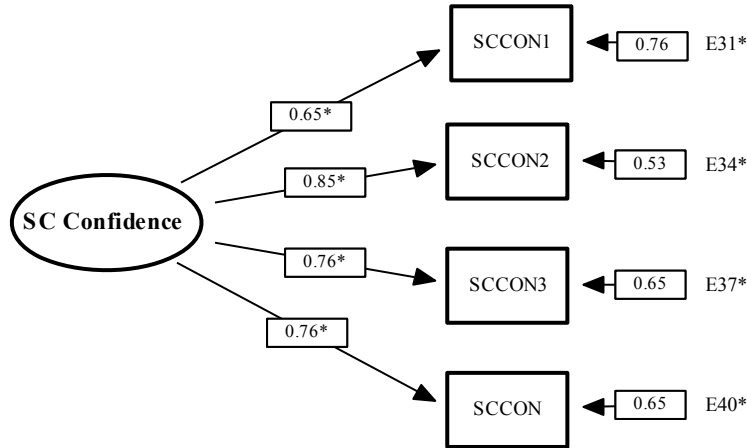
		Correlation
Pair 1	SCCON1 & RFCON 1	0.75** ($R^2 = 0.56$)
Pair 2	SCCON 2 & RFCON 2	0.89** ($R^2 = 0.79$)
Pair 3	SCCON 3 & RFCON 3	0.90** ($R^2 = 0.81$)
Pair 4	SCCON 4 & RFCON 4	0.83** ($R^2 = 0.69$)

SCCON = Single-case Appraisal Confidence RFCON = Relative-frequency Appraisal Confidence ** = $p < 0.01$

Table 24: Pearson-Product-Moment correlations between single-case and relative-frequency confidence (N = 376)

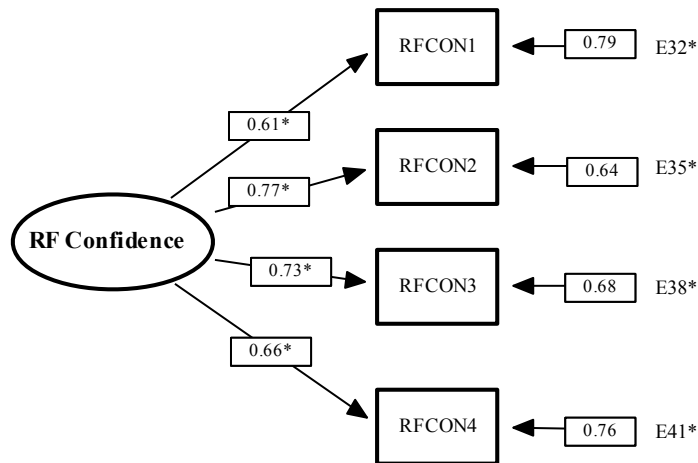
4.1.6 Models of single-case and relative-frequency appraisal confidence

Two CFAs for single-case appraisal confidence and relative-frequency appraisal confidence were performed. Both CFAs exhibited an excellent model fit. Figures 14 and 15 present the CFAs of single-case and relative-frequency appraisal confidence.



Chi-square ($\chi^2_{(1)} = 18.41$) $p = 0.00$ CFI = 0.97 RMSEA = 0.12 (90% CI = 0.10-0.15)
 SCCON = Single-case Appraisal Confidence

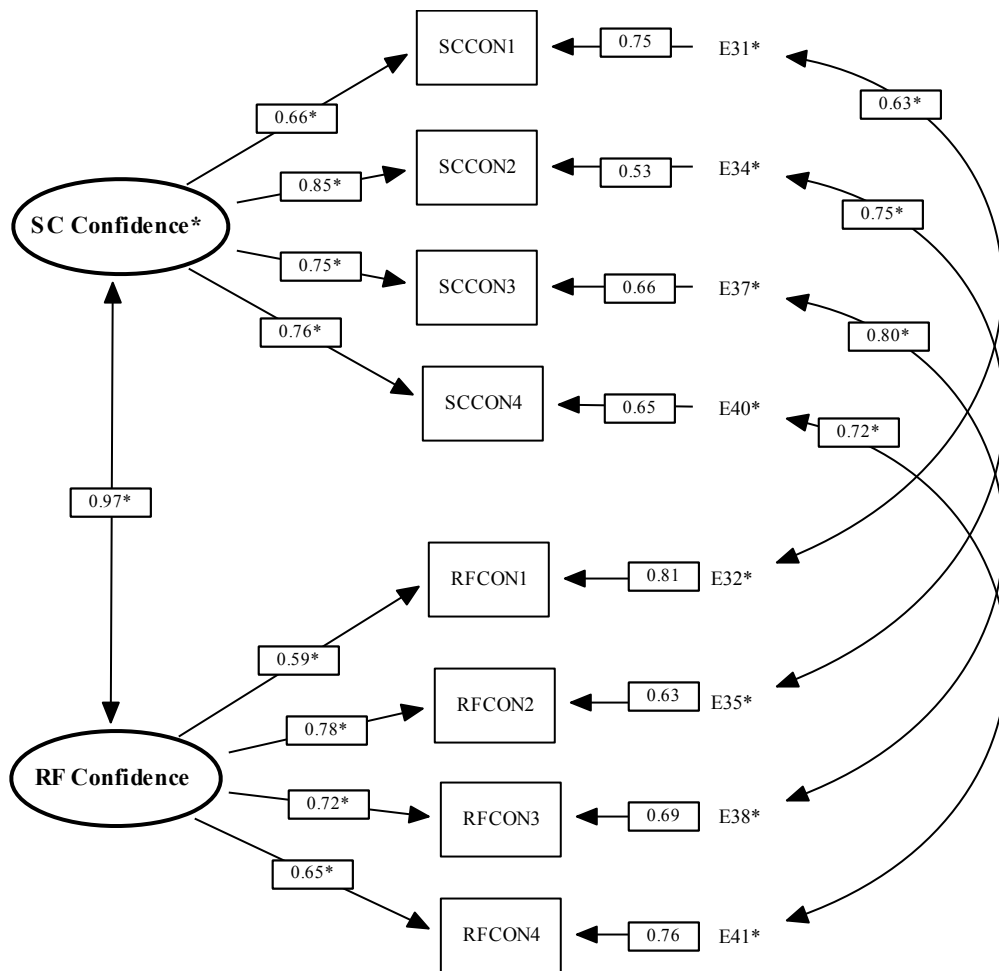
Figure 14: CFA of single-case appraisal confidence (N = 376)



Chi-square ($\chi^2_{(1)} = 5.06$) $p = 0.02$ CFI = 0.99 RMSEA = 0.10 (90% CI = 0.03-0.15)
 RFCON = Relative-frequency Appraisal Confidence

Figure 15: CFAs of relative-frequency appraisal confidence (N=376)

In Figure 14, h^2 was 0.58 (i.e., $0.65^2 + 0.85^2 + 0.76^2 + 0.76^2$) $\div 4 = 2.30/4$), whereas in Figure 15, h^2 was 0.48 (i.e., $0.61^2 + 0.77^2 + 0.73^2 + 0.66^2$) $\div 4 = 1.93/4$). Based on these h^2 values, the observed variables defined 58% and 48% of the latent single-case appraisal confidence and relative-frequency appraisal confidence, respectively. These total common factor variances were reasonable. SEM analysis that connected the two CFAs was next performed to examine the relationship between latent single-case appraisal confidence and relative-frequency appraisal confidence (Figure 16).



Chi-square ($\chi^2_{(13)} = 32.99$ $p = 0.002$, CFI = 0.99 RMSEA = 0.06 (90% CI = 0.04-0.09)
 SCCON = Single-case Appraisal Confidence RFCON = Relative-frequency Appraisal Confidence

Figure 16: The SEM model of the relationship between single-case appraisal confidence and relative-frequency appraisal confidence (N = 376)

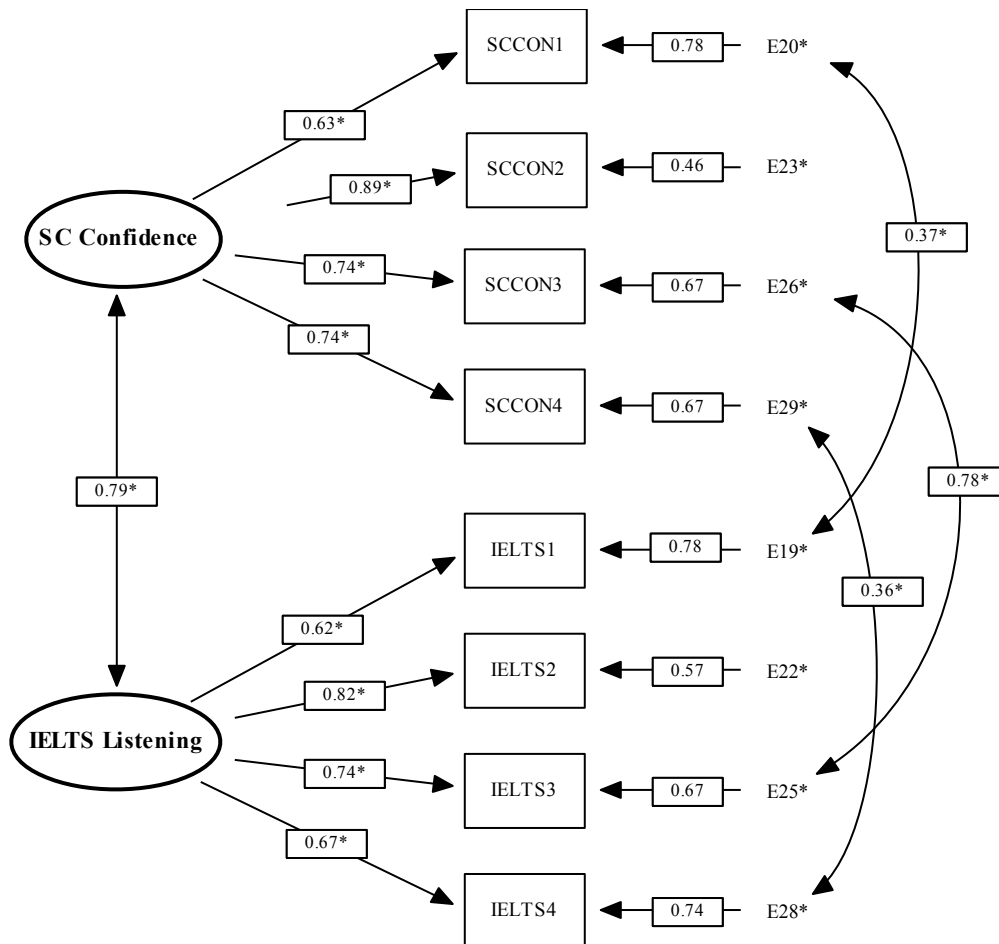
It should be noted that the SEM model in Figure 16 had been tested and re-tested prior to being reported here. In particular, it was found that there was some redundancy in the single-case and relative-frequency confidence in each test section and it might be that non-random errors associated with these variables were related (Bentler 2006). Hence, covariances for non-random measurement errors for each test section pair (i.e., E31 and E32; E34 and E35; E37 and E38; E40 and E41) were added in a re-hypothesised model. This re-hypothesised model was then tested and it was found that the model fitted much better with the data. The standard fit indices suggested a very good model fit of the re-specified SEM model (e.g., the Bentler-Bonett Normed Fit Index (NFI) = 0.99; Bentler-Bonett Non-Normed Fit Index (NNFI) = 0.99; Comparative Fit Index (CFI) = 0.99, and Root Mean-Square Error of Approximation (RMSEA) = 0.06 (90% Confidence Interval (CI) = 0.04, 0.09)). It should be noted that the chi-square, which is known to be sensitive to sample size, was statistically non-significant ($p > 0.001$). In SEM, the chi-square statistic needs to be non-significant to suggest that the data fit the model very well.

According to Figure 16, it was found that the correlation coefficient between latent single-case appraisal confidence and latent relative-frequency appraisal confidence was 0.96 ($R^2 = 0.92$), which was large. The SEM correlation coefficient suggests that test-takers who reported a high level of single-case appraisal confidence would also report a high level of relative-frequency appraisal confidence.

In retrospect, the high level of correlation found in the present study was not surprising since test-takers were asked to rate single-case appraisal confidence for each question before rating their relative-frequency appraisal confidence. Their performance appraisals for each question might have cumulatively influenced their evaluative decision on how well they had performed overall. They rated their relative-frequency appraisal confidence immediately after they had transferred a group of 10 answers with single-case confidence to the answer sheet, rather than at the end of the 40 questions, so test-takers might have been able to use more specific contextual information within a test section to evaluate their performance. The present findings are consistent with Kleitman and Stankov (2001), who found that intra-individual aspects of self-monitoring accounted for the relationship between single-case and relative-frequency appraisal confidence judgments.

4.1.7 SEM correlations between appraisal confidence and IELTS Listening test performance

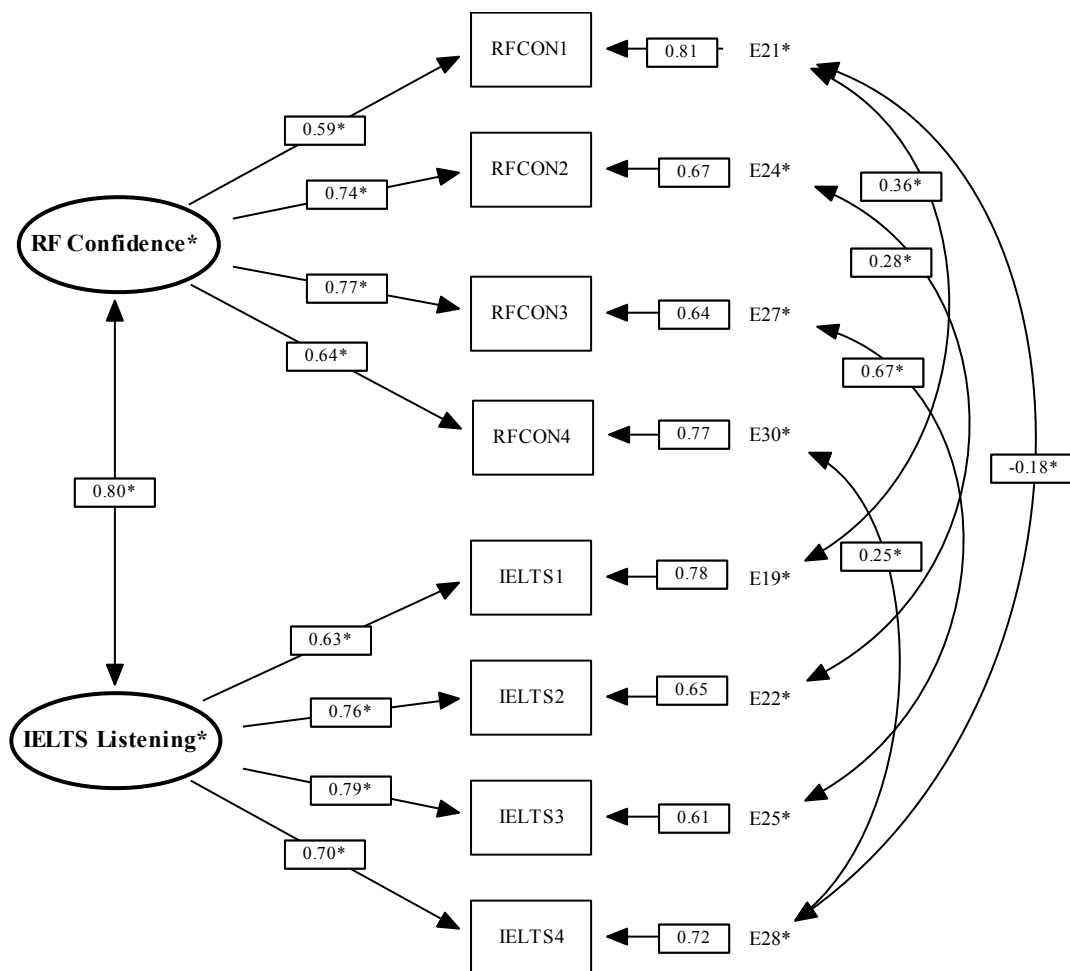
Given the analytical limitations of the subtraction formula and Pearson-Product-Moment correlational analysis discussed above, the relationships between appraisal confidence judgments and IELTS Listening performance were further explored through SEM analysis. The structural models of the latent single-case confidence (as in Figure 18) and the latent relative-frequency confidence (as in Figure 15) to the latent IELTS Listening test performance (as in Figure 13) had excellent model fits. Figures 17 and 18 present the SEM models of the relationship of single-case/relative-frequency confidence to IELTS Listening test performance, respectively. These SEM models had gone through model specifications, which finally included the corrected correlations between some observed single-case appraisal confidence variables and IELTS Listening variables.



Chi-square ($\chi^2_{(14)} = 58.19$ $p = 0.000$ CFI = 0.98 RMSEA = 0.09 (90% CI = 0.07-0.11)
 SCCON = Single-case Appraisal Confidence 1-4 = Number of Test Section

Figure 17: The SEM model of the relationship between the latent single-case appraisal confidence and the latent IELTS Listening test performance (N =376)

In Figure 17, the correlation coefficient between single-case appraisal confidence and IELTS Listening performance was 0.79 ($R^2 = 0.62$). In Figure 18, the correlation coefficient between relative-frequency confidence and IELTS Listening performance was 0.80 ($R^2 = 0.64$). It should be noted that the relationship between relative-frequency appraisal confidence and IELTS Listening test performance might have only been strong because of the preceding single-case appraisal confidence ratings. Both correlation coefficients were found to be larger than the Pearson-Product-Moment coefficient reported in Table 23 (0.73 and 0.70, respectively). The SEM correlation coefficients were found to be larger because SEM examines the relationships through the use of latent variables and the separation of non-random errors which affected parameter estimates. Nonetheless, SEM correlation coefficients between single-case appraisal confidence and IELTS Listening performance per each section were smaller than the corresponding Pearson-Product-Moment coefficients. For each test section, SEM r is computed by multiplying two factor loadings with the correlation value (e.g., in Section 1 (single-case confidence), a SEM correlation coefficient was 0.31 (i.e., $0.63 \times 0.62 \times 0.79$)).



Chi-square ($\chi^2_{(12)} = 27.70$) $p = 0.01$ CFI = 0.99 RMSEA = 0.06 (90% CI = 0.03-0.08)
 RFCON = Relative-frequency Appraisal Confidence 1-4 = Number of Test Section

Figure 18: The SEM model of the relationship between the latent relative-frequency appraisal confidence and the latent IELTS Listening test performance (N =376)

Table 25 compares the SEM correlation coefficients to the Pearson-Product-Moment correlation coefficients.

IELTS Listening	SEM <i>r</i> (single-case)	Pearson-Product-Moment <i>r</i> (single-case)	SEM <i>r</i> (relative-frequency)	Pearson-Product-Moment <i>r</i> (relative-frequency)
Section 1	0.31 ($R^2 = 0.10$)	0.55 ($R^2 = 0.30$)	0.30 ($R^2 = 0.09$)	0.55 ($R^2 = 0.30$)
Section 2	0.58 ($R^2 = 0.34$)	0.61 ($R^2 = 0.37$)	0.45 ($R^2 = 0.20$)	0.55 ($R^2 = 0.30$)
Section 3	0.43 ($R^2 = 0.18$)	0.79 ($R^2 = 0.62$)	0.48 ($R^2 = 0.23$)	0.76 ($R^2 = 0.58$)
Section 4	0.39 ($R^2 = 0.15$)	0.54 ($R^2 = 0.29$)	0.36 ($R^2 = 0.13$)	0.48 ($R^2 = 0.23$)
Overall	0.79 ($R^2 = 0.53$)	0.73 ($R^2 = 0.53$)	0.80 ($R^2 = 0.64$)	0.70 ($R^2 = 0.49$)

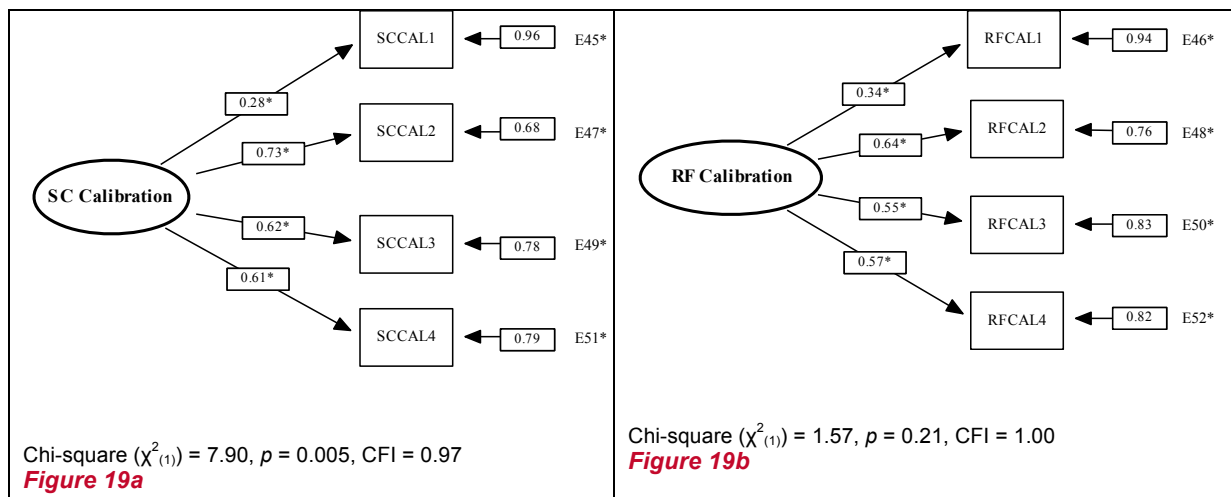
Table 25: Comparisons between SEM and Pearson-Product-Moment correlations (N = 376)

The findings indicate that the strength of the relationship between appraisal confidence judgments and IELTS Listening performance varied from section to section, i.e., they suggest that there are fluctuations in the accuracy of test-takers' appraisals. The comparisons between SEM correlation and Pearson-Product-Moment *r* coefficients reveal quite a pronounced difference between the two types of analytical measures. At the individual section, SEM correlation coefficients were smaller than Pearson-Product-Moment correlation coefficients. In single-case appraisal confidence, the shared variances ranged from 0.10 (Section 1) to 0.34 (Section 2). In relative-frequency appraisal confidence, the shared variances ranged from 0.09 (Section 1) to 0.23 (Section 3).

It should be noted that the SEM correlation coefficients in Section 2 were larger than those in Section 3, which was the other way around to those based on Pearson-Product-Moment analysis. Pearson-Product-Moment correlations used the raw scores, which were assumed to be free of errors to calculate the relationships, whereas SEM used latent variables to examine the relationships and considered the influence of non-random errors on parameter estimates. The SEM coefficients indicate that test-takers' confidence judgments were less predictive of IELTS test scores than Pearson-Product-Moment coefficients. Given the robust statistical procedures of SEM, researchers may be more inclined to use SEM coefficients than Pearson-Product-Moment coefficients.

4.1.8 CFA of appraisal calibration

Table 21 above has presented the appraisal calibration scores of the four IELTS Listening test sections. These scores were used to form observed variables of test-takers' calibration in CFAs. Figure 19 presents the CFAs of test-takers' calibration (i.e., based on single-case and relative frequency appraisal confidence). The fit indices indicate that both CFAs fit the data well. The relative-frequency calibration factor, however, had a better fit than the single-case calibration.

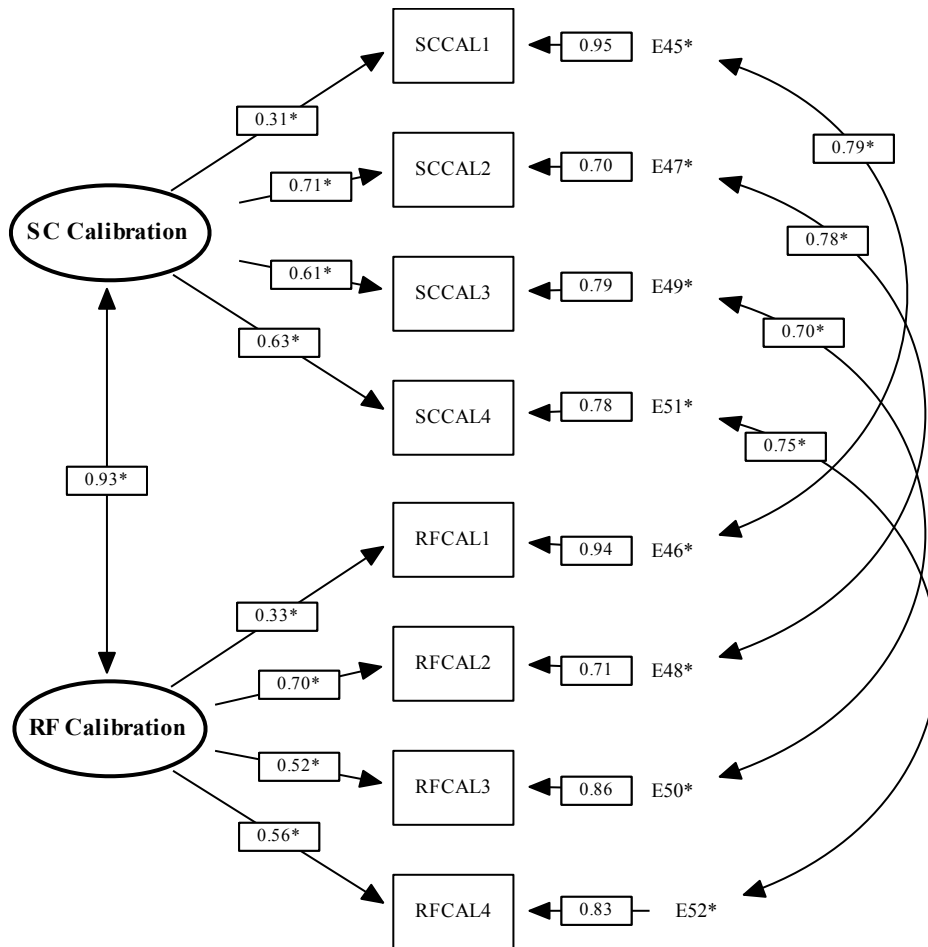


CALSC = Appraisal Calibration (Single-case) CALRF = Appraisal Calibration (Relative-frequency)

Figure 19: The CFAs of single-case appraisal calibration and relative-frequency appraisal calibration

In Figure 19, the factor loadings for appraisal confidence judgments in Section 1 were low (i.e., 0.28 and 0.34). The low factor loadings might be explained by the findings that test-takers were poorly calibrated in this test section. In Figure 19a, the total common factor variance (h^2) was 0.34 (i.e., $0.28^2 + 0.73^2 + 0.62^2 + 0.61^2$)/4 = 1.36/4), whereas in Figure 19b, h^2 was 0.29 (i.e., $0.34^2 + 0.64^2 + 0.55^2 + 0.57^2$)/4 = 1.15/4). Based on the h^2 values, the observed variables defined 34% and 29% of the latent single-case appraisal calibration and relative-frequency appraisal calibration, respectively. These values were considered low, but they accounted for more than 25% of the CFA model variance.

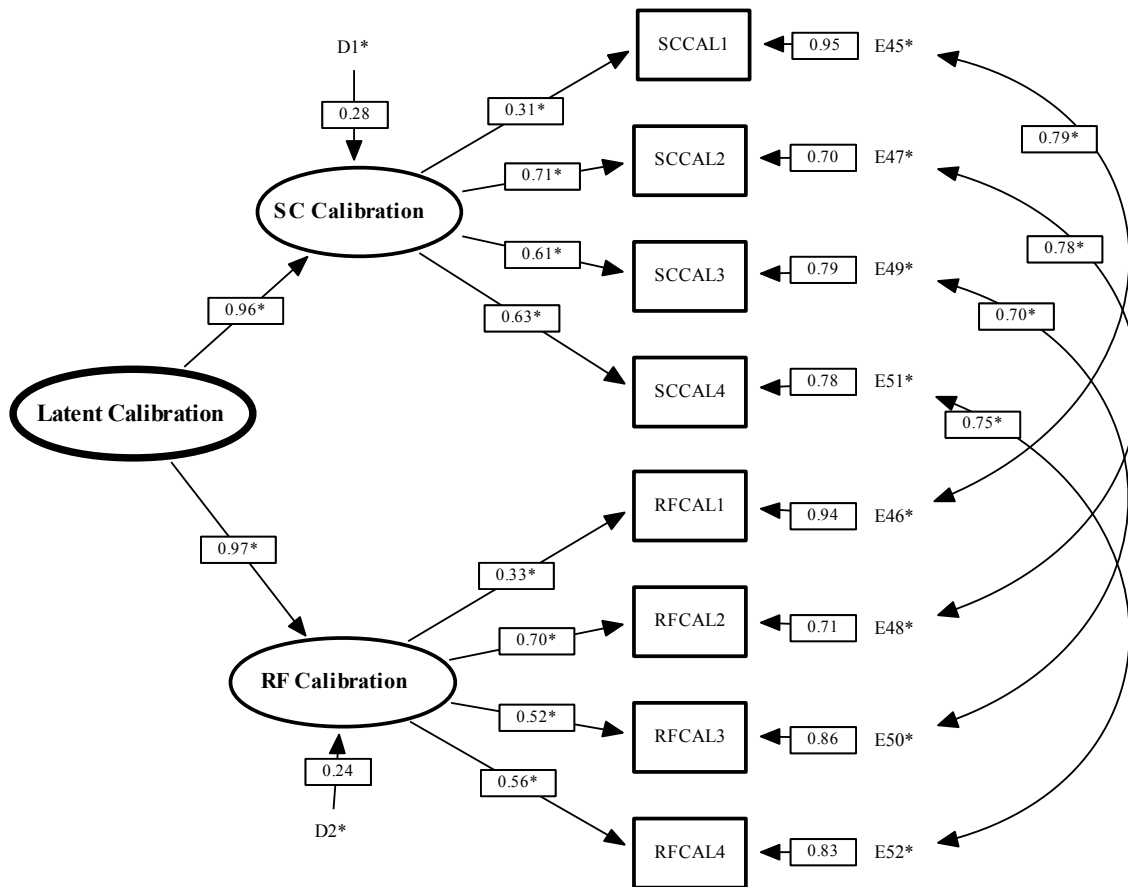
Figure 20 presents an SEM model that tested the interrelationship between latent single-case and relative-frequency calibration. It should be noted that in the model re-specification step, these errors were corrected using a correlation path because they were related to the same test section and had some content redundancy. The standard fit indices suggested very good model fit of the re-specified SEM model.



Chi-square ($\chi^2_{(13)} = 21.94$ $p = 0.06$ CFI = 0.99 RMSEA = 0.04 (90% CI = 0.00-0.07)
 CALSC = Appraisal Calibration (Single-case) CALRF = Appraisal Calibration (Relative-frequency)

Figure 20: SEM model of the relationship between latent single-case and relative-frequency appraisal calibration (N = 376)

A correlation coefficient of 0.93 ($R^2 = 0.86$) was found. This strong correlation suggests that these two factors might be explained by a higher-order latent calibration factor. A second-order confirmatory factor analysis was performed to examine this possibility. Figure 21 presents the second-order CFA of a general calibration factor. The fit statistics indicate very good model fit. Figure 21 shows that latent calibration directly and strongly explains single-case appraisal calibration (0.96; $R^2 = 0.92$) and relative-frequency calibration (0.97; $R^2 = 0.94$).



Chi-square ($\chi^2_{(11)} = 21.94$) $p = 0.02$ CFI = 0.99 RMSEA = 0.05 (90% CI = 0.02-0.08)
 CALSC = Appraisal Calibration (Single-case) CALRF = Appraisal Calibration (Relative-frequency)
 1-4 = Number of Test Section

Figure 21: The second-order CFA of a latent calibration factor (N = 376)

In summary, the first research question has examined the nature of test-takers' appraisal confidence and calibration in an IELTS Listening test. It was found that across different measures of appraisal calibration, test-takers were found to be miscalibrated, exhibiting a tendency to be overconfident in their test performance. This section has explored test-takers' appraisal calibration and confidence through CFAs and SEM analyses. Since Research Question 1 has focused on the whole group of test-takers, the next research question asks whether there are differences in appraisal confidence and calibration among test-takers when performing tasks at different levels of IRT difficulty.

4.2 What is the nature of test-takers' appraisal calibration in easy, moderately difficult, difficult and very difficult IELTS Listening questions?

The previous section suggests variability of test-takers' calibration across different sections. In particular, they tended to be overconfident in Section 1, which was the easiest section, and in Section 4, which was the most difficult section. They tended to have a better calibration in Section 3, which was moderately difficult.

Research Question 2 therefore examines the nature of their appraisal calibration at different test difficulty levels. As discussed in the method section, Rasch IRT was used to identify test difficulty levels (see Figure 5 and Table 12).

4.2.1 Appraisal confidence and performance based on test difficulty levels

Table 26 presents the descriptive statistics of test-takers' IELTS Listening performance and single-case appraisal confidence across different test difficulty levels.

Item	Minimum	Maximum	Mean	SD	Skewness	Kurtosis
EASYQ	0.00	100.00	81.84	19.69	-1.21	1.11
SCCON in EASYQ	0.00	100.00	83.93	16.85	-1.18	0.80
MDQ	0.00	100.00	61.75	25.26	-0.41	-0.75
SCCON in MDQ	13.64	100.00	70.40	20.78	-0.65	-0.43
DIFQ	0.00	100.00	42.93	23.02	0.28	-0.81
SCCON in DIFQ	0.00	100.00	53.61	22.85	-0.22	-0.73
VDIFQ	0.00	77.78	18.03	19.48	1.20	0.79
SCCON in VDIFQ	0.00	100.00	42.95	23.90	0.43	-0.70

EASYQ = Easy Questions MDQ = Moderately Difficult Questions DIFQ = Difficult Questions
VDIFQ = Very Difficult Questions SCCON = Single-case Appraisal Confidence

Table 26: Descriptive statistics of test-takers' IELTS Listening scores and single-case appraisal confidence according to IRT test difficulty levels (N = 376)

According to Table 26, the average performance was as follows: easy questions 82%; moderately difficulty 62%; difficult 43%; and very difficult questions 18%. The differences between single-case appraisal confidence and test performance appeared to be larger as test difficulty levels increased. This was clearly seen in the very difficult and extremely difficult questions.

4.2.2 Paired-samples *t*-tests between appraisal confidence and performance based on question difficulty levels

Table 27 presents the *t*-test results that compared differences between appraisal confidence and test performance based on test difficulty levels. The mean scores in Table 27 were also their appraisal calibration scores. It was found that appraisal confidence and performance at different difficulty levels were significantly different ($p < 0.05$), suggesting that single-case appraisal confidence judgments were larger than test performance. Except in the easy questions, which had a small effect size ($d = 0.14$), the effect sizes of other pairs were medium ($0.49 < d < 0.62$).

		Paired Differences				<i>t</i>	<i>d</i>	
		Mean	SD	SEM	95% Confidence Interval of the Difference			
					Lower			Upper
Pair 1	EASY SCCON - EASYQ	2.09**	15.65	0.81	0.51	3.68	2.59	0.14
Pair 2	MD SCCON - MDQ	8.65**	18.15	0.94	6.81	10.49	9.24	0.49
Pair 3	DIF SCCON - DIFQ	10.68**	21.17	1.09	8.53	12.82	9.78	0.50
Pair 4	VDIF SCCON - VDIFQ	24.92**	19.26	0.99	22.97	26.88	25.10	0.62

EASYQ = Easy Questions MDQ = Moderately Difficult Questions DIFQ = Difficult Questions
VDIFQ = Very Difficult Questions SCCON = Single-case Appraisal Confidence
SEM = Standard Error of Mean ** = $p < 0.01$ (2-tailed)

Table 27: The paired-sample *t*-test results between appraisal confidence and performance based on IRT test difficulty levels (N = 376)

According to Table 27, the average appraisal confidence for easy test questions was 2.1%, suggesting that test-takers in general were quite realistic when appraising their performance in easy questions. They were just overconfident in the moderately difficult questions (8.7%), generally overconfident (11%) in difficult questions and approached extremely overconfident in very difficult questions (25%). The calibration literature discusses the *hard-easy-effect* hypothesis in which individuals exhibit a tendency to be underconfident in easy questions but overconfident in difficult questions. The findings only support this hypothesis partially in that test-takers were found to be overconfident in difficult test questions. Figures 22 to 26 present test-takers' appraisal calibration diagrams based on difficulty level. These diagrams suggest that the majority of the test-takers were overconfident across test difficulty levels. It should be noted that in Figure 22, many test-takers whose test scores were above 70% tended to be underconfident in easy test questions, but those whose test scores were below 55% tended to be overconfident in easy test questions.

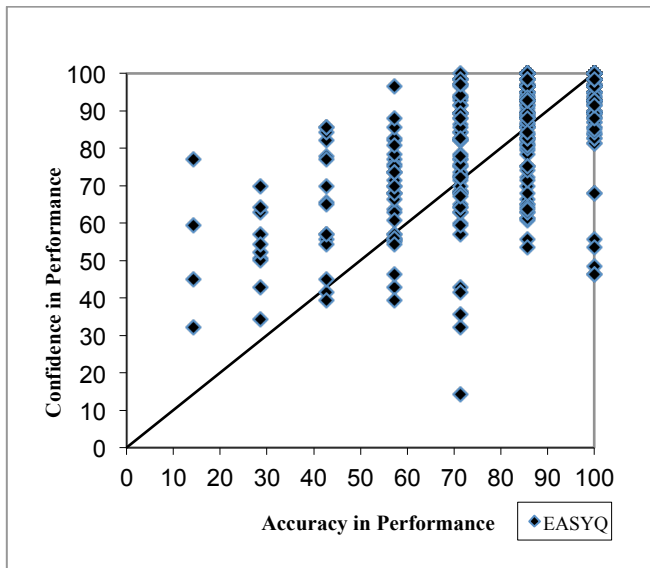


Figure 22: Test-takers' appraisal calibration diagram based on easy questions ($k = 7, N = 376$)

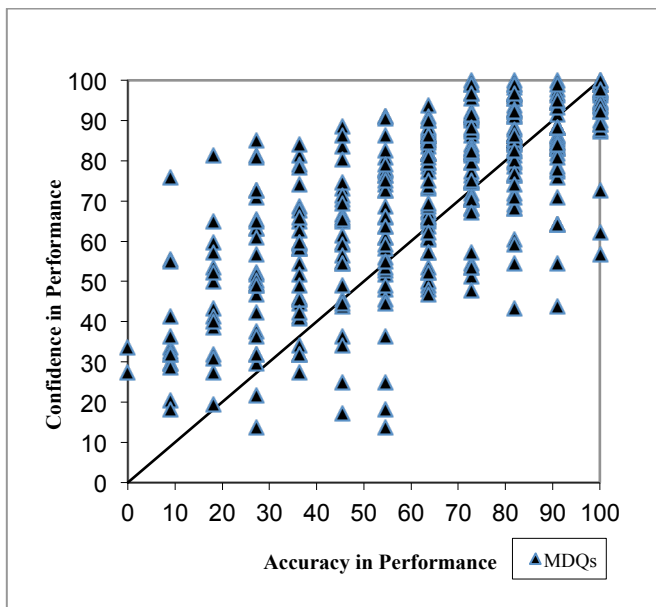


Figure 23: Test-takers' appraisal calibration diagram based on moderately difficult questions ($k = 11, N = 376$)

According to Figures 22 to 24, it appears that test-takers whose test scores were below 50% tended to be overconfident, whereas those whose test scores were above 70% tended to be underconfident. This warrants further analysis on the differences between the appraisal confidence levels of high performers and low performers. Figure 25 suggests that on average, test-takers were overconfident in all difficulty levels.

4.2.3 Correlations between appraisal confidence and performance based on IRT test difficulty levels

Having found the general tendency of test-takers to be overconfident across test difficulty levels, further analysis was carried out to determine the extent to which test-takers' appraisal confidence and performance based on difficulty levels were correlated. To achieve this aim, both SEM analysis and Pearson-Product-Moment correlational analysis were performed and the correlations coefficients will be compared.

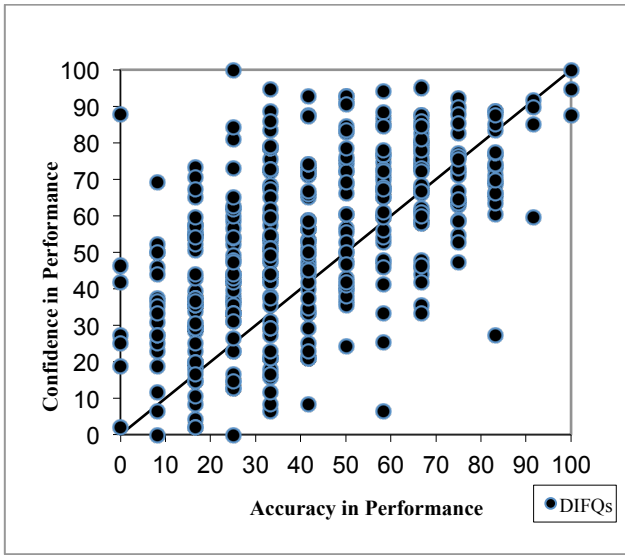


Figure 24: Test-takers' appraisal calibration diagram based on difficult questions ($k = 12, N = 376$)

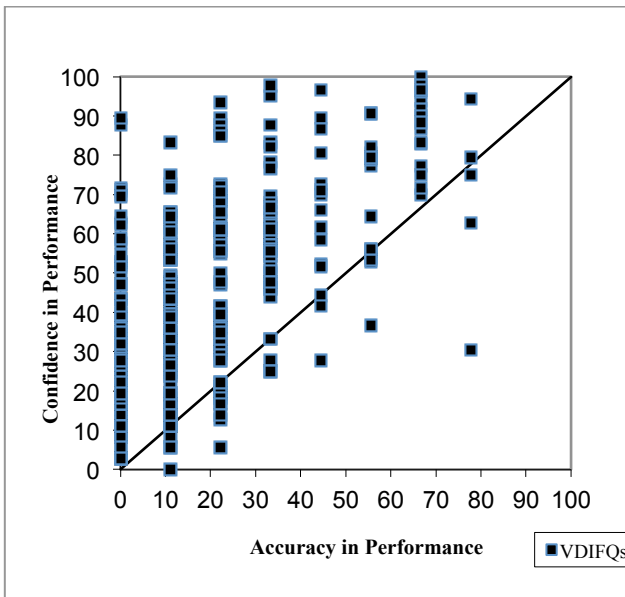


Figure 25: Test-takers' calibration diagram based on very difficult questions ($k = 9, N = 376$)

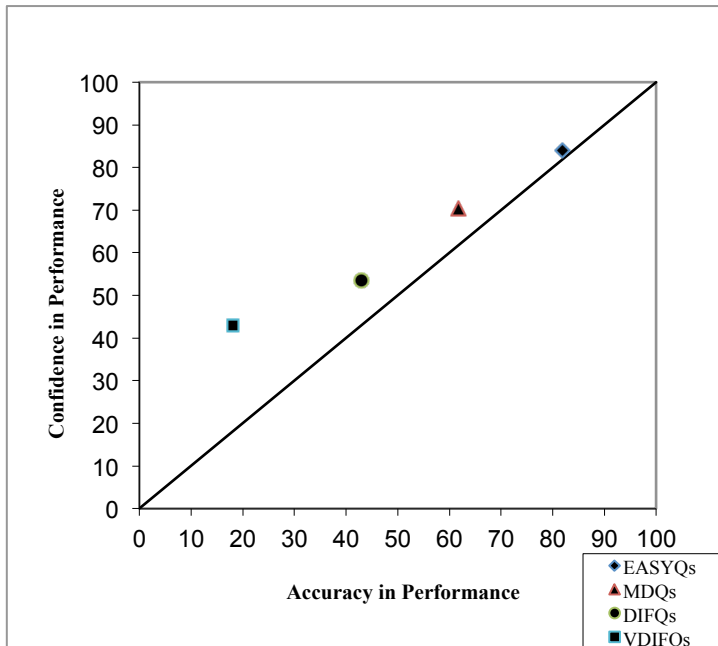


Figure 26: Test-takers' appraisal calibration diagram based on the four difficulty levels (N = 376)

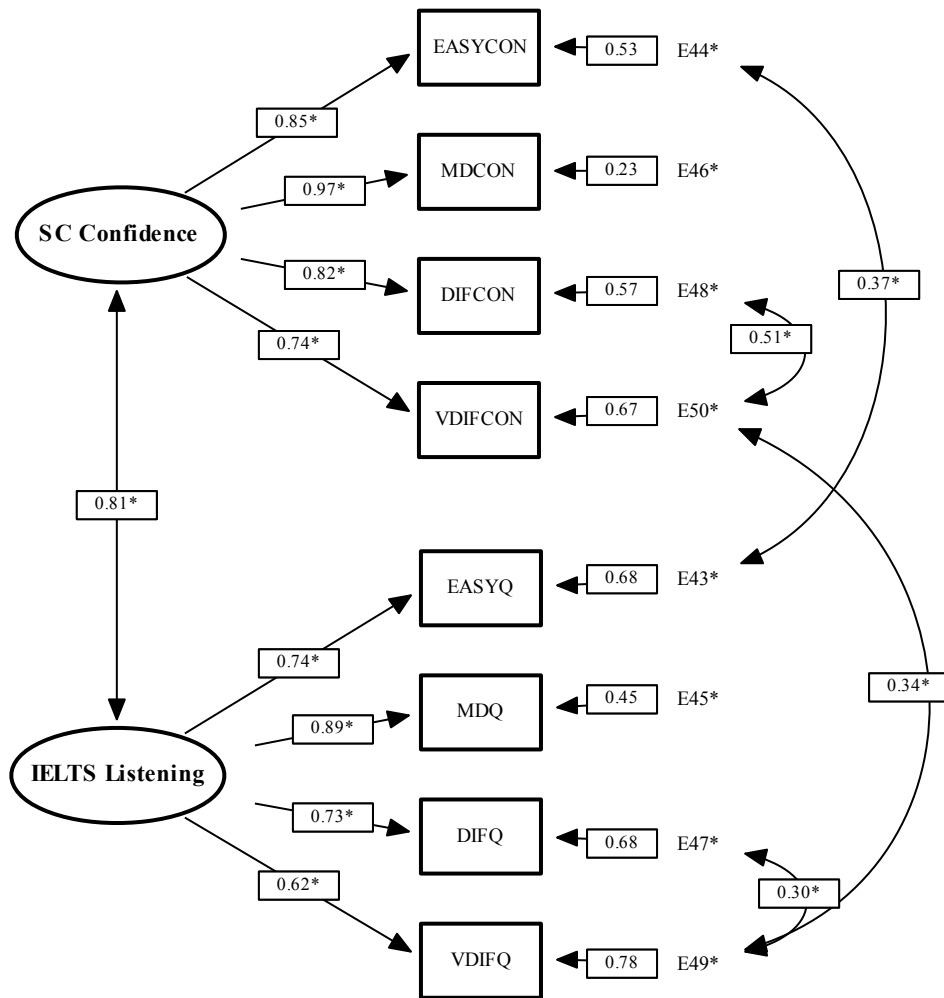
Figure 27 presents the SEM models of the relationship of single-case appraisal confidence to IELTS Listening test performance based on test difficulty levels. This SEM model had gone through model specifications, which in the end included the corrected correlations between some observed single-case appraisal confidence variables and IELTS Listening variables. Figure 27 has an excellent model fit. Compared to Figure 17, which is the SEM model based on IELTS Listening test sections. Figure 27 shows a slightly higher correlation coefficient between appraisal confidence and performance (0.79 versus 0.81).

Table 28 compares the SEM correlation coefficients to the Pearson-Product-Moment correlation coefficients in terms of the relationships between appraisal confidence and test performance based on IRT test difficulty levels. It was found that the SEM correlation coefficients were smaller than the Pearson-Product-Moment coefficients. It should be noted that SEM correlation coefficients were calculated on the basis of factor loadings between two variables, whereas Pearson-Product-Moment correlations were calculated based on raw scores. SEM correlation coefficients are therefore based on how well a construct of interest is measured. According to Table 28, test-takers' confidence judgments accounted for nearly 50% of IELTS test performance in moderately difficult questions. The coefficients from SEM and Pearson-Product-Moment were similar.

This observation that when answering questions at moderately difficult levels, test-takers tended to be realistic or calibrated is consistent with the findings in Research Question 1. However, when the test questions were either very easy or very difficult, test-takers tended to be severely miscalibrated.

In Table 28, the Pearson-Product-Moment correlation coefficients were higher than those produced by SEM. However, the two largely different coefficients were those of very difficult questions (14% shared variance for SEM versus 38% shared variance for Pearson-Product-Moment). Test-takers on average were 25% overconfident in very difficult questions, and the SEM coefficient tended to corroborate with the associated calibration score. Furthermore, since SEM has factored in the reliability of the appraisal confidence and IELTS performance observations, SEM coefficients might indicate a more valid inference about test-takers' calibration.

In summary, it was found that test-takers in general tended to be overconfident at all levels of test difficulty. Their calibration was very poor when answering very difficult questions. Figures 26 to 29 also indicate that test-takers' success levels might play a role in their calibration at various test difficulty levels. In particular, test-takers whose performance scores were above 70% tended to be underconfident, whereas test-takers who scored below 50% of the test tended to be highly overconfident. The next research question investigates whether there were differences in calibration scores between male and female test-takers.



Chi-square ($\chi^2_{(13)}$) = 57.43 $p = 0.000$ CFI = 0.99 RMSEA = 0.10 (90% CI = 0.07-0.12)
 EASYQ = Easy Questions MDQ = Moderately Difficult Questions DIFQ = Difficult Questions
 VDIFQ = Very Difficult Questions SCCON = Single-case Appraisal Confidence

Figure 27: The SEM model of the relationship between the latent single-case appraisal confidence and the latent IELTS Listening test performance based on test difficulty levels (N = 376)

IELTS Listening	SEM r (single-case)	Pearson-Product-Moment r (single-case)
EASYQ	0.50 ($R^2 = 0.25$)	0.64 ($R^2 = 0.41$)
MDQ	0.70 ($R^2 = 0.49$)	0.71 ($R^2 = 0.50$)
DQ	0.48 ($R^2 = 0.23$)	0.57 ($R^2 = 0.32$)
VDQ	0.37 ($R^2 = 0.14$)	0.62 ($R^2 = 0.38$)

EASYQ = Easy Questions MDQ = Moderately Difficult Questions DIFQ = Difficult Questions
 VDIFQ = Very Difficult Questions

Table 28: Comparisons between SEM and Pearson-Product-Moment correlations based on test difficulty levels (N = 376)

4.3 Do male and female test-takers differ in their appraisal confidence and calibration scores in an IELTS Listening test?

This question seeks to find out whether gender differences might be a factor that explains the nature of test-takers' appraisal calibration as presented in Research Questions 2 and 3. Second language acquisition research has robust evidence regarding gender differences in language learning, suggesting that females are better language learners than males (see e.g., Chavez 2001). The literature in language learning strategies also indicates that females use better strategies, particularly metacognitive strategies to help them acquire the target language (see e.g., Oxford 2011). Language testing and assessment research is also interested in explaining factors affecting language test scores and individual characteristics affecting language test performance have been a key research topic (e.g., Kunnan 1995).

In the present study, gender differences are examined in relation to appraisal confidence judgments and appraisal calibration. There is a scarcity in the literature in language testing and assessment that has examined whether males and females differ in confidence judgments and calibration in language test scores. However, two studies that did address this question are described now. Stankov and Lee (2014) conducted a large-scale study that investigated the nature of confidence judgment and calibration over 33 countries ($N = 6,544$). The researchers found that while males and females did not differ in their accuracy scores (like IELTS Listening scores in the current study), males were found to report significantly higher confidence in their judgments than their female counterparts (Cohen's $d = 0.25$, small effect size). This finding was consistent with the findings of the study by Pallier, Wilkinson, Danthiir, Klietman, Knezevic, Stankov and Roberts (2002). However, males were found to be more overconfident than females, suggesting that females were better calibrated (Cohen's $d = 0.18$, small effect size).

Due to the large number of variables that could be examined, it was decided that relative-frequency appraisal confidence variables and their associated calibration scores are not reported here given that the primary interest was in the processes of single-case appraisal confidence judgment during test taking. Table 29 presents the descriptive statistics for male and female test-takers. There were 138 males and 238 females in the study.

According to Table 29, female test-takers had higher scores than male test-takers in all the four IELTS Listening sections, as well as at the four test difficulty levels. Female test-takers' single-case appraisal confidence judgments were also higher than their male counterparts'.

Table 30 presents the appraisal calibration scores of male and female test-takers for each test section as well as at each test difficulty level. According to Table 30, both male and female test-takers tended to be overconfident in each test section and at each test difficulty level because their appraisal calibration scores were above 0% but not within $\pm 5\%$. However, female test-takers' calibration in the easy test questions was very good (0.78%) and male test-takers' calibration within the same category was below 5%. To find out whether male and female test-takers' mean scores significantly differ from each other, a one-way ANOVA was performed. It should be noted that initially multivariate analysis of variance (MANOVA) was to be used. However, the assumptions for MANOVA, such as equality of variance matrices, could not be met.

Table 31 presents the results of the test of the homogeneity of variances assumption. It is important that the Levene statistic for each dependent variable comparison is non-significant ($p > 0.05$) for this assumption not to be violated. According to Table 31, the assumption has been met for all dependent variables, except for three (i.e., the IELTS performance for Section 3, calibration score for Section 3 and the calibration score for the very difficult questions), which will be interpreted with caution.

		Mean	SD	SEM	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
IELTS1	Male	58.62	21.34	1.82	55.02	62.21	0.00	100.00
	Female	65.83	22.51	1.46	62.95	68.70	0.00	100.00
IELTS2	Male	43.41	23.00	1.96	39.53	47.28	0.00	100.00
	Female	46.97	22.58	1.46	44.09	49.86	0.00	100.00
IELTS3	Male	54.42	28.95	2.46	49.55	59.29	0.00	100.00
	Female	60.38	26.23	1.70	57.03	63.73	0.00	100.00
IELTS4	Male	31.38	19.30	1.64	28.13	34.63	0.00	90.00
	Female	32.73	21.32	1.38	30.01	35.45	0.00	100.00
SCCON1	Male	80.82	15.27	1.30	78.25	83.39	43.33	100.00
	Female	83.98	14.87	0.96	82.08	85.88	5.56	100.00
SCCON2	Male	50.24	25.36	2.16	45.97	54.51	0.00	100.00
	Female	53.35	25.72	1.67	50.06	56.63	0.00	100.00
SCCON3	Male	67.91	27.34	2.33	63.31	72.52	0.00	100.00
	Female	70.08	25.89	1.68	66.78	73.39	0.00	100.00
SCCON4	Male	40.58	24.58	2.09	36.45	44.72	0.00	91.00
	Female	44.63	25.67	1.66	41.35	47.90	0.00	100.00
EASYQ	Male	78.99	19.24	1.64	75.75	82.22	14.29	100.00
	Female	83.49	19.80	1.28	80.97	86.02	14.29	100.00
MDQ	Male	56.32	25.62	2.18	52.01	60.64	0.00	100.00
	Female	64.90	24.56	1.59	61.76	68.03	0.00	100.00
DIFQ	Male	40.58	23.00	1.95	36.71	44.45	0.00	91.67
	Female	44.29	22.97	1.49	41.36	47.23	0.00	100.00
VDIFQ	Male	17.79	18.53	1.58	14.67	20.91	0.00	66.67
	Female	18.16	20.04	1.30	15.60	20.72	0.00	77.78
SCCON in EASYQ	Male	83.34	16.67	1.42	80.53	86.14	32.14	100.00
	Female	84.28	16.98	1.10	82.11	86.44	14.29	100.00
SCCON in MDQ	Male	67.47	21.31	1.81	63.88	71.07	13.64	99.09
	Female	72.10	20.32	1.32	69.50	74.69	13.64	100.00
SCCON in DIFQ	Male	51.19	21.96	1.87	47.49	54.88	0.00	92.50
	Female	55.01	23.28	1.51	52.04	57.98	0.00	100.00
SCCON in VDIFQ	Male	41.66	24.16	2.06	37.60	45.73	0.00	96.67
	Female	43.70	23.77	1.54	40.66	46.73	2.78	100.00

SCCON = Single-case Appraisal Confidence EASYQ = Easy Questions MDQ = Moderately Difficult Questions
DIFQ = Difficult Questions VDIFQ = Very difficult Questions 1-4 = Number of Test Section

Table 29: Descriptive statistics of appraisal confidence and IELTS Listening performance between male and female test-takers (N = 376)

Table 32 presents the results of the one-way ANOVA for appraisal confidence and test performance. It was found that female test-takers significantly outperformed their male counterparts in two IELTS test sections: Section 1 ($F(1, 374) = 9.31, p < 0.01, d = -0.33$, small effect size) and Section 3 ($F(1, 374) = 4.17, p < 0.01, d = -0.22$, small effect size). A Cohen's d values of 0.20 and 0.30 are in the 58th (the compared distributions to have a non-overlap of 14.7%) and 62nd percentile (the compared distributions to have a non-overlap of 21.3%), respectively. It should be noted that the Levene Statistic was significant for Section 3, so caution is needed to generalise this finding.

Female test-takers were also found to have significantly better test performance than male test-takers in easy questions ($F(1, 374) = 4.62, p < 0.01, d = -0.23$, small effect size) and moderately difficult questions ($F(1, 374) = 10.31, p < 0.01, d = -0.34$, small effect size).

In regard to appraisal confidence, two appraisal confidence scores were found to be significantly higher for female test-takers than male ones (i.e., $F(1, 374) = 3.87, p = 0.05, d = -0.21$, small effect size for single-case appraisal confidence in Section 1 and $F(1, 374) = 4.37, p < 0.05, d = -0.22$, small effect size in moderately difficult questions). These significant differences had small effect sizes between 0.21 and 0.34. The rest of the variables did not differ significantly between male and female test-takers.

		Mean	SD	SEM	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
CALSC1	Male	22.20	17.74	1.51	19.22	25.19	-25.00	64.44
	Female	18.15	19.39	1.26	15.67	20.63	-61.11	75.00
CALSC2	Male	6.83	19.44	1.65	3.56	10.11	-42.50	57.00
	Female	6.37	22.64	1.47	3.48	9.26	-52.50	83.50
CALSC3	Male	13.49	16.09	1.37	10.79	16.20	-37.50	57.50
	Female	9.70	18.07	1.17	7.40	12.01	-47.50	65.50
CALSC4	Male	9.21	19.67	1.67	5.90	12.52	-40.00	65.00
	Female	11.90	23.97	1.55	8.83	6.00	-65.00	83.50
CALSC in EASYQ	Male	4.35	13.78	1.17	2.03	6.67	-39.29	62.86
	Female	0.78	16.52	1.07	-1.33	2.89	-57.14	45.00
CALSC in MDQ	Male	11.15	17.17	1.46	8.26	14.04	-28.18	54.09
	Female	7.20	18.57	1.20	4.83	9.57	-47.27	66.82
CALSC in DIFQ	Male	10.61	19.10	1.62	7.39	13.82	-33.33	56.67
	Female	10.75	22.29	1.44	7.90	13.60	-56.25	87.92
CALSC in VDIFQ	Male	23.87	16.89	1.44	21.03	26.71	-18.89	71.67
	Female	25.53	20.51	1.33	22.92	28.15	-47.22	89.44

CALSC = Calibration (Single-case) EASYQ = Easy Questions MDQ = Moderately Difficult Questions
DIFQ = Difficult Questions VDIFQ = Very Difficult Questions 1-4 = Number of Test Section

Table 30: Descriptive statistics of male and female test-takers' appraisal calibration scores (N = 376)

	Levene Statistic	df1	df2	Sig.
IELTS1	0.36	1	374	0.55
IELTS2	0.03	1	374	0.87
IELTS3	3.04	1	374	0.02
IELTS4	1.16	1	374	0.22
SCCON1	2.71	1	374	0.10
SCCON2	0.09	1	374	0.67
SCCON3	1.11	1	374	0.23
SCCON4	0.11	1	374	0.74
CALSC1	0.89	1	374	0.35
CALSC2	2.85	1	374	0.09
CALSC3	0.45	1	374	0.50
CALSC4	6.99	1	374	0.01
EASYQ	0.15	1	374	0.70
MDQ	1.77	1	374	0.18
DIFQ	0.14	1	374	0.71
VDIFQ	1.25	1	374	0.26
SCCON in EASYQ	0.11	1	374	0.74
SCCON in MDQ	0.56	1	374	0.46
SCCON in DIFQ	0.35	1	374	0.56
SCCON in VDIFQ	0.04	1	374	0.84
CALSC in EASYQ	0.96	1	374	0.33
CALSC in MDQ	0.34	1	374	0.56
CALSC in DIFQ	2.71	1	374	0.10
CALSC in VDIFQ	5.81	1	374	0.02

CALSC = Appraisal Calibration (Single-case) SCCON = Single-case Appraisal Confidence
EASYQ = Easy Questions MDQ = Moderately Difficult Questions DIFQ = Difficult Questions
VDIFQ = Very Difficult Questions 1-4 = Number of Test Section

Table 31: Test of homogeneity of variances

	Mean Square	F	Sig.	Cohen's d
IELTS1	4542.37	9.31	0.002	-0.33
IELTS2	1112.65	2.15	0.143	-0.16
IELTS3	3100.62	4.17	0.042	-0.22
IELTS4	160.21	0.38	0.539	-0.07
SCCON1	872.45	3.87	0.050	-0.21
SCCON2	843.52	1.29	0.257	-0.12
SCCON3	410.36	0.59	0.444	-0.08
SCCON4	1426.90	2.23	0.136	-0.16
EASYQ	1775.07	4.62	0.032	-0.23
MDQ	6419.62	10.31	0.001	-0.34
DIFQ	1204.26	2.28	0.132	-0.16
VDIFQ	11.75	0.03	0.861	-0.02
SCCON in EASYQ	77.40	0.27	0.602	-0.06
SCCON in MDQ	1868.30	4.37	0.037	-0.22
SCCON in DIFQ	1275.53	2.45	0.118	-0.17
SCCON in VDIFQ	360.85	0.63	0.427	-0.09

SCCON = Single-case Appraisal Confidence EASYQ = Easy Questions MDQ = Moderately Difficult Questions
DIFQ = Difficult Questions VDIFQ = Very Difficult Questions 1-4 = Number of Test Section

Table 32: Result of the one-way ANOVA for IELTS Listening scores and single-case appraisal confidence

Table 33 presents the ANOVA results for test-takers' appraisal calibration scores. It was found that female test-takers were better calibrated than their male counterparts in four out of eight appraisal calibration scores (IELTS Sections 1 and 3 (i.e., $F(1, 374) = 4.05, p < 0.05, d = 0.22$, small effect size for Section 1, and $F(1, 374) = 4.16, p < 0.05, d = 0.22$, small effect size, respectively) and easy and moderately difficult questions ($F(1, 374) = 4.58, p < 0.05, d = 0.24$, small effect size $F(1, 374) = 4.17, p < 0.05, d = 0.22$, small effect size, respectively).

It should be noted that both genders were overconfident in almost all appraisal calibration scores, except in easy questions for which female test-takers had an appraisal calibration score of 0.78, whereas male test-takers had an appraisal calibration score of 4.35, indicating that test-takers were quite realistic in their appraisal confidence for easy questions.

	Mean Square	F	Sig.	Cohen's d
CALSC1	1433.37	4.05	0.045	0.22
CALSC2	18.60	0.04	0.841	-0.02
CALSC3	1255.00	4.16	0.042	0.22
CALSC4	630.86	1.25	0.265	-0.12
CALSC in EASYQ	1111.15	4.58	0.033	0.24
CALSC in MDQ	1361.52	4.17	0.042	0.22
CALSC in DIFQ	1.79	0.00	0.950	-0.00
CALSC in VDIFQ	242.39	0.65	0.420	-0.09

CALSC = Appraisal Calibration (Single-case) SCCON = Single-case Appraisal Confidence
EASYQ = Easy Questions MDQ = Moderately Difficult Questions DIFQ = Difficult Questions
VDIFQ = Very Difficult Questions 1-4 = Number of Test Section

Table 33: Result of the one-way ANOVA for appraisal calibration scores

Figures 28 to 31 compare the appraisal calibration diagrams of male and female test-takers in the four significant variables.

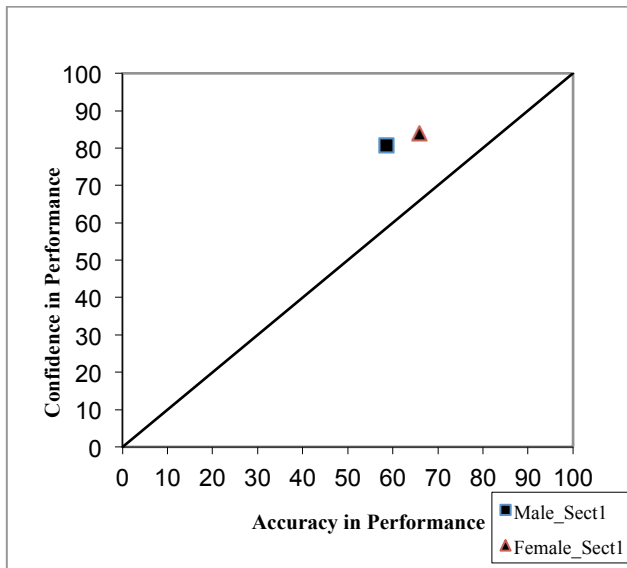


Figure 28: Male and female test-takers' appraisal calibration diagram in Section 1

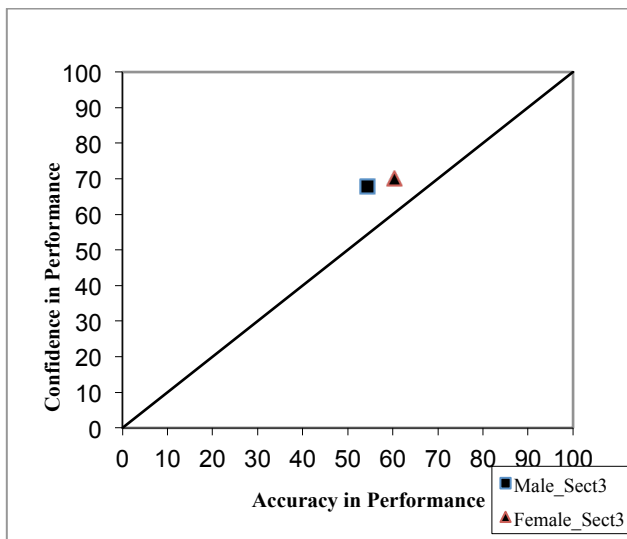


Figure 29: Male and female test-takers' appraisal calibration diagram in Section 3

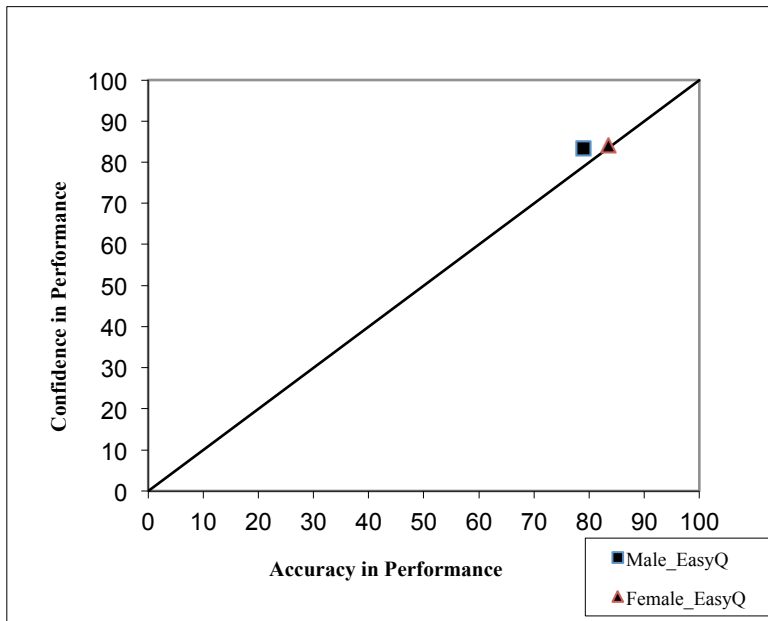


Figure 30: Male and female test-takers' appraisal calibration diagram in easy questions

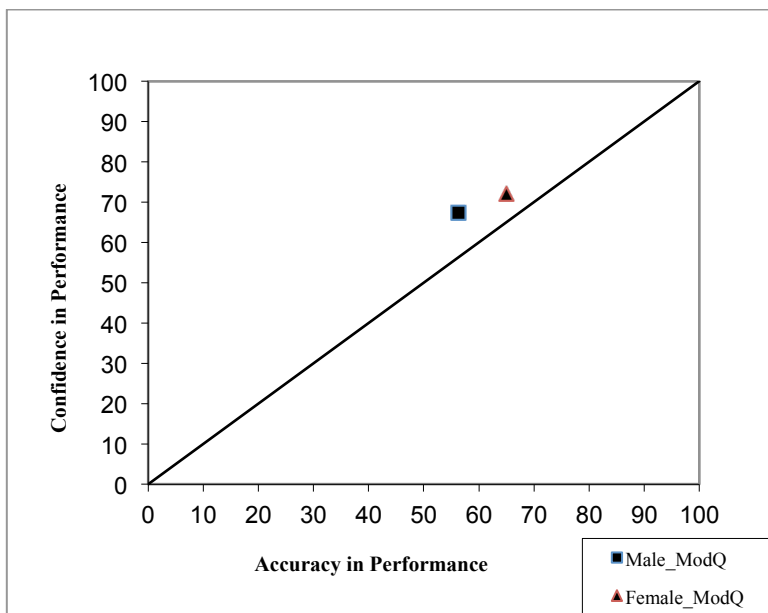


Figure 31: Male and female test-takers' appraisal calibration diagram in moderately difficult questions

In summary, both genders were found to be overconfident in their test performance, but female test-takers exhibited significantly better calibration scores than their male counterparts. The next section presents some statistical comparisons of calibration scores between test-takers with different ability levels as identified by the use of Rasch IRT analysis.

4.4 Do test-takers with different ability levels differ in their appraisal calibration scores?

One of the aims of this research question is to investigate whether success or ability levels explain differences in IELTS Listening test-takers' appraisal calibration scores. In order to achieve this aim, the Rasch IRT person-ability statistics (see Appendix 2) were used to classify test-takers into different ability groups. Rasch IRT assigns a logit score for each individual test-taker. The logit scores range across positive and negative values with 0 being assigned to test-takers with the average ability of the group in question (see Figure 5). Based on the person-ability statistics (logit scores), six groups of test-takers were identified: Group 1 (Logit 1.10 to Logit 3.76, N = 70), Group 2 (Logit 0.50 to Logit 0.94, N = 50), Group 3 (Logit 0.09 to Logit 0.37, N = 54), Group 4 (Logit -0.31 to Logit -0.04, N = 51), Group 5 (Logit -0.85 to Logit -0.44, N = 74) and Group 6 (Logit -2.78 to Logit -1.00, N = 77). Group 1 was the highest ability group, whereas Group 6 was the lowest ability group. Figure 32 presents the distribution of test-takers according to the IRT logit scores and Figure 33 presents the distribution of test-taker groups.

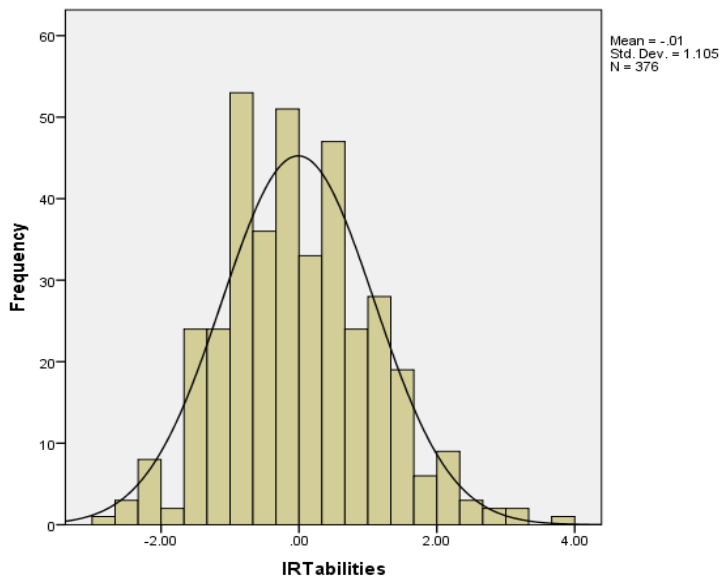


Figure 32: Distribution of test-takers based on IRT ability (N = 376)

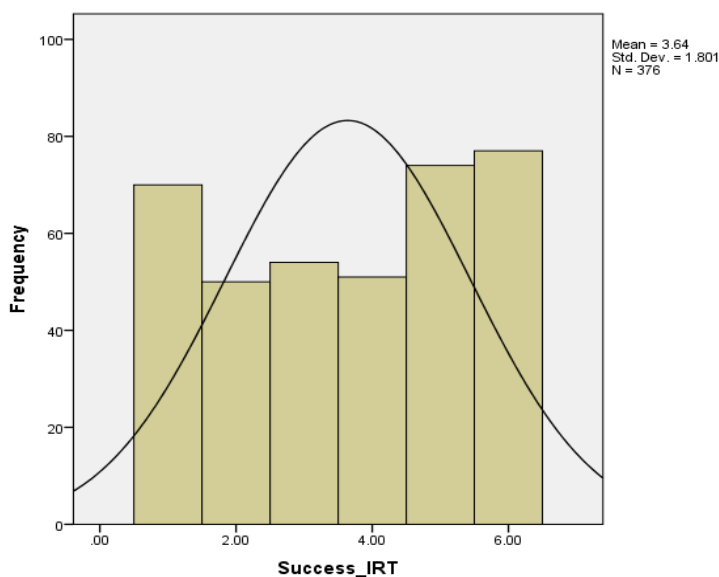


Figure 33: Distribution of the six test-taker groups based on the IRT ability (N = 376)

Initially, the study only aimed to classify test-takers into high-ability, medium-ability and low-ability groups on the basis of their overall raw scores. It was found that these classifications resulted in groups that could not be compared statistically due to their unequal numbers (high-ability = 83, medium-ability = 169 and low-ability = 124). Therefore, the use of IRT logit scores to divide test-takers into different ability groups was adopted. However, using a greater number of ability groups for this investigation caused some limitations in the statistical analyses that could be performed. Initially, it was intended that multivariate analysis of variance (MANOVA) be used to address this research question. Ability-level and gender variables were to be used as the two independent factors in MANOVA. It was found that the Box's M test of the assumption of equality of covariance matrices indicated statistical significance, suggesting that this assumption for MANOVA was violated. The Levene's statistics of equality of error variances were found to be significant for most of the dependent variables (e.g., appraisal calibration scores across different test sections). A one-way ANOVA, which focused on one independent factor was then employed to address this research question.

4.4.1 ANOVA results on appraisal calibration scores among the six ability groups

Table 34 presents the test of homogeneity of variances among the dependent variables. It was found that only three dependent variables met the homogeneity of variance assumption for ANOVA (i.e., CAL Section 1, CAL Section 2 and CALVDIF) since they were non-significant. Based on this, only three dependent variables will be interpreted.

	Levene Statistic	df1	df2	Sig.
CALSC1	2.025	5	370	0.074
CALSC2	1.793	5	370	0.113
CALSC3	3.736	5	370	0.003
CALSC4	4.718	5	370	0.000
CALSC in EASYQ	7.487	5	370	0.000
CALSC in MDQ	2.336	5	370	0.042
CALSC in DIFQ	5.096	5	370	0.000
CALSC in VDIFQ	1.283	5	370	0.270

CALSC = Appraisal Calibration (Single-case) EASYQ = Easy Questions MDQ = Moderately Difficult Questions
DIFQ = Difficult Questions VDIFQ = Very Difficult Questions 1-4 = Number of Test Section

Table 34: Test of homogeneity of variances

Table 35 presents the descriptive statistics of the eight calibration variables. It should be recalled that the mean score closer or equal to zero indicates that a group approaches good appraisal calibration.

		N	Mean	SD	SEM	95% Confidence Interval for Mean		Minimum	Maximum
						Lower Bound	Upper Bound		
CALSC1	1.00	70	5.71	14.54	1.74	2.25	9.18	-33.33	39.44
	2.00	50	18.41	15.32	2.17	14.06	22.77	-16.67	51.67
	3.00	54	18.14	17.03	2.32	13.49	22.79	-34.44	64.44
	4.00	51	18.31	15.82	2.22	13.86	22.76	-11.67	55.56
	5.00	74	21.25	18.39	2.14	16.98	25.51	-61.11	55.56
	6.00	77	33.48	18.58	2.12	29.26	37.69	-3.89	75.00
	Total	376	19.64	18.88	0.97	17.72	21.55	-61.11	75.00

SEM = Standard error of mean

Table 35: Descriptive statistics of test-takers' appraisal calibration scores (N = 376)

		N	Mean	SD	SEM	95% Confidence Interval for Mean		Minimum	Maximum
						Lower Bound	Upper Bound		
CALSC2	1.00	70	2.06	18.93	2.26	-2.46	6.57	-52.50	46.50
	2.00	50	3.08	18.76	2.65	-2.25	8.41	-45.00	45.00
	3.00	54	3.74	20.58	2.80	-1.88	9.36	-40.00	47.00
	4.00	51	7.09	25.87	3.62	-0.19	14.37	-45.00	60.00
	5.00	74	12.82	20.64	2.40	8.04	17.61	-32.50	61.50
	6.00	77	8.43	22.54	2.58	3.31	13.54	-37.50	83.50
	Total	376	6.54	21.49	1.11	4.36	8.72	-52.50	83.50
CALSC3	1.00	70	1.11	12.91	1.54	-1.96	4.19	-42.50	30.00
	2.00	50	7.90	14.74	2.08	3.71	12.09	-47.50	40.00
	3.00	54	9.78	15.68	2.13	5.50	14.06	-37.50	40.00
	4.00	51	14.34	18.57	2.60	9.12	19.57	-40.00	52.50
	5.00	74	17.93	16.86	1.96	14.02	21.83	-20.00	57.50
	6.00	77	14.45	19.37	2.21	10.05	18.84	-25.00	65.50
	Total	376	11.09	17.45	0.90	9.33	12.86	-47.50	65.50
CALSC4	1.00	70	4.36	20.21	2.42	-0.46	9.18	-65.00	46.00
	2.00	50	11.48	17.21	2.43	6.59	16.37	-40.00	45.50
	3.00	54	16.73	27.96	3.80	9.10	24.36	-50.00	83.50
	4.00	51	14.39	28.01	3.92	6.51	22.26	-35.00	83.00
	5.00	74	10.24	21.15	2.46	5.34	15.14	-40.00	58.00
	6.00	77	10.75	19.28	2.00	6.37	15.12	-25.00	65.00
	Total	376	10.91	22.50	1.16	8.63	13.19	-65.00	83.50
CALSC in EASYQ	1.00	70	-2.67	10.67	1.27	-5.22	-0.13	-46.43	20.00
	2.00	50	0.24	11.35	1.60	-2.98	3.47	-53.57	16.43
	3.00	54	-0.17	12.53	1.71	-3.59	3.25	-32.14	27.14
	4.00	51	-0.29	15.80	2.21	-4.74	4.15	-53.57	41.43
	5.00	74	5.44	16.98	1.97	1.50	9.37	-57.14	42.86
	6.00	77	7.58	19.98	2.28	3.04	12.11	-51.43	62.86
	Total	376	2.09	15.65	0.81	0.51	3.68	-57.14	62.86
CALSC in MDQ	1.00	70	-2.25	13.37	1.60	-5.43	0.94	-43.18	26.36
	2.00	50	5.07	13.77	1.95	1.16	8.99	-38.64	35.91
	3.00	54	2.80	15.86	2.16	-1.53	7.13	-47.27	30.00
	4.00	51	8.96	17.78	2.49	3.96	13.96	-25.00	53.64
	5.00	74	13.75	18.06	2.10	9.56	17.93	-40.91	54.09
	6.00	77	19.86	18.69	2.13	15.62	24.11	-36.36	66.82
	Total	376	8.65	18.14	0.94	6.81	10.49	-47.27	66.82

CALSC = Appraisal Calibration (Single-case) EASYQ = Easy Questions MDQ = Moderately Difficult Questions
DIFQ = Difficult Questions VDIFQ = Very Difficult Questions 1-4 = Number of Test Section

Table 35: Descriptive statistics of test-takers' appraisal calibration scores (N = 376) (continued)

		N	Mean	SD	SEM	95% Confidence Interval for Mean		Minimum	Maximum
						Lower Bound	Upper Bound		
CALSC in DIFQ	1.00	70	-1.34	17.05	2.04	-5.41	2.73	-56.25	42.50
	2.00	50	5.78	14.21	2.01	1.75	9.82	-33.33	28.33
	3.00	54	13.02	26.20	3.57	5.87	20.18	-52.08	61.25
	4.00	51	15.92	23.61	3.31	9.28	22.56	-32.92	75.00
	5.00	74	15.25	19.00	2.21	10.85	19.66	-25.00	56.67
	6.00	77	15.37	20.17	2.30	10.79	19.95	-27.08	87.92
	Total	376	10.70	21.15	1.09	8.55	12.84	-56.25	87.92
CALSC in VDIFQ	1.00	70	20.70	20.24	2.42	15.87	25.52	-47.22	64.44
	2.00	50	29.27	18.25	2.58	24.08	34.45	-8.33	69.44
	3.00	54	31.30	20.56	2.80	25.68	36.91	-9.44	72.22
	4.00	51	26.17	22.36	3.13	19.88	32.45	-16.67	89.44
	5.00	74	25.42	16.19	1.88	21.67	29.17	-16.67	54.44
	6.00	77	20.17	16.91	1.93	16.33	24.01	-11.11	70.00
	Total	376	24.92	19.26	0.99	22.97	26.88	-47.22	89.44

CALSC = Appraisal Calibration (Single-case) EASYQ = Easy Questions MDQ = Moderately Difficult Questions
DIFQ = Difficult Questions VDIFQ = Very Difficult Questions 1-4 = Number of Test Section

Table 35: Descriptive statistics of test-takers' appraisal calibration scores (N = 376) (continued)

From the descriptive statistics in Table 35, it can be seen that in Section 1, the highest ability group approached good appraisal calibration (+5.71%), whereas the rest of the groups were generally overconfident, with the lowest ability group being extremely overconfident (+33.48%). In Section 2, Groups 1 to 3 had good appraisal calibration scores (within a ±5% range), whereas Groups 4 to 6 were just overconfident (less than +10%). In Section 4, all test-taker groups tended to be overconfident. Regarding the appraisal calibration scores based on test difficult levels, Group 1 tended to have good appraisal calibration scores in the first three difficulty levels (within a ±5% range), although they were underconfident. In very difficult questions, all groups exhibited extreme overconfidence, with Group 3 being 31% overconfident while Groups 1 and 6 having surprisingly similar appraisal calibration scores.

The ANOVA test suggests that of the three dependent variables that had not violated the homogeneity of equal variance assumption, only one statistically significant difference was found. The appraisal calibration scores in Section 1 were found to be significant across the six groups ($F(5, 370) = 20.32, p < 0.01$). In order to identify where differences occurred, the *Scheffe post hoc* test was performed. Table 36 presents the *post hoc* test results, which reveal that Group 1 was the most calibrated group. Its calibration score was significantly different from those of the other five test-taker groups, with large effect sizes (-0.79 to -1.68). Group 6 was the least calibrated group as its calibration scores were significantly larger than the other five groups, with almost all having large effect sizes (except Group 5, which had a medium effect size).

It appears that ability level plays an important role in influencing test-takers' appraisal calibration. The *Scheffe post hoc* test also detected statistically significant differences among three other dependent variables that had not met the homogeneity of variance assumption (i.e., calibration scores in Section 3 and calibration scores in moderately difficult and difficult sections).

Dependent Variable	(I) Success_ IRT	(J) Success_ IRT	Mean Difference (I-J)	SEM	Sig.	95% Confidence Interval		Cohen's d
						Lower Bound	Upper Bound	
CALSC1	1.00	2.00	-12.70**	3.12	0.006	-23.13	-2.27	-0.85
		3.00	-12.42**	3.05	0.006	-22.63	-2.22	-0.79
		4.00	-12.60**	3.10	0.006	-23.00	-2.23	-0.83
		5.00	-15.53**	2.81	0.000	-24.92	-6.14	-0.94
		6.00	-27.76**	2.78	0.000	-37.07	-18.46	-1.68
	2.00	3.00	0.27	3.30	1.000	-10.78	11.33	N/A
		4.00	0.10	3.35	1.000	-11.11	11.31	N/A
		5.00	-2.84	3.08	0.974	-13.15	7.48	N/A
		6.00	-15.07**	3.07	0.000	-25.30	-4.84	-0.89
	3.00	4.00	-0.17	3.29	1.000	-11.17	10.83	N/A
		5.00	-3.11	3.01	0.957	-13.19	6.97	N/A
		6.00	-15.34**	2.99	0.000	-25.34	-5.34	-0.86
	4.00	5.00	-2.93	3.06	0.969	-13.19	7.32	N/A
		6.00	-15.17**	3.04	0.000	-25.34	-5.00	-0.88
	5.00	6.00	-12.23**	2.74	0.002	-21.40	-3.06	-0.66
CALSC3	1.00	4.00	-13.23**	3.05	0.002	-23.44	-3.02	-0.84
		5.00	-16.81**	2.76	0.000	-26.06	-7.56	-1.13
		6.00	-13.34**	2.74	0.000	-22.49	-4.17	-0.83
CALSC in MDQ	1.00	4.00	-11.20*	3.04	0.020	-21.38	-1.02	-0.72
		5.00	-15.99**	2.76	0.000	-25.21	-6.77	-1.02
		6.00	-22.12**	2.73	0.000	-31.24	-12.98	-1.40
	2.00	6.00	-14.79**	3.00	0.000	-24.84	-4.75	-0.91
	3.00	5.00	-10.94*	2.96	0.019	-20.84	-1.05	-0.65
		6.00	-17.06**	2.94	0.000	-26.88	-7.24	-0.99
	4.00	6.00	-10.91*	2.98	0.022	-20.89	-0.92	-0.60
CALSC in DIFQ	1.00	3.00	-14.36*	3.66	0.010	-26.62	-2.10	-0.66
		4.00	-17.26**	3.73	0.001	-29.72	-4.79	-0.85
		5.00	-16.59**	3.37	0.000	-27.88	-5.31	-0.92
		6.00	-16.71**	3.34	0.000	-27.89	-5.53	-0.90

CALSC = Appraisal Calibration (Single-case) EASYQ = Easy Questions MDQ = Moderately Difficult Questions
DIFQ = Difficult Questions VDIFQ = Very Difficult Questions 1-4 = Number of Test Section
* $p < 0.05$ ** $p < 0.01$

Table 36: The Scheffe post hoc test in Sections 1 and 3, moderately difficult questions and difficult questions among the six ability groups (N = 376)

It was found that Group 1 had a better appraisal calibration score than Groups 4, 5 and 6 in Section 3 of the IELTS Listening test and moderately difficult and difficult questions. The effect sizes were found to be generally large. Figure 34 is an appraisal calibration diagram of Groups 1 and 6 on Section 1 of the IELTS Listening test. Based on this figure, the test-takers in Group 6 were found to be mostly overconfident.

It should be noted that in Figure 34, there were some overlapping points between Groups 1 and 6 test-takers. This was because the ability levels were determined on the basis of the whole test and as noted earlier, Section 1 was the easiest section.

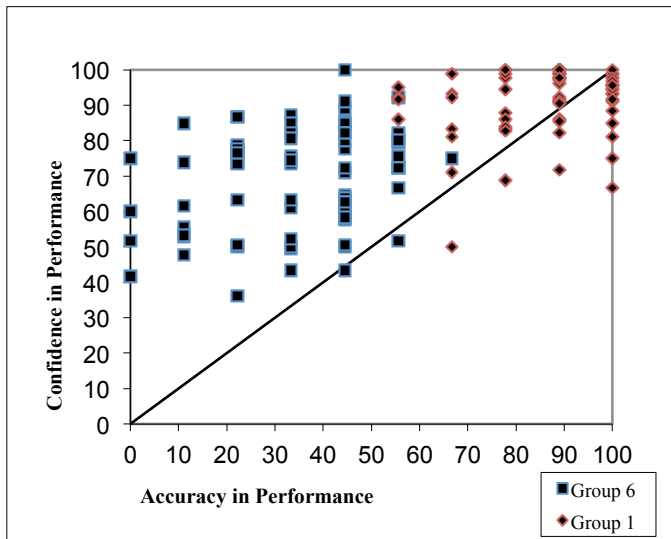


Figure 34: A calibration diagram of Groups 1 and 6 on Section 1 of the IELTS Listening test

Four test-takers (top two test-takers for Groups 1 and 6) were chosen for examination at the individual level. Tables 37 and 38 provide information about these test-takers and their performance and confidence scores. These four test-takers were all female.

Test-taker # IRT3.76 (Group 1)	Test performance (%)	SCCON (%)	Test-taker # IRT3.24 (Group 1)	Test performance (%)	SCCON (%)
Section 1	88.89	97.78	Section 1	88.89	100.00
Section 2	100.00	87.50	Section 2	80.00	100.00
Section 3	100.00	90.00	Section 3	100.00	100.00
Section 4	90.00	84.00	Section 4	100.00	100.00
EASYQ	100.00	100.00	EASYQ	100.00	100.00
MDQ	100.00	93.64	MDQ	100.00	93.64
DIFQ	100.00	87.50	DIFQ	100.00	87.50
VDIFQ	77.78	79.44	VDIFQ	77.78	79.44

SCCON = Single-case Appraisal Confidence EASYQ = Easy Questions MDQ = Moderately Difficult Questions
 DIFQ = Difficult Questions VDIFQ = Very Difficult Questions 1-4 = Number of Test Section

Table 37: Summary of two of the highest IRT ability test-takers' performance and appraisal confidence

Test-taker # IRT-2.78 (Group 6)	Test performance (%)	SCCON (%)	Test-taker # IRT-2.49 (Group 6)	Test performance (%)	SCCON (%)
Section 1	11.11	61.67	Section 1	11.11	55.56
Section 2	11.11	61.67	Section 2	0.00	0.00
Section 3	10.00	22.50	Section 3	20.00	31.50
Section 4	10.00	25.00	Section 4	10.00	31.50
EASYQ	28.57	52.14	EASYQ	28.57	70.00
MODQ	9.09	28.64	MDQ	9.09	18.18
DIFQ	8.33	22.92	DIFQ	0.00	18.75
VDIFQ	0.00	22.22	VDIFQ	11.11	23.89

EASYQ = Easy Questions MDQ = Moderately Difficult Questions DIFQ = Difficult Questions
VDIFQ = Very Difficult Questions 1-4 = Number of Test Section

Table 38: Summary of two of the lowest IRT ability test-takers' performance and confidence

Figures 35 to 38 show the calibration diagram of these four test-takers.

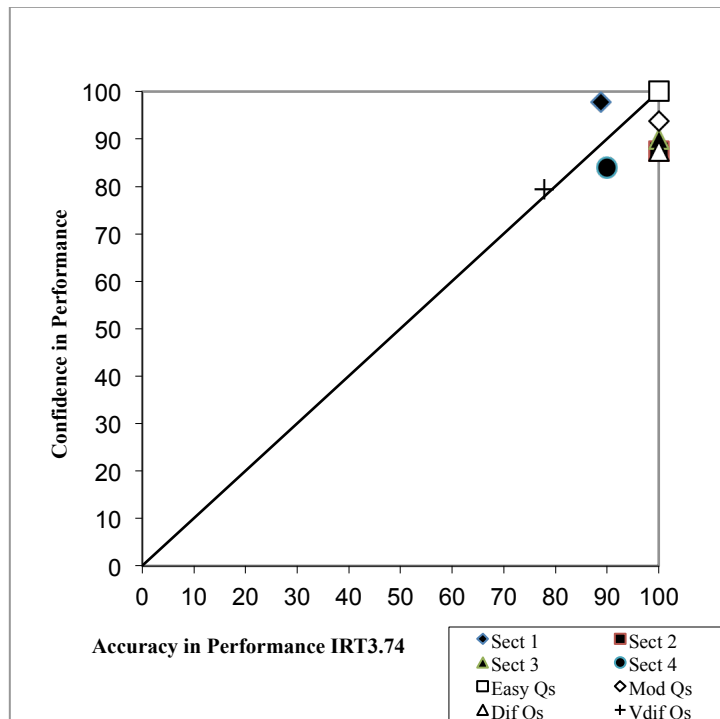


Figure 35: Appraisal calibration diagram of test-taker IRT logit 3.76 (Group 1)

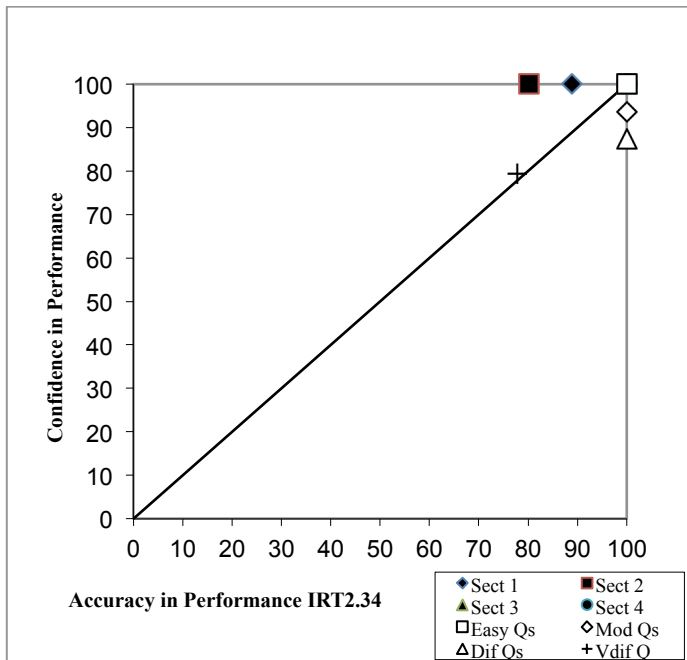


Figure 36: Appraisal calibration diagram of test-taker IRT logit 3.24 (Group 1)

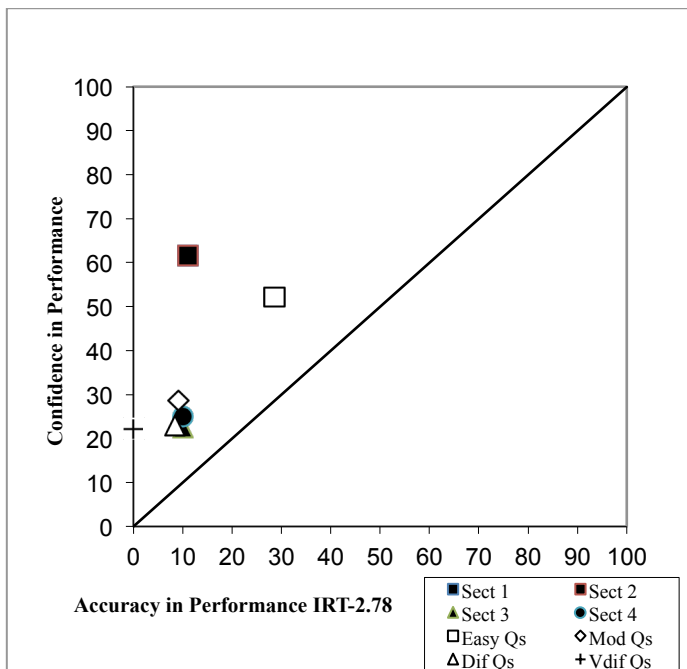


Figure 37: Appraisal calibration diagram of test-taker IRT logit -2.78 (Group 6)

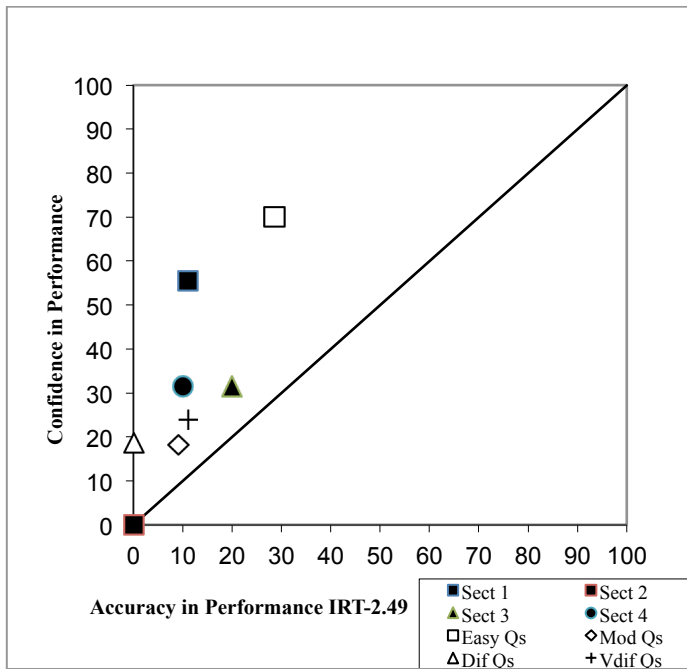


Figure 38: Appraisal calibration diagram of test-taker IRT logit -2.49 (Group 6)

In summary, test-takers in Group 1 (high-ability test-takers) were found to be quite realistic in their performance appraisals, whereas test-takers in Group 6 (low-ability test-takers) were the most unrealistic in their performance appraisals. These findings are related to the Kruger-Dunning effect, which will be discussed in Section 5.

The following section addresses the last research question, which examines the structural relationships among IELTS Listening performance, appraisal confidence judgments, and trait and state strategy use and listening test difficulty.

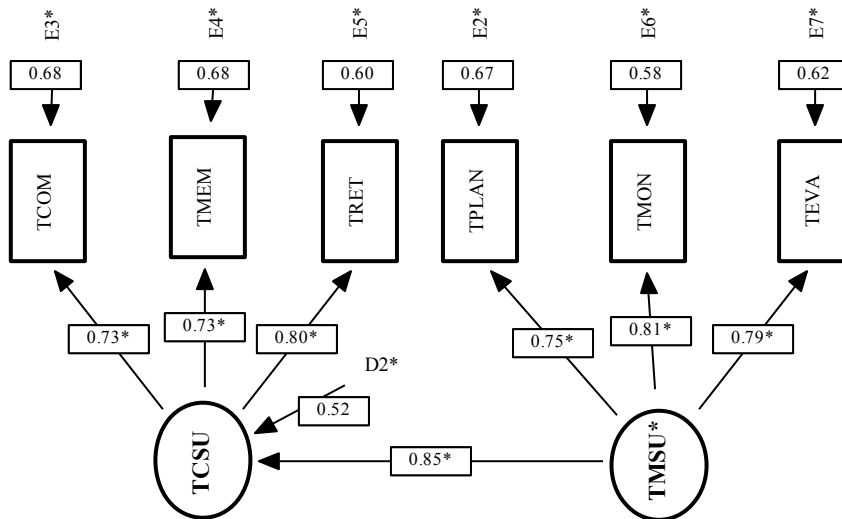
4.5 What are the structural relationships among test-takers' appraisal confidence, calibration, trait and state cognitive and metacognitive strategy use, IELTS Listening test difficulty, and IELTS Listening performance?

In regard to Research Question 1, several SEM analyses on the relationship of single-case and relative-frequency appraisal confidence to IELTS Listening test performance were presented. Research Question 5 aims to examine the structural relationship among test-takers' appraisal confidence, calibration and reported trait and state strategy use through further SEM analyses. All the SEM analyses in this section started with simple measurement models for parameter estimates, and then were developed into a more complex structural model by the addition of new measurement models and variables. This approach allows the detection of potentially incorrect model specifications. Prior to data analysis, the data were screened and tested for univariate skewness and kurtosis statistics and multivariate normality through Madia's normalised estimate (Bentler 2006). The research method section has discussed the SEM procedures, and item-level and reliability analyses of test, questionnaire and appraisal confidence data.

4.5.1 Trait cognitive and metacognitive strategy use

As discussed in the review of the literature and the research method sections, test-takers' strategic knowledge about their strategy use in IELTS Listening tests was measured through a trait strategy use questionnaire, and their strategic regulation during the actual IELTS Listening test used in the study was measured through a state strategy use questionnaire. Trait strategy use is a general tendency of test-takers to process information (e.g., using cognitive and metacognitive strategies) across contexts. State strategy use is the actual use of strategies test-takers perceive using during a particular test situation. It has been argued that both facets of test-takers' strategic competence should be measured in order to comprehensively understand the role of strategic competence on language test performance (Phakiti 2007b).

For the purpose of the present study, the trait strategy use model by Phakiti (2008a) was replicated. In this SEM model, it was hypothesised that trait metacognitive strategy use (TMSU), which was made up of trait planning, trait monitoring and trait evaluating strategy variables, had a direct influence on trait cognitive strategy use (TCSU), which comprises trait comprehending, trait memory and trait retrieval strategy variables. In the hypothesised model, a regression path was used to connect TMSU to TCSU. Figure 39 presents the SEM model of the relationship between TMSU and TCSU. The SEM standard fit indices suggested a very good model fit.



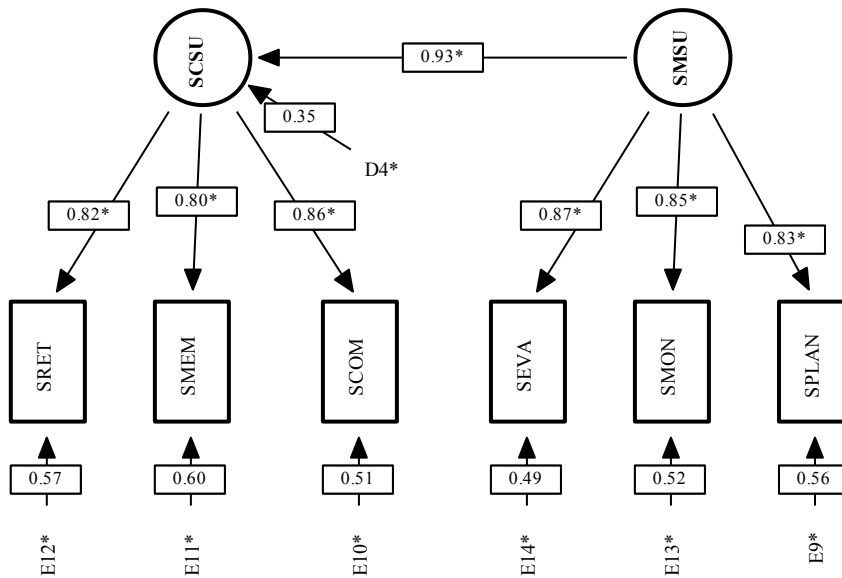
Chi-square ($\chi^2_{(6)}$) = 48.58, $p = 0.000$, CFI = 0.97, RMSEA = 0.11 (90% CI = 0.07-0.14)
 S = State T = Trait CSU = Cognitive Strategy Use MSU = Metacognitive Strategy Use
 COM = Comprehending Strategy MEM = Memory Strategy RET = Retrieval Strategy
 PLAN = Planning Strategy MON = Monitoring Strategy EVA = Evaluating Strategy

Figure 39: The SEM model of the relationship between trait MSU and trait CSU (N =376)

The SEM models in the present study are based on the *standardised solution* (Bentler, 2006). Therefore, all observed variables (Vs), latent variables (Fs), non-random errors (Es), and disturbances of prediction (Ds) were rescaled to have a variance of 1.0. The values of these variables are the same as *correlation coefficients*. Using these models, it is easy to interpret the variables in the linear structural equation system. The total common factor variances (h^2) of TMSU and TCSU were 0.61 (i.e., $0.75^2 + 0.81^2 + 0.79^2$) $\div 3 = 1.84/3$) and 0.57 (i.e., $0.73^2 + 0.73^2 + 0.80^2$) $\div 3 = 1.71/3$), respectively. Based on the h^2 values, the observed variables defined 61% and 57% of the TMSU and TCSU, respectively. The regression coefficient from TMSU to TCSU was found to be high ($\gamma = 0.85$; $R^2 = 0.72$, large effect size). The regression coefficient suggests that both trait metacognitive and trait cognitive strategy use constructs work closely together. This SEM model implies that the nature of test-takers' strategic knowledge of how they generally use metacognitive strategies to regulate their cognitive strategies during IELTS Listening tests informs their trait metacognitive strategy use. The next section presents how this SEM model statistically holds for state metacognitive and cognitive strategy use.

4.5.2 State cognitive and metacognitive strategy use

State metacognitive strategy use (SMSU) and state cognitive strategy use (SCSU) were measured at the end of the test. The hypothesised model was similar to that in Figure 40, which presents the SEM model of the relationship between SMSU and SCSU. The SEM standard fit indices indicated a very good model fit.



Chi-square ($\chi^2_{(6)} = 33.98, p = 0.000, CFI = 0.99, RMSEA = 0.10 (90\% CI = 0.06-0.13)$
 S = State CSU = Cognitive Strategy Use MSU = Metacognitive Strategy Use COM = Comprehending Strategy
 MEM = Memory Strategy RET = Retrieval Strategy PLAN = Planning Strategy MON = Monitoring Strategy
 EVA = Evaluating Strategy

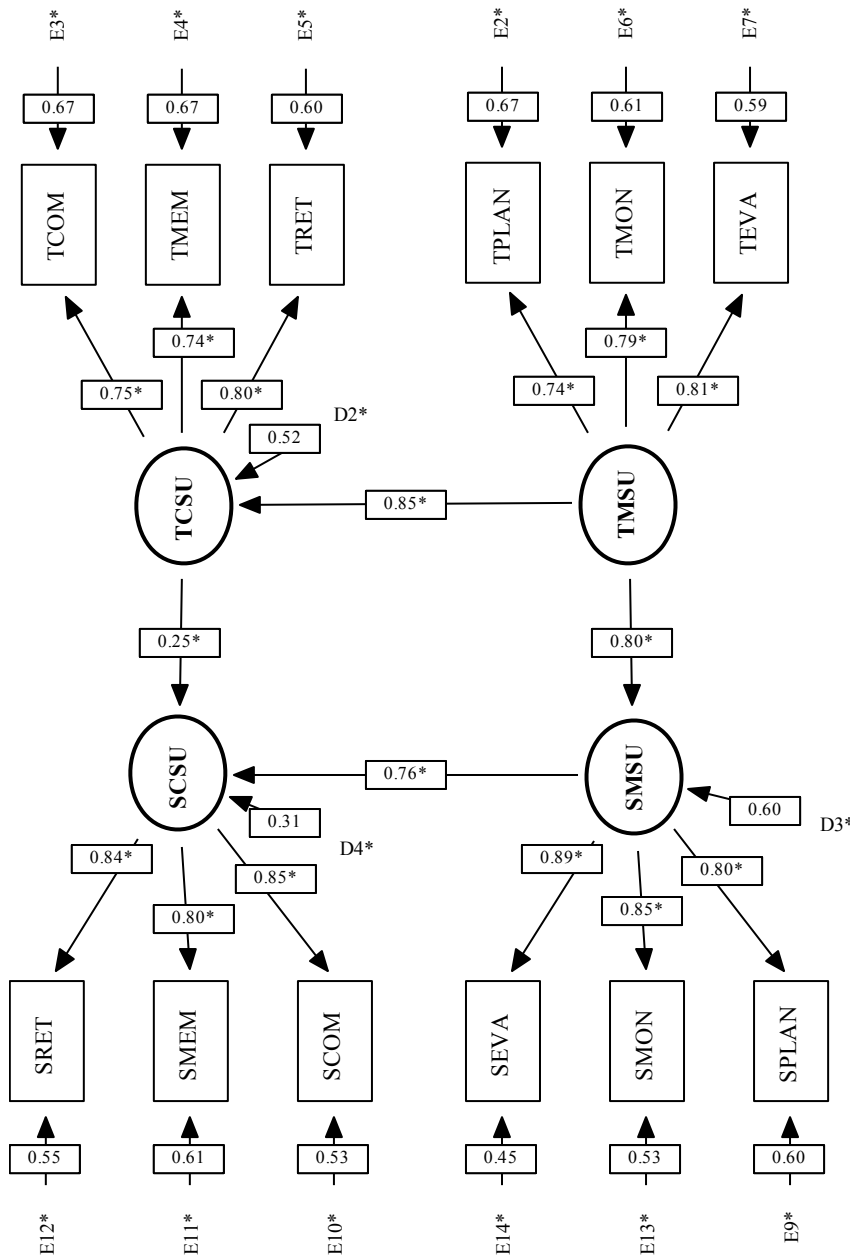
Figure 40: The SEM model of the relationship between state MSU and state CSU (N =376)

The total common factor variances (h^2) of SMSU and SCSU were 0.72 (i.e., $0.87^2 + 0.85^2 + 0.83^2$) $\div 3 = 2.17 \div 3$) and 0.68 (i.e., $0.82^2 + 0.80^2 + 0.86^2$) $\div 3 = 2.05 \div 3$), respectively. The total common factor variances in state strategy use were found to be larger than those for trait strategy use. The regression coefficient from SMSU to SCSU was found to be high ($\gamma = 0.93; R^2 = 0.86$, large effect size). This regression coefficient suggests that during the IELTS Listening test, state metacognitive strategy use closely regulated cognitive strategy use. This strong relationship might be partly explained by the fact that test-takers were asked to rate their appraisal confidence in their test performance, which in turn might have triggered their metacognitive awareness of their cognitive processes during test-taking.

The next section reports on how trait MSU and CSU may influence state MSU and CSU.

4.5.3 The relationships between trait and state MSU and CSU

To examine the inter-relationship between trait MSU and CSU and state MSU and CSU, a regression path from TMSU to SMSU and a regression path from TCSU to SCSU were added to the model. Figure 41 presents the SEM model of the relationship between TMSU and SMSU and TCSU and SCSU. It should be noted that there were two correlation coefficients between two pairs of errors that were corrected in the model re-specifications: E9 and E2 = 0.40 and E13 and E6 = 0.42.



Chi-square ($\chi^2_{(44)} = 201.49, p = 0.000, CFI = 0.97, RMSEA = 0.09$ (90% CI = 0.08-0.11)
 S = State T = Trait CSU = Cognitive Strategy Use MSU = Metacognitive Strategy Use
 COM = Comprehending Strategy MEM = Memory Strategy RET = Retrieval Strategy
 PLAN = Planning Strategy MON = Monitoring Strategy EVA = Evaluating Strategy

Figure 41: The SEM model of the relationship between trait and state MSU and CSU (N =376)

According to Figure 41, first, TMSU had a strong, direct positive effect on SMSU (regression coefficient (γ) = 0.80; $R^2 = 0.64$, large effect size). This means that TMSU (i.e., strategic knowledge of metacognitive strategy use) accounted for 64% of the SMSU (i.e., strategic regulation of metacognitive strategy use) variance. This finding is similar those found in Phakiti (2008a) and Bi (2014). This coefficient can be interpreted as follows: test-takers with a strong perception of their use of a set of metacognitive strategies in IELTS Listening tests in general also report high use of metacognitive strategies in a given context. However, the values of the R^2 suggest that TMSU does not account for all the ways SMSU is used in a specific setting. Specific characteristics or conditions within a testing context partly determine the degree of state metacognitive strategy use.

Second, TCSU was found to have a weak influence on SCSU ($\beta = 0.25$, $R^2 = 0.06$; small effect size). This finding was similar to previous research (e.g., Bi, 2014; Phakiti 2008a). In a reading test context, Phakiti (2008a) found regression coefficients of 0.22 (Time 1) and 0.25 (Time 2). Theoretically, unlike TMSU and SMSU, TCSU does not have an executive function, so what test-takers think they generally do cognitively in the IELTS Listening tests may not necessarily have a significant impact on what they actually do in a specific IELTS Listening test. In Phakiti's (2008) longitudinal model, it was found that TCSU had a much lower stability over time, compared to TMSU. Furthermore, it might well be that the strong regression coefficient from TMSU to TCSU might be responsible for the regression coefficient from TCSU to SCSU.

Third, based on Figure 41, SMSU had a strong positive influence on SCSU ($\beta = 0.76$; $R^2 = 0.58$; large effect size). That is, SMSU and SCSU had 58% shared variance. The findings imply that while students engage in actual information processing during the IELTS Listening test, online metacognitive strategies work hand-in-hand with online cognitive strategies to accomplish the test tasks. It should be noted that the strength of the influence of SMSU on SCSU was reduced from 0.93 (Figure 40) to 0.76 (Figure 41). This is not surprising because more variables such as TMSU and TCSU have been added to this model, which resulted in more information for parameter estimates. It should be noted that the relationship between SMSU and SCSU is similar to that between TMSU and TCSU.

In both trait and state strategy use, the present findings lend empirical support to the executive functions of MSU in both the long-term and working memories of human information processing.

4.5.4 The relationships among trait and state MSU and CSU and appraisal confidence

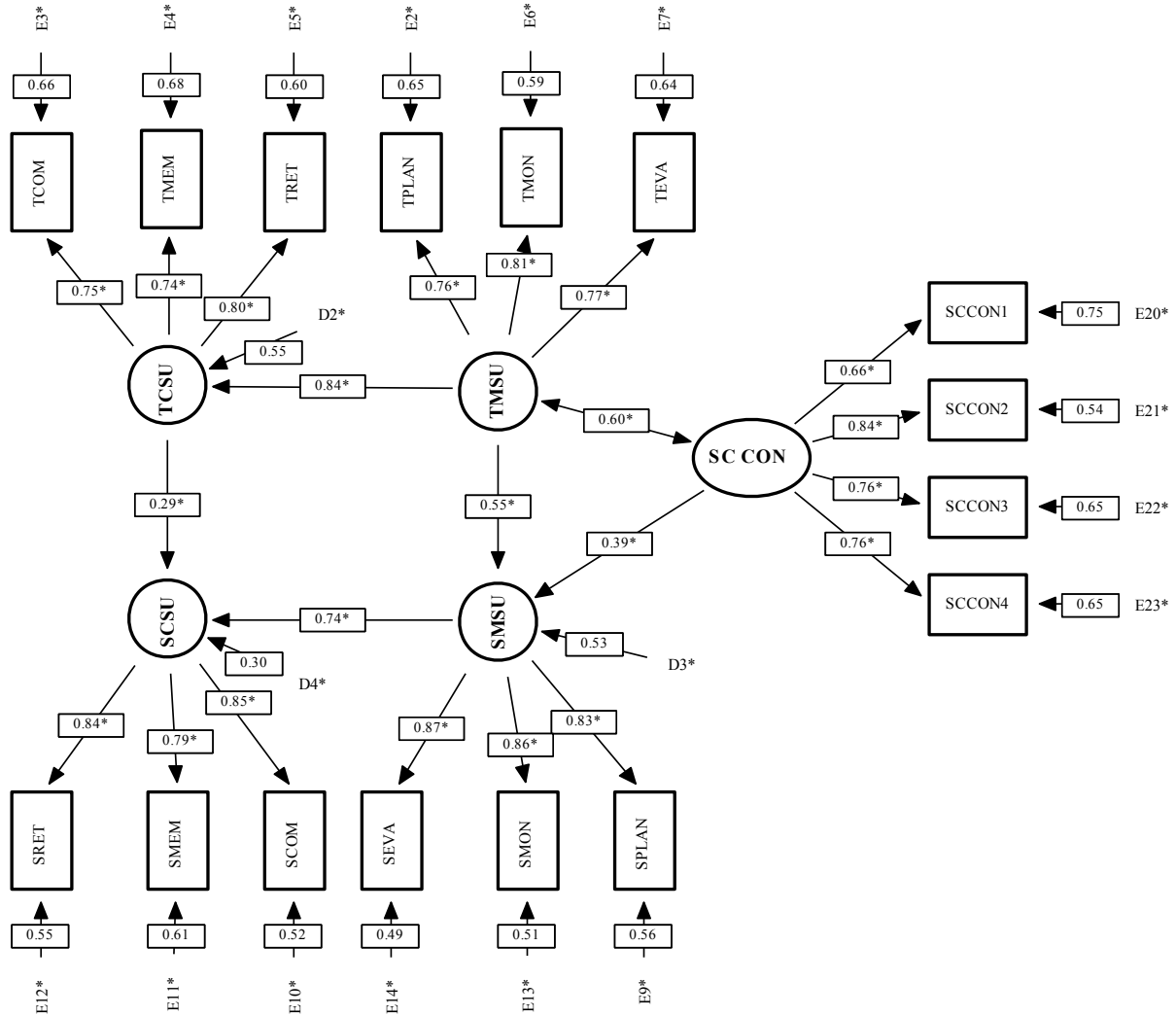
It has been hypothesised that both single-case appraisal confidence and relative-frequency appraisal confidence judgments should be strongly related to trait and state MSU and CSU because accurate appraisal confidence informs test-takers how to best complete the given test tasks at hand. In the present study, since the trait strategy use questionnaire was answered before the IELTS Listening test, trait strategy use should be considered an independent factor. Test-takers' performance appraisals as measured by single-case appraisal confidence should result in strategic behaviours to tackle the given test tasks.

Single-case appraisal confidence should be considered a state independent factor. However, relative-frequency appraisal confidence should be considered a state dependent factor because it was measured after single-case appraisal confidence in each test section.

Figure 42 presents the SEM model that tested the relationships of single-case appraisal confidence to trait and state strategy use. The SEM model had a very good model fit.

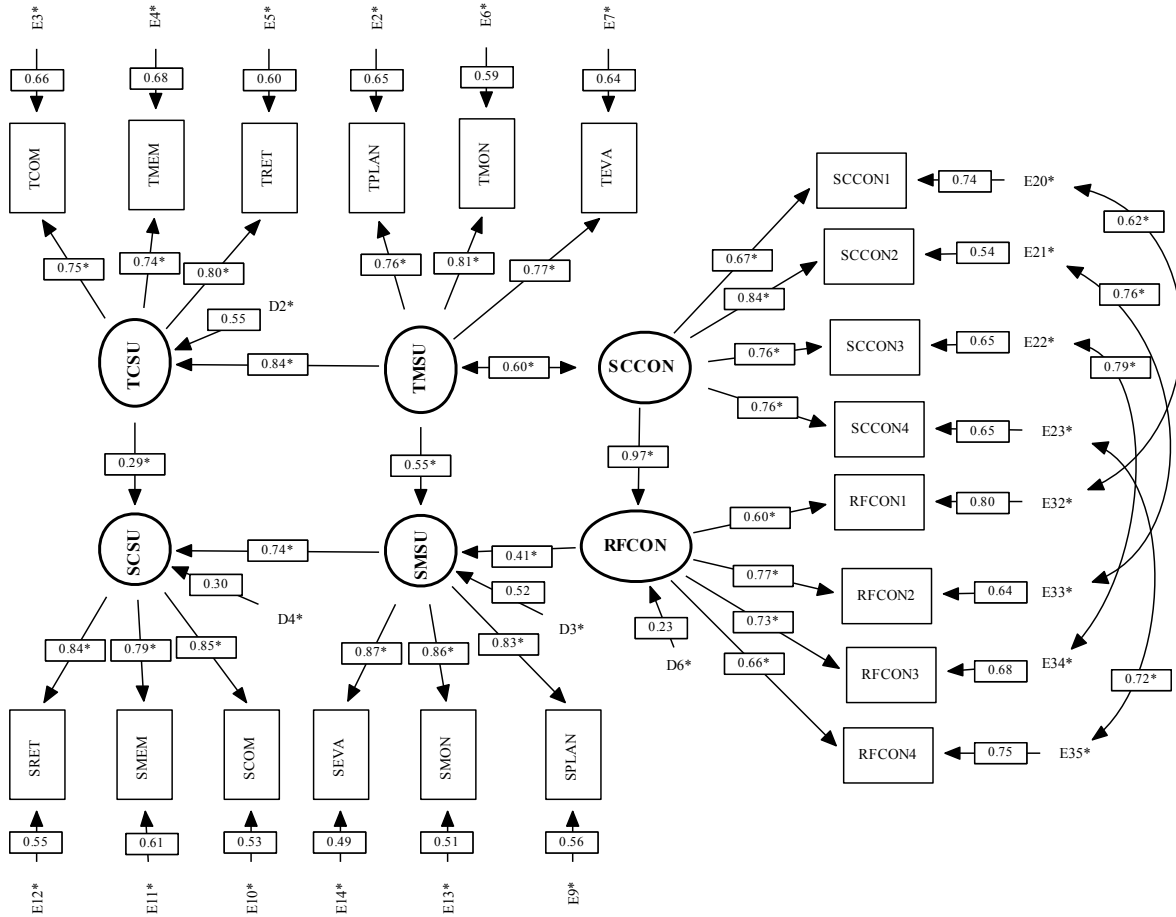
When single-case appraisal confidence was added to the SEM model presented in Figure 42, the following observations can be made. First, the correlation coefficient between single-case appraisal confidence (SCCON) and trait metacognitive strategy use (TMSU) was positive and moderate (0.60; $R^2 = 0.36$; medium effect size). The coefficient might indicate that there was a common monitoring factor underlying SCCON and TMSU. Second, single-case confidence was found to directly and positively influence state metacognitive strategy (SMSU). The regression coefficient was 0.39 ($R^2 = 0.15$; small effect size). It should be recalled that in relation to Research Questions 1 to 4, it was found that test-takers were miscalibrated and tended to be overconfident in their performance. The small regression coefficient found in Figure 42 was not surprising because a lack of performance appraisal accuracy could influence cognitive processes that did not address a test task at hand. This finding shows an empirical connection between poor performance appraisals and metacognitive strategy use.

In order to investigate further, relative-frequency appraisal confidence as presented in Figure 15 was added to the model. In this SEM model, relative-frequency appraisal confidence was treated as a dependent factor. As discussed above, single-case and relative-frequency appraisal confidence judgments appeared to be closely associated with each other. Figure 43 presents the SEM model that includes the relative-frequency appraisal confidence. It should be noted that the choice to construct a direct regression path from relative-frequency appraisal confidence to SMSU was found to be more meaningful than the choice to have both regression paths from single-case confidence and relative-frequency appraisal confidence to SMSU. This SEM model had a very statistical good fit. The correlation coefficients between the following three error pairs that do not appear in the model should be noted: E9 and E2 = 0.32; E13 and E6 = 0.36; E14 and E7 = 0.47.



Chi-square ($\chi^2_{(90)}$) = 264.91, p = 0.000, CFI = 0.98, RMSEA = 0.07 (90% CI = 0.06-0.08)
 S = State T = Trait CSU = Cognitive Strategy Use MSU = Metacognitive Strategy Use
 COM = Comprehending Strategy MEM = Memory Strategy RET = Retrieval Strategy
 PLAN = Planning Strategy MON = Monitoring Strategy EVA = Evaluating Strategy
 SCCON= Single-case Appraisal confidence

Figure 42: The SEM model of the relationship of single-case appraisal confidence to trait and state MSU and CSU (N =376)



Chi-square ($\chi^2_{(150)}$) = 341.42, $p = 0.000$, CFI = 0.99, RMSEA = 0.06 (90% CI = 0.05-0.07)
 S = State T = Trait CSU = Cognitive Strategy Use MSU = Metacognitive Strategy Use
 COM = Comprehending Strategy MEM = Memory Strategy RET = Retrieval Strategy
 PLAN = Planning Strategy MON = Monitoring Strategy EVA = Evaluating Strategy
 SCCON = Single-case Appraisal Confidence RF = Relative-frequency Appraisal Confidence

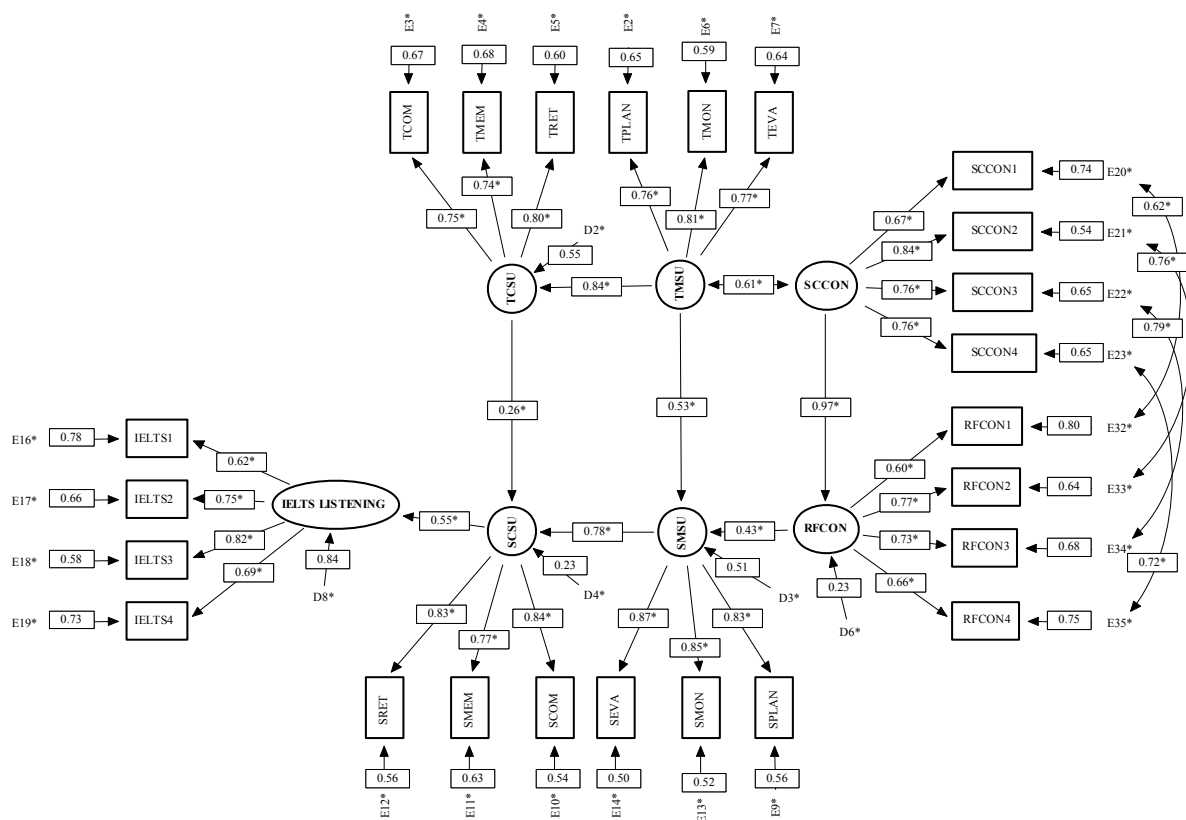
Figure 43: The SEM model of the relationship of single-case and relative-frequency appraisal confidence to trait and state MSU and CSU (N=376)

According to Figure 43, it was found that relative-frequency appraisal confidence was directly influenced by single-case appraisal confidence ($\gamma = 0.97$; $R^2 = 0.94$; large effect size), which in turn directly contributed to SMSU. Single-case appraisal confidence hence indirectly influenced SMSU via TMSU ($\gamma = 0.33$; $R^2 = 0.11$; small effect size) and RFCON ($\gamma = 0.40$; $R^2 = 0.16$; small effect size). The addition of RFCON to this SEM model is significant in that an overall performance appraisal is included to explain the variances in this human information processing model. It was also found that the parameter estimates among latent variables had become quite stable when compared to the preceding SEM models.

The next section presents the SEM model that includes the IELTS Listening test performance.

4.5.5 Trait and state cognitive strategy use, appraisal confidence, and IELTS Listening test performance

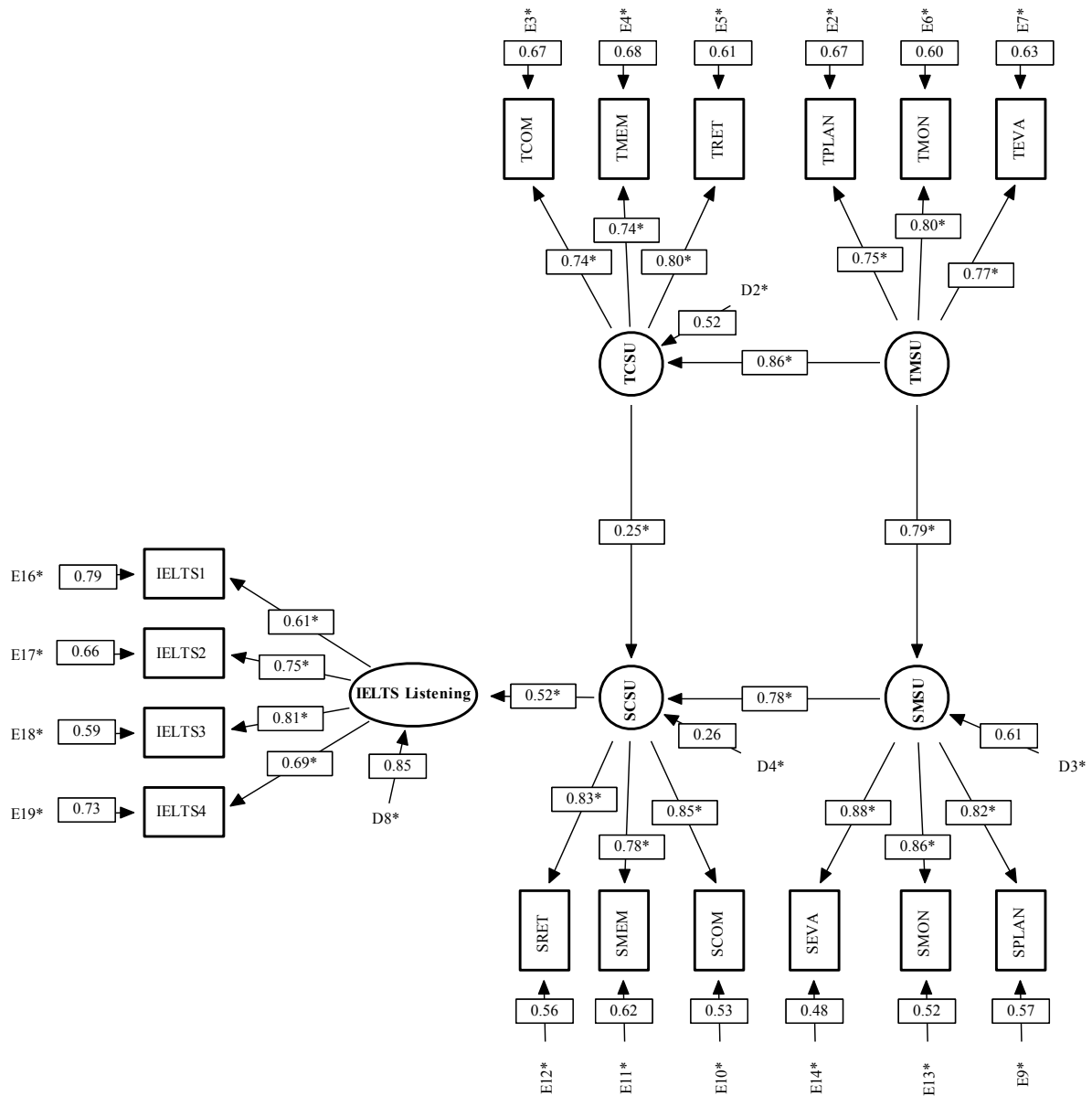
Phakiti (2008a, 2016) has theoretically and empirically illustrated the direct and indirect relationships between individual facets of strategy use (i.e., trait and state) and language test performance. Based on Phakiti's (2007b) human information processing model (presented in the literature review section), trait strategy use is hypothesised to be indirectly related to a specific language test performance via state strategy use or other cognitive and affective aspects within a specific language use/test context. Therefore, the SEM model in Figure 43 was extended by adding a regression path from SCSU to IELTS Listening performance. Figure 44 presents this SEM model, which had a very good fit.



Chi-square ($\chi^2_{(230)} = 847.35, p = 0.000, CFI = 0.97, RMSEA = 0.08$ (90% CI = 0.07-0.09)
 S = State T = Trait CSU = Cognitive Strategy Use MSU = Metacognitive Strategy Use
 COM = Comprehending Strategy MEM = Memory Strategy RET = Retrieval Strategy
 PLAN = Planning Strategy MON = Monitoring Strategy EVA = Evaluating Strategy
 SCON = Single-case Appraisal Confidence RFCON = Relative-frequency Appraisal Confidence

Figure 44: SEM model of trait and state cognitive strategy use, appraisal confidence, and IELTS test performance (N =376)

According to Figure 44, SCSU had a direct positive influence on IELTS Listening test performance. The regression coefficient was 0.55 ($R^2 = 0.30$; medium effect size). This finding was similar to that found in Phakiti (2008a), in which it was found that SCSU explained 30% of lexico-grammatical performance variance in Phase 1 of the study. As previous research did not include a confidence factor in the SEM models adopted, it was decided to find out how much SCSU could explain the IELTS Listening test results without the confidence factors in the model. Figure 45 presents the SEM model that excluded the confidence factors in the parameter estimates. This SEM model also had an excellent model fit.



Chi-square ($\chi^2_{(91)} = 283.35, p = 0.000, CFI = 0.98, RMSEA = 0.08$ (90% CI = 0.07-0.09)

S = State T = Trait CSU = Cognitive Strategy Use MSU = Metacognitive Strategy Use
 COM = Comprehending Strategy MEM = Memory Strategy RET = Retrieval Strategy
 PLAN = Planning Strategy MON = Monitoring Strategy EVA = Evaluating Strategy

Figure 45: SEM model of trait and state cognitive strategy use and IELTS test performance (N =376)

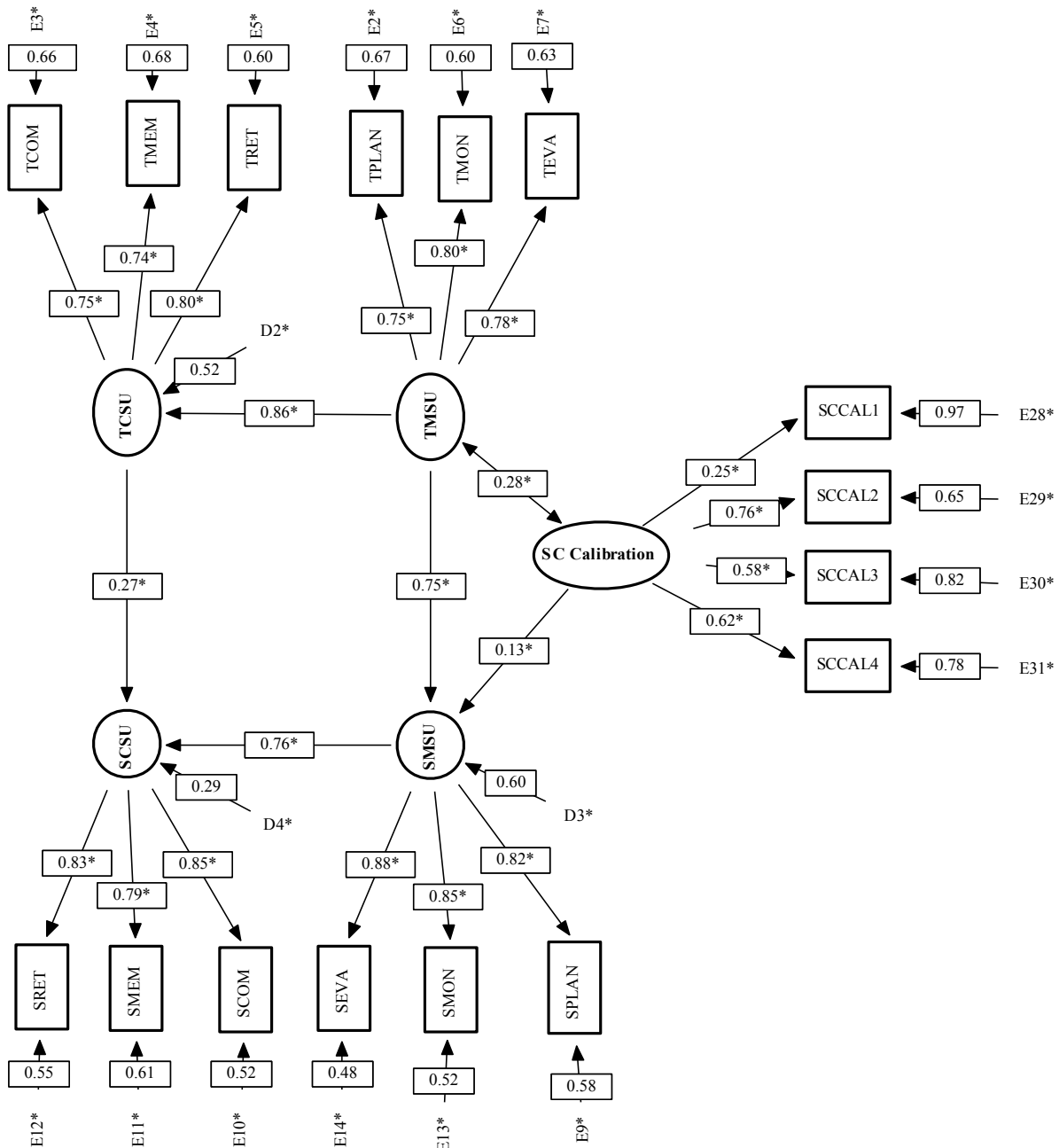
It was found that the relationship between TMSU and SMSU was back to 0.79 (similar to that of the earlier model presented in Figure 41). It can be seen that when single-case and relative-frequency confidence factors were not considered, the regression coefficient was much larger (0.79 versus 0.55). However, while the regression coefficient from SMSU to SCSU remained stable, the regression coefficient from SCSU to IELTS Listening was reduced to 0.52 ($R^2 = 0.27$; medium effect size). The change in the regression coefficient was not dramatic. On the basis of the two SEM models (Figures 43 and 44), it might be inferred that when confidence judgment variables are factored into test-takers' cognitive network, IELTS Listening test performance was more fully explained.

The next section presents the SEM model that examines how test-takers' calibration was related to TMSU and SMSU.

4.5.6 Trait and state MSU and CSU and appraisal calibration

In the section on Research Question 1, the CFA model of test-takers' single-case calibration in the IELTS Listening test was presented (Figure 19a). In order to examine the relationships between single-case calibration and trait and state strategy use, Figure 41 was connected to Figure 19a. It was hypothesised that single-case calibration and trait metacognitive strategy use were independent factors that influenced other factors. Hence, a correlation coefficient path between TMSU and calibration was connected. Calibration was hypothesised to directly influence SMSU, which plays an executive role. Figure 50 presents the SEM model that examined how calibration could be connected to trait and state strategy use. It should be noted that the correlation coefficients between the following three error pairs do not appear in the model: E9 and E2 = 0.35; E13 and E6 = 0.39; E14 and E7 = 0.45.

According to Figure 46, the correlation coefficient between calibration and TMSU was 0.28 ($R^2 = 0.08$; small effect size) and the regression coefficient from calibration to SMSU was 0.13 ($R^2 = 0.02$; small effect size). Both coefficients were statistically significant at 0.05. Phakiti (2016) found that the regression coefficient from calibration to SMSU was 0.33 ($R^2 = 0.11$; small effect size). In Phakiti, TMSU was not considered in the modeling used because a trait strategy use questionnaire was not used. The present findings in the small coefficients might imply that when test-takers are not well-calibrated in their performance appraisals, they cannot employ metacognitive strategies effectively during IELTS test-taking. The next section presents the final SEM model in which all variables including trait and state listening difficulty factors were included.



Chi-square ($\chi^2_{(90)}$) = 247.27, $p = 0.000$, CFI = 0.98, RMSEA = 0.07 (90% CI = 0.06-0.08)

S = State T = Trait CALS = Calibration in Section

CSU = Cognitive Strategy Use MSU = Metacognitive Strategy Use

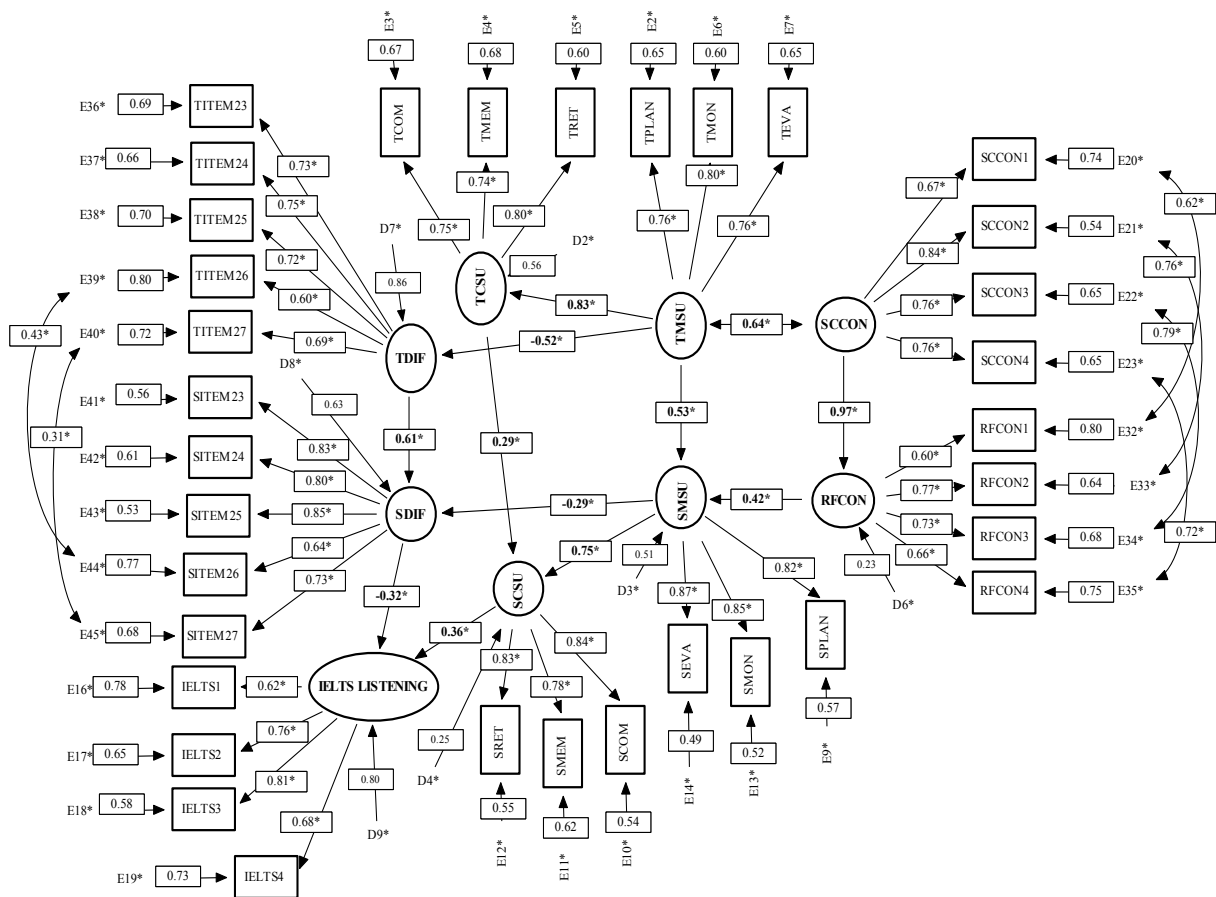
COM = Comprehending Strategy MEM = Memory Strategy RET = Retrieval Strategy

PLAN = Planning Strategy MON = Monitoring Strategy EVA = Evaluating Strategy

Figure 46: SEM model of trait and state cognitive strategy use and appraisal calibration (N =376)

4.5.7 Trait and state cognitive strategy use, appraisal confidence, trait and state IELTS Listening test difficulty, and IELTS test performance

In this section, all the variables were simultaneously tested in the SEM analysis. The SEM model here was an extension of the SEM model in Figure 45, but the trait and state IELTS Listening difficulty factors were added. Prior to this addition, CFAs of trait and state IELTS Listening difficulty variables were performed and the model fit was found to be very good. A subsequent structural model between trait and state IELTS Listening difficulty was tested. A regression path was used to connect trait IELTS Listening difficulty to state IELTS Listening difficulty because it was hypothesised that a trait influences a state. The SEM model had excellent model fit (CFI = 0.99, RMSEA = 0.06) and the regression coefficient was 0.74 ($R^2 = 0.55$; large effect size). To connect this trait and state IELTS Listening model to the SEM model in Figure 47, it was further hypothesised that state IELTS Listening difficulty would have a direct, negative impact on IELTS Listening test performance (i.e., a regression path was added from SDIF to IELTS Listening). State metacognitive strategy use (SMSU) was hypothesised to directly influence state IELTS Listening test performance since metacognitive strategies have a monitoring function that enables the detection of any affective difficulty (i.e., a regression path was added from SMSU to SDIF). Finally, trait metacognitive strategy use (TMSU) was hypothesised to directly influence trait IELTS Listening difficulty (TDIF).



Chi-square ($\chi^2_{(497)}$) = 1219.65, $p = 0.000$, CFI = 0.98, RMSEA = 0.06 (90% CI = 0.05-0.07)

S = State T = Trait

CSU = Cognitive Strategy Use MSU = Metacognitive Strategy Use

COM = Comprehending Strategy MEM = Memory Strategy RET = Retrieval Strategy

PLAN = Planning Strategy MON = Monitoring Strategy EVA = Evaluating Strategy

SCCON = Single-case Appraisal Confidence RFCON = Relative-frequency Appraisal Confidence

Figure 47: The SEM model of trait and state cognitive strategy use, appraisal confidence, trait and state IELTS Listening test difficulty, and IELTS test performance (N =376)

Figure 47 presents the network of trait and state cognitive strategy use, appraisal confidence, IELTS test performance, and perceived IELTS Listening test difficulty. It should be noted that the statistically significant correlations among independent variables (errors) that were not included in this figure were: E9 and E2 (0.31), E13 and E6 (0.37) and E14 and E7 (0.48). The data fitted the hypothesised model very well (Bentler-Bonett NFI (= 0.96), Bentler-Bonett NNFI = 0.97; CFI = 0.98), and RMSEA = 0.06 [90% Confidence Interval = 0.05, 0.07]).

First, it was found that the regression coefficient between TDIF and SDIF was 0.65 ($R^2 = 0.42$, medium effect size) and SDIF in turn was found to negatively influence IELTS Listening test performance ($\beta = -0.32$, $R^2 = 0.10$; small effect size). Second, it was found that TMSU had a direct negative influence on TDIF ($\gamma = -0.52$; $R^2 = 0.27$, medium effect size), whereas the regression coefficient of SMSU to SDIF was found to be -0.29 ($R^2 = 0.08$, small effect size). This negative regression suggests that more use of state metacognitive strategies could reduce generally perceived listening difficulty in IELTS Listening tests. To find out the effect of TDIF and SDIF on IELTS Listening test performance, the regression path from SMSU to SDIF was dropped and the model was retested.

It was found that while the regression coefficient from SDIF to IELTS Listening test performance did not change (further discussed below), the regression coefficient from TDIF to SDIF changed to 0.76. On this basis, it could be inferred that the use of metacognitive strategies during an IELTS Listening test could filter the impact of trait IELTS Listening difficulty on state IELTS Listening difficulty. TMSU and SMSU might well be operating together to reduce the negative influence of trait and state IELTS Listening difficulty on IELTS Listening test performance.

Finally, by simultaneously modeling all the variables in a single SEM analysis, it was found that the regression coefficient from SCSU to IELTS Listening test performance was reduced to 0.36 ($R^2 = 0.13$; small effect size). In spite of additional variables considered in the SEM analysis, SCSU (which was regulated through trait and state metacognitive strategy use and single-case and relative-frequency appraisals) remained to have a significant impact on IELTS Listening performance.

Based on Figure 47, the decomposition of the IELTS Listening test performance (direct and indirect effects) can be summarised as follows:

$$\begin{aligned} \mathbf{F9 (IELTS Listening)} &= 0.102 \mathbf{F2 (TCSU)} + 0.363 \\ &\mathbf{F3 (SMSU)} + 0.358 \mathbf{F4 (SCSU)} + 0.153 \mathbf{F6} \\ &\mathbf{(RFCON)} - 0.198 \mathbf{F7 (TDIF)} - 0.322 \mathbf{F8 (SDIF)} + \\ &0.380 \mathbf{F1 (TMSU)} + 0.149 \mathbf{F5 (SCCON)} + 0.057 \mathbf{D2} \\ &+ 0.186 \mathbf{D3} + 0.091 \mathbf{D4} + 0.035 \mathbf{D6} - 0.169 \mathbf{D7} - \\ &0.203 \mathbf{D8} + 0.802 \mathbf{D9} \end{aligned}$$

In summary, several SEM models have suggested that IELTS Listening test performance can be affected by several factors including performance appraisal accuracy (calibration), trait and state cognitive and metacognitive strategy use, and generally perceived and situation-specific listening difficulty in IELTS Listening. It was found that the relationships between performance appraisals and other strategic competence facets can be highly complex. Several SEM models indicate the multiple interactions among confidence, trait and state strategy use and listening difficulty. When single-case and relative-frequency appraisal confidence judgments are accurate (i.e., highly calibrated), cognitive and metacognitive strategy use may have more power to positively influence IELTS Listening test performance. It was also found that metacognitive strategy use has a positive impact on IELTS test performance by filtering test-takers' perceived difficulty out during information processing.

5 DISCUSSION

The present study has primarily focused on the appraisal calibration and strategy use aspects of strategic competence – non-linguistic factors that play a critical role in determining test-takers' academic listening levels. The importance of strategic competence has been recognised throughout the history of communicative competence theories and has been empirically supported by numerous studies in language learning and testing. However, only recently has IELTS research begun to examine cognitive factors such as strategies that influence IELTS Listening test performance (e.g., Badger & Yan 2009; Field 2009; Winke & Lim 2014).

Unlike previous IELTS Listening studies, this study examines test-takers' performance appraisals and the extent to which they are calibrated with actual listening test performance, as well as how appraisals may be linked to metacognitive strategy use. The study has been largely informed by substantive theories and research in cognitive and educational psychology of human metacognitive processes, decision-making processes, performance judgments and monitoring accuracy. The study postulates that if test-takers do not know how well they are performing in a given test, they cannot properly and efficiently complete the test tasks involved in that test. Some key variables (e.g., ability levels, gender and task difficulty) that may determine the nature of test-takers' calibration have been examined.

As discussed in the literature review section, L2 listening is a highly complex cognitive process and includes a perception phase, a parsing phase, and an utilisation phase (e.g., Field 2013; Vandergrift & Goh 2012). Successful L2 listening requires an interactive combination of bottom-up and top-down processing. It is well known that success in L2 listening can be determined by both listener factors (e.g., linguistic knowledge, world knowledge, strategic ability, motivation, L1 background) and contextual factors (e.g., types of audio text, test task complexity and the speed of text delivery). L2 listening requires both automatic and control processes because automaticity, such as automatic word recognition and automatic monitoring, can help ease the information processing demand in the

working memory, while conscious control processing, such as purposeful monitoring, evaluating and decision-making can help listeners move beyond decoding the text to integrating new information with existing information. Field (2009, 2013) and Vandergrift and Goh (2012) point out that low-ability listeners are likely to struggle with issues of listening fluency and meaning construction as they are likely to have to deal with basic linguistic features and cognitive overloading, whereas high-ability listeners can go beyond such a linguistic threshold to construct meaning and monitor comprehension.

In L2 listening models, the roles of self-monitoring and metacognition are well recognised (e.g., Buck 2001; Field 2013; Vandergrift & Goh 2012). It is important for L2 listeners to accurately monitor the incoming audio text in order to select certain information to process, as well as to integrate and use it carry out a given task. Monitoring accuracy is particularly significant in a high-stakes test as a minor mistake (e.g., a misspelling of an answer) can result in a poor test score.

In line with such models of L2 listening, the current revised model of communicative language ability by Bachman and Palmer (2010) underlines the important role of performance appraisals as part of strategic competence during test taking and language use. Performance appraisals are crucial because test-takers need to know that their responses are correct or appropriate by using some criteria to judge their performance. Performance appraisals are investigated through appraisal confidence judgments. Based on Efklides (2011), appraisal confidence is a form of task-specific metacognitive experience that is vital for good performance. However, little is empirically known about monitoring accuracy or the accuracy of performance appraisals in L2 listening in general and L2 listening tests in particular.

To examine test-takers' performance appraisals and strategic processing in the IELTS Listening test, the present study asked test-takers to:

1. report on their general perceived cognitive and metacognitive strategy use, and level of IELTS Listening difficulty (this is a trait-like measure)
2. complete IELTS Listening tasks and rate their single-case and relative-frequency appraisal confidence
3. report on their perceived cognitive and metacognitive strategy use and level of difficulty during the listening test.

The data were then analysed to examine their calibration, strategy use and level of listening difficulty. The following section summarises the key findings for each research question and discusses them in light of relevant theories and empirical research.

5.1 Discussion of the findings

5.1.1 Research question 1: The nature of test-takers' appraisal confidence and calibration in IELTS Listening test tasks

First, according to Table 18, test-takers did quite well in the IELTS Listening Section 1 (performance above 63% on average) and Section 3 (58% on average). They did not perform as well in Section 2 (46%) or Section 4 (32%). It was clear that test-takers did better in conversational–transactional listening tasks (Sections 1 and 3) than in monologue tasks (Sections 2 and 4). The *t*-tests indicate significant differences in test scores across the four test sections (see Table 19). The difference in the mean scores for Sections 1 and 3 was 31%, whereas the difference in the mean scores between Sections 3 and 4 was 26%. The magnitudes of the effect sizes were medium to large.

Second, test-takers' appraisal confidence judgments (both single-case and relative-frequency) were found to be consistently higher than their associated test performances (see Table 18). Before examining whether test-takers were calibrated in their performance appraisals, the differences between single-case and relative-frequency appraisal confidence were examined. It was found that in three out of four test sections, single-case appraisal confidence was significantly higher than relative-frequency appraisal confidence. However, the effect sizes were small. The relationships among single-case appraisal confidence judgments for each test section, as well as between latent single-case relative-frequency confidence judgments, suggest that there is a *general appraisal confidence factor* because appraisal confidence in different test sections significantly and positively correlated with

one another. Calibration research in cognitive and educational psychology has found that appraisal confidence judgments in various tests were correlated with one another (see Kleitman & Stankov 2007; Stankov, Lee, Luo & Hogan 2012). Stankov et al. (2012) pointed out that this general confidence factor is analogous to the *general cognitive ability factor 'g'*.

Third, it was found that test-takers were overconfident in their test performance for all four test sections. Based on single-case appraisal confidence, test-takers were found to be nearly 20% overconfident in Section 1 and 10% overconfident in Section 4. However, they were found to be most realistic in Section 2 (6.5% overconfident). The paired-samples *t*-tests (Table 22) suggest that appraisal confidence was significantly higher than test performance (the magnitudes of the differences were medium to large). In this study, test-takers would be considered realistic when their calibration was within $\pm 5\%$. When examining the frequencies of their calibration scores within $\pm 5\%$, it was found that only a small percentage of test-takers were calibrated (i.e., Section 1 = 54 (14%), Section 2 = 88 (23%), Section 3 = 96 (26%) and Section 4 = 76 (20%). On the basis of these findings, 74% (Section 3) to 86% (Section 1) of the test-takers were miscalibrated (either over- or underconfident).

According to Stankov and Lee (2014a), the majority of individuals (approximately 70%) were overconfident in their abilities. Stankov and Lee suggested that the correlations between accuracy, confidence and calibration scores may indicate whether individuals' overconfidence was an outcome of *cognitive bias* or *motivational bias*. If the calibration scores are correlated more highly with accuracy than with confidence, then cognitive bias may be present. If the correlation is higher with confidence than with accuracy, then motivational bias may be present. In order to consider this question of bias, Pearson-Product-Moment correlations were computed and are reported in Table 39.

	IELTS scores	Confidence score
Section 1	-0.75** ($R^2 = 0.56$)	0.15** ($R^2 = 0.02$)
Section 2	-0.33** ($R^2 = 0.11$)	0.54** ($R^2 = 0.29$)
Section 3	-0.37** ($R^2 = 0.14$)	0.27** ($R^2 = 0.07$)
Section 4	-0.31** ($R^2 = 0.10$)	0.64** ($R^2 = 0.41$)

** = $p < 0.01$

Table 39: Pearson-Product-Moment correlations between appraisal calibration and IELTS Listening accuracy and appraisal confidence (N = 376)

A negative sign of the correlations suggests that bias was higher for low ability test-takers. The correlational analysis suggests that cognitive bias might be present in Sections 1 and 3 and that motivational bias might be present in Sections 2 and 4. According to Stankov and Lee (2014a), cognitive bias implies that people cannot accurately assess their confidence, resulting in a biased self-conception (e.g., Kruger & Dunning 1999). *Motivational bias* suggests that people are overconfident because there is a psychological benefit of overconfidence, e.g., it helps improve task motivation (Pajares 1996) and a social benefit of overconfidence (e.g., it helps people convince others that they are more able than they actually are (Anderson, Brion, Moore & Kennedy 2012).

In the present study, as earlier analysis suggested that test-takers performed significantly better in Sections 1 and 3 than in Sections 2 and 4, the presence of motivational bias was not surprising, but was an important discovery. That is, when test-takers perceive they are faced with a difficult task, they are likely to feel more confident, which helps them improve their motivation to improve their task completion. The correlation coefficient of 0.64 in Section 4 (compared to 0.54 in Section 2) is consistent with this assumption because Section 4 was the most difficult section for this group of test-takers.

These findings of overconfidence are in line with calibration research, which finds that people are generally overconfident in their performance (e.g., Epstein et al. 1984; Hattie 2013; Hadwin & Webster 2013; Maki & Serra 1992; Moore & Healy 2008; Schraw et al. 2013; Soll 1996; Stankov, Lee & Paek 2009; Weaver & Bryant 1995). It should be noted that the present study included a 0% confidence scale in the confidence measure, whereas some other studies did not have this scale, which might have resulted in a finding of overconfidence in those studies (this is then an artefact of researchers' confidence measures). However, the present study did not consider the influence of guessing test answers when calculating test-takers' appraisal calibration. That is, it is possible that a test-taker guessed the answer to the given test question and expressed a 0% appraisal confidence, for example. However, if the guessed answer were correct, that test-taker would be found to be 100% underconfident. Future research should consider this and integrate the guessing factor into calibration studies.

In addition to calibration scores, correlational analysis was performed to examine the relationships between appraisal confidence and test performance. It was found that there were statistically significant relationships between appraisal confidence and performance across the four listening test sections (see Tables 23 and 25). A high correlation indicates that test-takers can accurately monitor their performance. A zero or non-significant correlation suggests that there is no association between their confidence and their actual performance. The latter case suggests test-takers have a serious metacognitive monitoring problem. There were variations in the significant relationships between appraisal confidence and performance, and appraisal confidence for each test section was not a strong predictor of the associated test performance.

In the present study, when the overall correlation analysis between appraisal confidence and performance was performed for the overall test, the correlation coefficients were found to be large (0.73 for Pearson-Product-Moment correlation and 0.79 for SEM correlation). Phakiti (2016) found the SEM correlation coefficient of 0.61 ($R^2 = 37$) between confidence and performance of Thai EFL university students. The correlation coefficients indicate that, as success in test performance increases, appraisal confidence also increases. Although test-takers were not very accurate in their performance appraisals, the coefficients show that higher-ability test-takers reported higher confidence than lower-ability ones did. The present study, therefore, found a strong association between appraisal confidence and test performance when they examined as a whole test, but the associations were weaker when considered at a test section level.

According to van Loon, de Bruin, van Gog and van Merriënboer (2013), previous research found correlations between confidence and accuracy to be less than 0.25. Maki and Serra (1992) found the correlations to be less than 0.35. Weaver and Bryan (1995) found a correlation to be as large as 0.69. The present findings for the correlations were larger than those discussed in Dunlosky and Lipko (2007). Stankov and Lee (2008) reported Pearson-Product-Moment correlation coefficients of 0.61 (reading 1), 0.52 (reading 2), 0.45 (listening 1) and 0.48 (listening 2). Stankov, Lee, Luo and Hogan (2012) reported Pearson-Product-Moment correlations of 0.48 (English Grammar), 0.56 (English vocabulary) and 0.49 (English reading comprehension) between confidence and accuracy.

The correlation coefficients in the present study were similar to those found in these previous studies.

The present findings can help inform research on self-assessment as discussed in Oscarson (1997, 2014), Ross (1998) and Matsuno (2009), who found that L2 learners' self-assessment was poorly related to language performance across language skills. In self-assessment research, Blanche and Merino (1989) suggested that low-proficient learners tended to overestimate their skills and high-proficient learners tended to underestimate their skills. The study by Trofimovich, Isaacs, Kennedy, Saito and Crowther (2016) found that L2 learners' self-assessment on L2 speech were inaccurate, suggesting that low-ability learners overestimated their performance and high-ability learners underestimated their speech performance.

Finally, unlike previous research, the present study examines the structural relationship between single-case appraisal confidence and relative-frequency appraisal confidence through SEM. It was found that the two types of confidence judgment were highly correlated (0.93, see Figure 20), suggesting that there could be a general calibration factor (see Figure 21). The present study suggests that in calibration research, data on both forms of confidence should be collected as they allow researchers to have more comprehensive information about test-takers' performance appraisals.

5.1.2 Research question 2: The nature of confidence and calibration in easy, moderately difficult, very difficult and extremely difficult questions

Since Research Question 1 focuses on overall appraisal confidence and calibration across test sections, it is important to examine test-takers' appraisal confidence and calibration based on test difficulty level. The study rigorously addressed this objective through the use of Rasch IRT analysis. The analysis reveals several useful findings that could explain the nature of calibration found in Research Question 1.

First, it was found that in easy questions, the calibration score was within 5%, suggesting that test-takers were likely to be able to estimate their performance success. However, their calibration scores deviated more and more from zero as the test difficulty level increased (see Table 27). It was found that in very difficult questions, test-takers were 25% overconfident (Cohen's $d = 0.62$).

In the calibration literature, the *hard-easy effect* has been discussed (Stankov & Lee 2008b). The hard-easy effect phenomenon can be observed when people are overconfident in difficult tasks but underconfident in easy tasks. Although underconfidence was not observed in this analysis, it was found that the hard-easy effect can be highly complex because a closer look at the calibration diagram in Figure 22 suggests that many high-ability test-takers (above 70% performance) tended to be underconfident in easy questions, whereas many low-ability test-takers (below 50% performance) tended to be overconfident in easy questions. The overconfidence phenomenon in the very difficult questions was clearly pronounced (see Figure 25).

A closer examination of test-takers' calibration scores within $\pm 5\%$ across the four difficulty levels indicates that the number of calibrated test-takers was larger for easy questions than for very difficult questions (N = 157 (42%) for easy questions, N = 108 (29%) for moderately difficult questions, N = 66 (18%) for difficult questions, and N = 54 (7%) for very difficult questions). On the basis of this analysis, up to 93% of the test-takers were miscalibrated in very difficult questions.

Moore and Healy (2008) argued that the interaction between an individual's ability and task difficulty level has a potential impact on variation in calibration. In addition to the use of calibration scores, Pearson-Product-Moment and SEM correlations were computed to examine the relationships between appraisal confidence and performance across the four difficulty levels (see Table 28). Based on SEM correlation coefficients, it was found that test-takers' confidence judgment explained only 25% of the test performance variance in easy questions, 49% in moderately difficult questions, 23% in difficult questions and 14% in the very difficult questions.

To examine whether test-takers' confidence at these difficulty levels were the result of cognitive bias or motivational bias (Stankov & Lee 2014a), Pearson-Product-Moment correlations were computed and reported in Table 40. A negative sign of the correlation implies that low-ability groups had higher bias scores. The correlational analysis suggests that cognitive bias might be present in easy questions ($r = -0.57$ between accuracy and calibration) and moderately difficult questions ($r = -0.58$ between accuracy and calibration), whereas motivational bias might be present in very difficult questions ($r = 0.61$ between confidence and calibration).

	IELTS scores	Confidence score
Easy questions	-0.57** (R ² = 0.32)	0.27** (R ² = 0.07)
Moderately difficult questions	-0.58** (R ² = 0.34)	0.16** (R ² = 0.03)
Difficult questions	-0.47** (R ² = 0.22)	0.45** (R ² = 0.20)
Very difficult questions	-0.24** (R ² = 0.06)	0.61** (R ² = 0.37)

** = $p < 0.01$

Table 40: Pearson-Product-Moment correlations between appraisal calibration and IELTS Listening accuracy and appraisal confidence based on difficulty levels (N = 376)

Both cognitive and motivational bias appeared to be present in equal measure in difficult questions ($r = -0.47$ versus 0.45). As pointed out above, when faced with difficult tasks, appraisal confidence might serve as a motivational reason to carry out the given test, thereby affecting calibration.

On the basis of these analyses in regard to Research Question 2, the 'hard-easy' effect might partly be responsible for appraisal calibration variation among test-takers in this IELTS Listening test. According to Suantak, Bolger, and Ferrell (1996), the hard-easy effect could be an outcome of poor performance appraisals. The researchers argued that individuals can be insensitive to changing levels of task difficulty and, therefore, may fail to adjust their internal response criteria to changes in task demands during the course of test completion. According to Stone (2000), overconfidence indicates that individuals fail to detect an increase in cognitive task difficulty when their appraisal confidence is consistently high. Underconfidence, on the contrary, results from their failure to detect a decline in task difficulty when their confidence is consistently low. Hadwin and Webster (2013) and Kleitman and Stankov (2001), for example, point out that those who are overconfident in one task are likely to exhibit overconfidence in other tasks as well.

Clearly in a language testing and assessment situation, it is desirable for test-takers to be more calibrated when they are faced with difficult questions than with easy ones. Being overconfident in one's performance in difficult questions/tasks has a negative impact on test performance. According to Moore and Healy (2008), the phenomenon of over- and under- confidence can be highly complex. That is, when individuals overestimate their test performance in difficult questions, they believe that their performance is worse than that of other people, but when they underestimate their performance in easy question, they believe that their performance is better than that of others.

5.1.3 Research question 3: Gender differences in appraisal confidence and calibration scores

In addition to test task difficulty level, a gender factor was considered as part of a plausible reason for the finding of poor appraisal calibration among test-takers (Research Question 1). It was found that both male and female test-takers tended to be overconfident in their test performance across different test sections, as well as across different test difficulty levels (see Table 30). Gender differences were compared in different test sections as well as test difficulty levels.

In this study, through the use of ANOVA, it was found that male and female test-takers did not differ in their appraisal calibration in Sections 1 and 3, nor in difficult and very difficult questions. However, female test-takers outperformed their male counterparts in the listening test scores of Sections 1 and 3 (see Table 32) and they were also found to have better calibration scores in these two sections than their male counterparts (see Table 33). That is, not only did they outperform their male counterparts in easy and moderately difficult questions, but they also had better calibration scores than their male counterparts at these two difficulty levels (see Table 33). The magnitude of the effect sizes, however, was small.

It is possible that gender differences may become more significant in cohorts with a different nationality make-up or larger sample sizes in the revised report.

The gender variable might, nonetheless, partly explain variation in the overconfidence phenomenon among the test-takers. The descriptive statistics suggest that male test-takers tended to report higher appraisal confidence, as well as to be less calibrated than their female counterparts.

The present findings were similar to those found in Pallier et al. (2002), Pallier (2003) and Stankov and Lee (2014a), who found that males were more overconfident than females, and females had a better calibration score. Stankov and Crawford (1996, 1997) did not find any significant gender differences in appraisal confidence and calibration. In L2 research, a considerable amount of empirical work has been undertaken to examine gender differences as a source of successful language learning (Chavez, 2001; Phakiti 2003a).

Generally speaking, females tend to be more successful than males in language learning, suggesting that gender is an independent factor that has a potential to affect differences in learning success and behaviours.

It should be noted that findings regarding gender differences in test performance and other psychological attributes do not directly imply that a test is unfair and biased against one gender. Such findings only suggest that male and female test-takers may have different thought processes or abilities. However, if males in general were found to be less calibrated, as well as less successful than females, male students may be provided with some feedback on their appraisal ability and a remedial activity should be developed to help them improve their use of metacognitive strategies.

5.1.4 Research question 4: Test-takers with different success levels and their appraisal calibration scores

As found in Research Questions 1 and 2, test-takers' success levels might interact with their overall appraisal calibration. To further examine whether differences in success level produces differences in calibration, Rasch IRT analysis was performed to identify the levels of test-takers. Originally, it was intended that test-takers would be grouped into high-ability, medium-ability and low-ability groups. However, it was found that the number of test-takers per group was imbalanced. It was decided that IRT logit scores be used to group test-takers into six distinctive-ability levels (Group 1 being the highest ability and Group 6 being the lowest ability).

Test-takers in Group 1, in particular, tended to have better appraisal calibration scores, compared to test-takers in other ability groups. The majority of test-takers in each group were overconfident in their performance.

The findings suggest that the relationship between appraisal confidence and accuracy may be *quadratic* (Jackson & Kleitman 2014), which implies that optimal behaviour can be expected when people are calibrated (e.g., in the case of Group 1).

Test-takers at all ability levels were found to be overconfident in their performance, except in some test sections or at the high difficulty level, at which the high-ability test-takers were a just underconfident or quite realistic. The *Kruger and Dunning effect* was partly observed, especially with the low-ability test-takers. The unskilled and unaware effect was also observed since test-takers tended to be highly overconfident in difficult tasks resulting in a low level of accuracy, but relatively well-calibrated in easy tasks resulting in a high level of accuracy (Ehrlinger & Dunning 2003; Ehrlinger et al. 2008).

In Field's (2013) five-level L2 listening model, low-ability listeners are likely to process listening text at the first three levels of listening comprehension (i.e., input decoding, lexical searching and parsing). However, the monitoring and decision judgment processes of audio text can only take place adequately at level five (i.e., discourse construction). It might well be that test-takers in Groups 4 to 6 could not adequately reach levels four and five of listening comprehension. Given this, they would be unlikely to be realistic about their test performance.

Nonetheless, appraisal calibration cannot merely be accounted for by ability level because all test-taker groups were found to be overconfident across different test sections and difficulty levels.

The phenomenon of overconfidence observed in the present study suggests a complex interaction between language ability level and metacognition. The present findings were similar to those in Stankov and Lee (2014a), Stankov et al. (2009) and Stankov et al (2012), which found that overconfidence was present not only for low-ability students, but also for high-ability students, but overconfidence was less pronounced among high-ability groups, compared to that of low-ability groups. Lichtenstein and Fischhoff (1977) were among the earliest studies on human appraisal calibration, investigating whether those who know more also know more about how much they know.

The present study found that the highest ability group tended to know how much they knew or how well they could perform better than those who knew less. However, overconfidence still tends to be manifested among different ability levels as well as at different test difficulty levels.

5.1.5 Research question 5: The structural relationships among test-takers' confidence, calibration, trait and state cognitive and metacognitive strategy use, IELTS listening test difficulty, and IELTS Listening performance

One of the key goals of this study is to be able to understand the connections between performance appraisals, trait and state strategy use and perceived listening difficulty through simultaneous data analysis of all these variables. Previous calibration research focused on how confidence and calibration were related to other psychological traits, such as self-concept, anxiety, self-efficacy, self-protection, and self-enhancement (e.g., Jiang & Kleitman 2015; Stankov et al. 2012). Little is known about how calibration and confidence judgments are related to strategy use. In the present study, several SEM models were tested from the simplest structural models to the most complex ones. For the purpose of this section, only two SEM models are discussed because they provide insight into the relationship between appraisal confidence, trait and state strategy use and IELTS test performance.

In Figure 44, it was found that single-case appraisal confidence and relative-frequency appraisal confidence had a significant relationship to both trait and state metacognitive strategy use. The correlation coefficient between single-case appraisal confidence and trait metacognitive strategy use was 0.61 ($R^2 = 0.37$, large effect size), whereas the coefficient between relative-frequency appraisal confidence and trait metacognitive strategy use was 0.43 ($R^2 = 0.18$, medium effect size). Single-case appraisal confidence had an indirect influence on state metacognitive strategy use via relative-frequency appraisal confidence (0.42, $R^2 = 0.17$, medium effect size) and trait metacognitive strategy use (0.32, $R^2 = 0.10$, medium effect size).

On the basis of these findings, although test-takers' appraisal confidence judgments were significantly related to trait and state metacognitive strategy use, the strengths of the detected relationships were not very large. It can be inferred that test-takers' appraisal confidence judgments did not have a high impact on their metacognitive strategy use during this listening test.

Single-case appraisal confidence was found to have a small indirect impact on IELTS Listening test performance via (1) relative-frequency appraisal confidence and state metacognitive and cognitive strategy use ($\gamma = 0.18$, $R^2 = 0.03$), and (2) trait and state metacognitive strategy use, state cognitive strategy use ($\gamma = 0.14$, $R^2 = 0.02$). The addition of appraisal confidence to the SEM model changed the regression coefficient from state cognitive strategy use only slightly, from 0.52 to 0.55 (see Figure 49). The reason for this may well be explained by the finding that the test-takers were miscalibrated and overconfident in their test performance, especially in very easy or difficult questions.

A similar strength of the indirect influence of single-case appraisal confidence on EFL test performance via state metacognitive and cognitive metacognitive strategy use was also found in Phakiti (2016; $\gamma = 0.15$, $R^2 = 0.02$). Again the effect of appraisal confidence on reported strategy use was not found to be strong, implying that test-takers might not have adjusted their strategy use according to the complexity or the demands of the listening test tasks. It might be that overconfidence mediates the effect of inaccurate performance appraisals since test-takers found it difficult to assess the accuracy of their own performance estimates, thereby failed to use appropriate strategies to tackle the tasks at hand.

Metacognitive and self-regulated learning theories have highlighted the importance of knowing what one knows, as it is central to successful learning (Efklides, 2008; Tobias & Everson 2009). For example, accurate monitoring can lead to appropriate self-regulation (e.g., maintaining motivation to learn, adopting appropriate help-seeking behaviours). Hattie (2013) pointed out that underconfident students may have low self-efficacy and spend unnecessary time and effort on task completion and that overconfidence can negatively impact deep learning and examination preparation and academic project completion. According to Butler and Winne (1995) and Hardwin and Webster (2013), when students are overconfident, they are unlikely to suitably regulate their strategies to facilitate their performance. Consequently, they may fail to realise when they should be actively regulating or trying out new strategies to increase the likelihood of achieving their goal. Hattie (2013) points out that students may be able to know what they know, but may be less able to judge what they do not know.

Figure 46 illustrates how test-takers' appraisal calibration was related to trait and state metacognitive strategy use. It was found that single-case appraisal calibration had a positive yet small relationship to trait and state metacognitive strategy use. The SEM correlation coefficient between appraisal calibration and trait metacognitive strategy use was 0.28 ($R^2 = 0.08$), whereas the regression coefficient from single-case appraisal calibration to state metacognitive strategy use was 0.13 ($R^2 = 0.02$). Both coefficients had a small effect size, suggesting that test-takers' appraisal calibration was not well connected to how test-takers reported using trait and state metacognitive strategies, which in turn regulated trait and state cognitive strategies. The regression coefficient between trait metacognitive strategy use and state metacognitive strategy use changed from 0.80 (Figure 41) to 0.55 (Figure 43) when the appraisal confidence variables were added to the SEM model.

This finding implies that inaccurate performance appraisals might reduce the strength of the relationship between trait metacognitive strategy use as part of strategic knowledge and state metacognitive strategy use as part of strategy regulation (see Phakiti 2007b). Previous test-taking strategy research, which relied on the use of questionnaires found a weak relationship between reported strategy use and L2 performance (see e.g., Bi 2014; Phakiti 2008; Purpura 1999; Zhang & Zhang 2013). The missing explanation for such weak relationships may well be test-takers' poor appraisal calibration.

Figure 47 has presented the most complex SEM model that takes all the variables into account in order to estimate the relationships among appraisal confidence, trait and state strategy use and listening difficulty, and IELTS Listening performance. This model provides estimates of the various factors that are related to IELTS Listening test performance. In this model, it was found that when trait and state perceived IELTS Listening difficulty were simultaneously analysed, state and trait IELTS Listening difficulty had a negative impact on the IELTS test performance and accounted for 10% and 4% of the IELTS Listening test and trait IELTS Listening difficulty, respectively.

The findings imply that test-takers' perceptions of test item difficulty level were predictive of their performance.

In Figure 47, the prediction of state IELTS Listening difficulty was nearly as much as that by state cognitive strategy use ($\beta = 0.36$, $R^2 = 0.13$). It was, however, found that state metacognitive strategy use could reduce the impact of trait IELTS Listening difficulty on state IELTS Listening difficulty (from 0.61 to 0.76 when the regression path between state metacognitive strategy use and state IELTS Listening difficulty was dropped from a subsequent analysis). The SEM model in Figure 51 also indicates that trait and state metacognitive strategy use had a negative relationship to trait and state IELTS Listening difficulty ($\gamma = 0.52$ and $\beta = -0.29$, respectively).

The findings suggest that as test-takers report more use of trait and state metacognitive strategies in IELTS Listening tests, their perceived trait and state IELTS Listening difficulty can be reduced. In this analysis, it was also found that both single-case and relative-frequency confidence played an indirect role to reduce the influence of trait and state IELTS difficulty on the IELTS Listening test performance.

5.2 Limitations of the present study

The present study has some key limitations that should be noted. First, the current findings relied on the test-takers' accounts of their reported performance appraisals and their motivation to complete the test. Since there was no impact of the test scores on test-takers' academic grades, some test-takers might not have made much effort to do the listening test well or to report their appraisal confidence and strategy use with any great thought.

Second, the findings were skewed not only by the instruments used (e.g., the IELTS Listening test, the test conditions, the appraisal confidence rating scales), but also their reliability and validity. Although the participants had the appraisal confidence rating scale explained to them and practised rating appraisal confidence prior to the data collection, they might still have been unfamiliar with this kind of appraisal rating, which would have affected the current findings. Furthermore, it may well be that these test-takers were not used to judging their test performance and hence by being asked to do so, they may not have been able to translate their performance appraisals into the appraisal confidence scales adequately. Hattie (2013) pointed out that measurement errors can be frequent when research instruments that measure individuals' behaviours, attitudes and feelings are employed in research.

In relation to appraisal confidence ratings, Hadwin and Webster (2013) pointed out that data involving self-reported ratings of appraisal confidence often suffer from a restricted range because participants tend to limit appraisal confidence ratings to a small range of possible values. Without detailed discussion and explanation, participants may misinterpret the values of the confidence scales (Hadwin & Webster 2013; Stankov et al. 2009).

Third, it is undeniable that strategic competence is highly complex and appears to have several facets that operate both at the conscious and unconscious levels. Accordingly, the present study and any research of this sort is limited by the fact that performance appraisals are treated as being at a conscious level, while in fact numerous performance appraisals may operate at an unconscious and unreflective level.

In the current study, if some performance appraisals were operating at an unconscious and unreflective level, test-takers might have failed to report their confidence judgments accurately. Difficulty in translating unconscious performance appraisals can lead to findings of miscalibration. Nevertheless, metacognitive theories (e.g., Efklides 2008, 2011; Hacker, Dunlosky & Graesser 2009) suggest that people are likely to be conscious about a cognitive task when they have a feeling of uncertainty or difficulty.

The level of conscious and non-conscious awareness of metacognition is, nonetheless, not easy to resolve. Efklides (2008) pointed out that the association of metacognition with consciousness is no doubt absolutely necessary to help researchers understand how people take control of their cognitive activities, especially when automaticity fails. Efklides (2008) also pointed out that as explicit conscious awareness and implicit non-conscious awareness do not necessarily function on the same continuum, "the dissociation between conscious and non-conscious awareness does not mean that there is no regulation of cognition at an implicit level" (p. 281). Roderer and Roebers (2014) found that appraisal confidence judgments can be located somewhere at the end of the sometimes quite complex memory retrieval and memory monitoring processes as they either follow sequentially or occur in parallel. Hence, access to performance appraisals can be difficult and people may be found to be miscalibrated as a consequence.

Fourth, while comprehensive, the findings in the present study are mainly quantitative. Qualitative data such as verbal reports would yield further insight into performance appraisals, trait and state strategy use and the reasons for appraisal confidence judgments and the calibration of test-takers. A hybrid approach combining both quantitative and qualitative findings may lead to new findings and improved theories.

In the current study, together with the current quantitative data, data from a small-scale study that combines both quantitative and qualitative data through case studies have been collected. Participants who had taken the current test were recruited for retrospective, semi-structured individual interviews. They were also asked to complete a different IELTS Listening test and rate their confidence judgments. This small-scale study was conducted to further examine the potential reasons behind their confidence and the factors affecting their confidence ratings and calibration.

This qualitative data set can assist validation for variation and deviation from the norm (i.e., the quantitative findings) and provide new information and insights into the issues that are unforeseen or unique to a particular test-taker and context. However, since the quantitative data analysis has been highly complex with several new research questions for IELTS Listening tests, it was not possible to include details of this study in this report. The rationale and level of qualitative analysis needs to be discussed and presented thoroughly in an independent report. Hence, the analysis of this dataset remains to be done and reported in subsequent publications.

6 CONCLUSIONS AND IMPLICATIONS

This study has examined the nature of test-takers' performance appraisals of the correctness of their answers, their calibration, trait and state strategy use and IELTS listening difficulty in a simulated IELTS Listening test. The study has focused on appraisal calibration, which is the alignment between test-takers' appraisal confidence judgments and performance accuracy. This ability is considered an essential component of strategic competence that had not been much explored.

It was found that test-takers' calibration scores varied according to the test difficulty level, gender and test-taker ability level. The participant test-takers were generally miscalibrated in their performance appraisals and exhibited a tendency to be overconfident across the four test sections.

Their appraisal calibration scores were particularly poor in Section 4 and in the most difficult questions. The concurrent validity of test-takers' performance appraisals is critical because accurate appraisals have consequences for other subsequent metacognitive processes, such as planning, goal setting and monitoring. For example, if test-takers believe that their performance is already very good, they are less likely to worry about their performance. However, if they are incorrect in their performance appraisals (that is, their performance is in fact not as good as they believe), they are unlikely to take extra action or to make an effort to improve their work or to use more advanced metacognitive strategies (Tobia & Everson 2009). Performance appraisals can, therefore, influence other affective schemata (e.g., motivation, self-efficacy, and effort), and cognitive strategy use.

An examination of whether successful test-takers differ from unsuccessful ones in terms of their appraisal calibration in a language test situation is also crucial to explain the qualitative difference of strategic competence among test-takers with varying language abilities. Performance appraisals and appraisal calibration are multi-faceted and multi-level constructs that interact closely with other facets of strategic competence.

Furthermore, unlike previous strategic research, which asked test-takers to report their strategy use at either the beginning or at the end of the whole test, this study has asked test-takers to report both at the beginning (trait strategy use), during the test taking (performance appraisals) and at the end of the test (state strategy use). This method has allowed the researcher to comprehensively investigate the cognitive validity of the IELTS Listening test (e.g., performance appraisals and monitoring accuracy in the course of IELTS listening).

6.1 Implications for the IELTS Listening test

This study has shed light on the validity of IELTS Listening test tasks. According to Enright and Jamieson (2008, 2010), validation research aims to gather evidence of targeted language abilities (i.e., evaluation), evidence of score consistency across different test tasks or questions (i.e., generalisation), and evidence of listening scores that reflect target language proficiency and feedback (i.e. explanation).

Strategic competence has long been theorised to be part of communicative language ability (e.g., Bachman & Palmer 2010). While this competence is not directly included in a test score, it is assumed to affect test score variation. That is, good test performance implies good strategic competence. It has been found that several facets of strategic competence, such as trait and state strategy use and performance appraisals were related to IELTS Listening test performance.

For IELTS Listening tests, the finding that test-takers were likely to be overconfident in their test performance, particularly in difficult test questions, can explain why some test-takers are unsuccessful in the IELTS Listening test. Poor appraisal calibration as found in the current study is indicative of a cognitive problem that clearly goes beyond any language proficiency tests. Research on individuals' appraisal confidence judgments and appraisal calibration has found extensive evidence indicating that overconfidence and miscalibration is a common phenomenon among test-takers (see e.g., Stankov & Lee 2014a).

6.2 Implications for language teaching and IELTS test preparation

Although the present study involves basic research and does not focus on teaching implications, it has yielded some implications related to washback since teaching and assessment are intertwined ingredients in education. While it may be common for students to be overconfident (e.g., Hattie 2013; Stone 2000), overconfidence may not be *universal*. Hattie (2013) pointed out that students rarely receive external evaluation of their progress and have little idea of what success on the task should look like.

It can be argued that, with some intervention, students can become more calibrated in their metacognitive appraisals. An experimental study by Cao and Nietfeld (2005), for example, suggested that high-performing students made significantly better judgments of learning and monitoring accuracy than low-performing students. The researchers also found that increased correlations between judgments of learning and monitoring accuracy resulted from weekly monitoring exercises, suggesting that students could be trained to judge their performance more accurately.

Nietfield and Schraw (2002), through two experimental studies, found that strategy training and activation of prior knowledge improved mathematics students' monitoring accuracy and self-efficacy (see also Bol et al. 2005; Gutierrez & Schraw 2014; Kleitman & Costa 2014). Koriat, Lichtenstein and Fischhoff (1980) found that overconfidence occurs when people ignore or disregard contradictory information present in a given context. Koriat et al (1980) demonstrated how calibration can be improved by, for example, asking people why they chose a particular answer over other possible answers.

In an L2 context, the integration of performance appraisal training will be of significant value to L2 learners/test-takers. Like training human scorers in writing or speaking tests, students may be better calibrated by, for instance, receiving explicit feedback on whether they are realistic, overconfident or underconfident. Metacognitive appraisal instruction may vary according to the level of language skills of the students because the nature of cognitive processing, feedback and task demands can vary (Kostons et al. 2012; VanPatten 1994). It may be effective for each learner to keep a record of their calibration graphs as a reminder of their calibration development (see Example of feedback to students in Appendix 1, A1.9). One outcome from this kind of instruction, as Zimmerman (1994) asserted, is that metacognitive learners/test-takers will be made more aware of what they know and what they do not know.

Through the application of the cognitive processing and appraisal confidence level generation presented in Figure 3, it is possible in an IELTS Listening test preparation to integrate appraisal confidence ratings in test tasks. A student-friendly version of this flowchart can be provided to students, so that they can identify the level of their appraisal confidence that best represents the likelihood of their performance success. It can be argued that providing a confidence judgment requires test-takers to assess their accuracy level more overtly than if no confidence estimate is required. Confidence judgments require them to monitor their performance and by doing so, may initiate the use of appropriate metacognitive strategies to address different test difficulty levels. Incorporating confidence judgments into IELTS test tasks allows test-takers not only to develop the ability to achieve their desired performance levels, but also, with some *post hoc* metacognitive feedback, to have a greater awareness of what they know and what they do not know.

An inclusion of appraisal confidence in IELTS Listening practice tasks can promote positive washback because students will be more overt about appraising their test performance. In order to prepare for an official IELTS test, assessing appraisal confidence can be seen as a *debiasing technique* that helps test-takers to approach better appraisal calibration, which subsequently improves test performance.

The use of appraisal confidence ratings can allow language teachers and researchers to understand the reasons why test-takers choose a wrong answer during a test. Such information can lead to the adoption of remedial activities in the classroom (Caleon & Subramaniam 2010; Stankov et al. 2012). An experimental research design that compares the differences between students receiving appraisal confidence training and those who have not will allow the usefulness and practicality of this metacognitive intervention to be evaluated.

It is also important to point out that, according to Kostons et al. (2012), teachers are not always equipped with the skills to help students develop their monitoring accuracy and metacognitive appraisals. That is, many language teachers may not know how to effectively provide cognitive feedback to their students (on whether their students are overconfident, underconfident or realistic, for example). For this reason, a calibration curriculum design for language teacher training should be considered and developed.

6.3 Recommendations for future research

The nature of the present study remains exploratory given the novel nature of the application of appraisal calibration concepts in language testing and assessment research. Several aspects in this study need to be replicated using various IELTS Listening tests and extended to other IELTS skills tests, as well as other language tests and assessments. Much remains to be done with the current quantitative data, especially with the numerous different ways that exist to examine test-takers' appraisal calibration (e.g., Schraw et al. 2013; Schraw et al. 2014). However, the study has opened several avenues for future research.

First, future research needs to consider the extent to which the nature of appraisal calibration found in the present study holds across other groups of test-takers and in other test contexts.

More research that examines test-takers' appraisal calibration and state and trait strategy use in IELTS Listening tests is needed in different contexts and at different proficiency levels. Since the participants in the current study were mainly Chinese and had an average overall IELTS and IELTS Listening scores of 5.67 (SD = 0.75) and 5.71 (SD = 0.94; N = 225), respectively, there remains a need for IELTS validation research that includes test-takers from non-Chinese backgrounds and with other score bands.

If the extent to which IELTS Listening test-takers are calibrated (or miscalibrated) is well-understood empirically across various test-taker groups and contexts, we will begin to better understand the complex nature of language test performance and strategic competence during test taking and to better use IELTS test scores to infer performance in non-test language use situations. Future research should also address the other components of the IELTS test beyond the listening test.

Second, it is reasonable to argue that appraisal calibration of test-takers' performance appraisals is not only determined by an individual (i.e., individual-driven factors, such as strategic competence levels, age, educational backgrounds, motivation and test-anxiety) but also by the context in which the test takes place (i.e., context-driven factors, such as the kinds of language test tasks, language modes and the nature of the stakes associated with the test scores). Future research should aim to understand the influences of these factors on test-takers' appraisal calibration and to identify the conditions under which test-takers can be calibrated in their appraisals.

Third, future research would help advance our understanding of the relationships between calibration and other cognitive constructs. In cognitive and educational psychology, cognitive constructs such as self-conception, self-efficacy, motivation, test anxiety and epistemological beliefs have recently been investigated in relation to appraisal confidence (e.g., Jiang & Kleitman 2015; Morony et al. 2013; Stankov & Lee 2014a, Stankov et al. 2012). Jiang and Kleitman, for example, found that students who engage in self-protection behaviours may lower their performance appraisals of their ability, which in turn may affect their on-task appraisal confidence. A longitudinal research design and/or a mixed methods design would allow researchers to generalise the patterns of appraisal calibration and miscalibration over time across different IELTS Listening test sections.

Fourth, since the present study is a pioneer study on test-takers' appraisal calibration in IELTS Listening tests, it is still the case that little is known and understood about appraisal calibration and the phenomena of overconfidence, underconfidence, hard-easy effects, and Kruger-Dunning effects. Little is understood about the reciprocal relationships between performance appraisals, trait and state strategy use, and perceived listening difficulty. This line of research should be also extended to an examination of cross-cultural differences in appraisal calibration. According to Stankov and Lee (2014b), since academic self-beliefs, such as self-efficacy and self-concept, are related to appraisal confidence (Morony, Kleitman, Lee & Stankov 2013) and since people from Confucian countries appear to be more unforgiving than people from the rest of the world (e.g., Stankov 2009), cross-cultural comparisons, for example between Confucian Asians and North Americans or Europeans would shed light on whether culture plays a role in confidence judgments. A study by Yates, Shinotsuko, Patalano and Sieck (1998) on decision-making found that Confucians were more overconfident than North Americans, whereas Stankov and Lee (2014a) had mixed findings.

Finally, although miscalibration is indicative of inaccurate strategic competence, more empirical research is needed to identify and understand the sources of miscalibration, as well as the reasons for miscalibration. Research, including the present study, has suggested that too much overconfidence can be detrimental to learning or task performance (e.g., Kruger & Dunning 1999; Stankov & Lee 2014b), and some cognitive researchers have begun to explore why people tend to be overconfident.

To date, some plausible reasons that have been put forward to explain why some people are overconfident include the desire to: *maintain high self-regard* (Moore & Healy 2008, Stankov & Lee 2014a); *enhance self-esteem* (Jiang & Kleitman 2015); *serve as a compensatory strategy* to regulate motivational state (Klassen 2002); *enhance social status* (Anderson et al. 2012); *appear according to social desirability* (Roderer & Roebbers 2014); and to *act as a cushion against feelings of failure* (Jiang & Kleitman 2015).

Little is understood regarding the reasons for some people to be underconfident (see Stankov, Pallier, Danthir & Morony 2012) and therefore more research in test-takers' calibration in language assessment contexts is needed.

6.4 Concluding remarks

The ability to be calibrated when taking a high-stakes test with test items of varying levels of difficulty can explain individual differences in language test performance and hence success in strategy use. Good appraisal calibration is especially important when the language situation requires test-takers to be aware of their actions or performance and the potential consequences of their performance on their future. Although calibration does not necessarily guarantee test accuracy, it is arguably a desirable ability for successful test taking and language use.

ACKNOWLEDGMENTS

I would like to express my appreciation to IDP, IELTS Australia for funding this research project and the students who took part in this study.

Throughout this research journey, I am extremely grateful to the following people:

- Jenny Osborne and Stephanie Bethencourt and their team at IDP, IELTS Australia for their generous support throughout the completion of this project
- The two anonymous reviewers
- Patrick Pheasant, Georgiana Toma, Katherine Olston, Paula Gothelf, Rosie Giddings and Paul Mahony for allowing me to recruit participants at their institutes for this project
- Patrick Brownlee, Jane Harvey and Sue Goodwin in the Faculty Research Office for their support
- Brian Paltridge, Janette Bobis, Sabina Kleitman, Lazar Stankov, Jim Purpura, Andrew Cohen, Antony Kunnan and Talia Isaacs for their engaging discussion on this topic
- Guy Middleton for critically and patiently editing and commenting on this long report
- Nick Bi for being such a great research assistant in this project.

Several parts of this research have been presented at the *Language Testing Research Colloquium* in Toronto (2015), the *Asian Association for Language Assessment Conference* in Thailand (2015), a *Research Seminar in TESOL and Language Studies* at the University of Sydney (2015), and at the ALAA/ALANZ/ALTTANZ Conference in Adelaide, Australia (2015).

REFERENCES

- Ackerman, PL & Wolman, SD, 2007, 'Determinants and validity of self-estimates of abilities and self-concept measures', *Journal of Experimental Psychology: Applied*, vol. 13, no. 2, pp. 57–78.
- Alexander, PA, 2013, 'Calibration: What is it and why it matters? An introduction to the special issue on calibrating calibration', *Learning and Instruction*, vol. 24, no. 1, pp. 1–3.
- Alexander, PA, Graham, S & Harris, KR, 1998, 'A perspective on strategy research: Progress and prospects', *Educational Psychology Review*, vol. 10, no. 2, pp. 129–153.
- Anderson, C, Brion, S, Moore, DA & Kennedy, JA, 2012, 'A status-enhancement account of overconfidence', *Journal of Personality and Social Psychology*, vol. 103, no. 4, pp. 718–735.
- Andrade, MS, 2006, 'International students in English-speaking universities', *Journal of Research in International Education*, vol. 5, no. 2, pp. 131–154.
- Andrade, MS, 2010, 'Increasing accountability: Faculty Perspectives on the English language competence of non-native English speakers', *Journal of Studies in International Education*, vol. 14, no. 3, pp. 221–239.
- Aryadoust, V, 2011, 'Validity arguments of the speaking and listening modules of International English Language Testing System: A synthesis of existing research', *The Asian ESP Journal*, vol. 7, no. 1, pp. 28–54.
- Aryadoust, V, 2013, *Building a validity argument for a listening test of academic proficiency*, Cambridge Scholars, Cambridge.
- Bachman, LF, 1990, *Fundamental considerations in language testing*, Oxford University Press, Oxford.
- Bachman, LF, 2000, 'Modern language testing at the turn of the century: assuring that what we count counts', *Language Testing*, vol. 17, no. 1, pp. 1–42.
- Bachman, LF & Palmer, AS, 1996, *Language testing in practice*, Oxford University Press, Oxford.
- Bachman, LF & Palmer, AS, 2010, *Language assessment in practice*, Oxford University Press, Oxford.
- Badger, R & Yan, X, 2009, 'The use of tactics and strategies by Chinese students in the listening component of IELTS', *IELTS Research Reports*, vol. 9, pp. 67–96, IELTS Australia, Canberra and British Council, London.
- Bentler, PM, 1985–2016, *EQS Version 6 for Windows* [Computer software], Multivariate Software, Encino, CA.
- Bentler, PM, 2006, *EQS structural equation program manual*, Multivariate Software, Encino, CA.
- Bi, NZ, 2014, *A multi-dimensional examination of strategic competence in the lexico-grammar test performance of Chinese EFL university students*, PhD thesis submitted to the University of Sydney, NSW, Australia.
- Björkman, M, 1994, 'Internal use theory: Calibration and resolution of confidence in general knowledge', *Organizational Behavior and Human Decision Processes*, vol. 58, no. 3, pp.386–405.
- Bol, L, Hacker, DJ, O'Shea, P & Allen, D, 2005, 'The influence of overt practice, achievement level, and explanatory style on calibration accuracy and performance', *Journal of Experimental Education*, vol. 73, no. 4, pp. 269–290.
- Bollen, KA, 1989, *Structural equation modeling with latent variables*, Wiley, New York.
- Bond, TG & Fox, CM, 2007, *Applying the Rasch model: Fundamental measurement in the human sciences*, 2nd edn, Routledge, New York.
- Buck, G, 2001, *Assessing listening*, Cambridge University Press, Cambridge.
- Butler, DL & Winne, PH, 1995, 'Feedback and self-regulated learning: A theoretical synthesis' *Review of Educational Research*, vol. 65, no. 3, pp. 245–281.
- Byrne, BM, 2006, *Structural equation modeling with EQS and EQS/Windows: Basic concepts, applications, and programming*, Psychology Press, New York and London.
- Caleon, I & Subramaniam, R, 2010, 'Development and application of a three-tier diagnostic test to assess students' understanding of waves', *International Journal of Science Education*, vol. 32, no. 7, pp. 939–961.
- Canale, M & Swain, M, 1980, 'Theoretical bases of communicative approaches to second language teaching and testing', *Applied Linguistics*, vol. 1, no. 1, pp. 1–47.

- Chapelle, CA, Enright, MK & Jamieson, J, 2010, 'Does an argument-based approach to validity make a difference?' *Educational Measurement: Issues and Practice*, vol. 29, no. 1, pp. 3–13.
- Chapelle, CA, Enright, MK & Jamieson, J, (eds) 2008, *Building a validity argument for the test of English as a foreign language*, Routledge, London.
- Chavez, M, 2001, *Gender in the language classroom*, McGraw Hill, Boston.
- Cleary, TJ, 2009, 'Monitoring trends and accuracy of self-efficacy beliefs during interventions: Advantages and potential applications to school-based settings'. *Psychology in the Schools*, vol. 46, no. 2, pp.154–171.
- Cleary, TJ, 2011, 'Emergence of self-regulated learning microanalysis: Historical overview, essential features, and implications for research and practice. In BJ Zimmerman & DH Schunk (eds.), *Handbook of self-regulation of learning and performance*, Routledge, New York and London.
- Cohen, AD & Upton, TA, 2007, 'I want to go back to the text': Response strategies on the reading subtest of the new TOEFL', *Language Testing*, vol. 24, no. 2, pp. 209–250.
- Cohen, AD, 2011, *Strategies in learning and using a second language*, 2nd edn, Longman, London.
- Cohen, AD, 2014, Using test-taking strategies for task development. In AJ Kunnan (ed.), *Companion to language assessment*, Wiley-Blackwell, London.
- Cohen, J, 1988, *Statistical power analysis for the behavioral sciences*, Sage, Newbury Park, CA.
- Cohen, J, 1992, 'A power primer', *Psychological Bulletin*, vol. 112, no. 1, pp. 155–159.
- Crisp, V, Sweiry, E, Ahmed, A & Pollitt, A, 2008, 'Tales of the expected: the influence of students' expectations on question validity and implications for writing exam questions', *Educational Research*, vol. 50, no. 1, pp. 95–115.
- Cullen, P, French, A & Jakeman, V, 2014, *The official Cambridge guide to IELTS for academic & general training*, Cambridge University Press, Cambridge.
- Davies, A, 2008, *Assessing academic English*, Cambridge University Press, Cambridge.
- Dinsmore, DL & Parkinson, MM, 2013, 'What are confidence judgments made of? Students' explanations for their confidence ratings and what that means for calibration', *Learning and Instruction*, vol. 24, no. 1, pp. 4–14.
- Dunlosky, J & Lipko, AR, 2007, 'Metacomprehension: A brief history and how to improve its accuracy', *Current Directions in Psychological Science*, vol. 16, no. 4, pp. 228–232.
- Dunlosky, J, Serra, MJ, Matvey, G & Rawson, KA, 2005, 'Second-order judgments about judgments of learning', *Journal of General Psychology*, vol. 132, no. 4, pp. 335–346.
- Edwards, W, 1967, *Statistical methods*, 2nd ed, Holt, Rinehart and Winston, New York.
- Efklides, A, 2008, 'Metacognition: Defining its facets and levels of functioning in relation to self-regulation and co-regulation', *European Psychologist*, vol. 13, no. 4, pp. 277–287.
- Efklides, A, 2011, 'Interactions of metacognition with motivation and affect in self-regulated learning: The MASRL model', *Educational Psychologist*, vol. 46, no. 1, pp. 6–25.
- Ehrlinger, J & Dunning, D, 2003, 'How chronic self-views influence (and potentially mislead) estimates of performance', *Journal of Personality and Social Psychology*, vol. 84, no. 1, pp. 5–17.
- Ehrlinger, J, Johnson, K, Banner, M, Dunning, D & Kruger, J, 2008, 'Why the unskilled are unaware: Further explanations of (absent) self-insight among the incompetent', *Organizational Behavior and Human Decision Processes*, vol. 105, no. 1, pp. 98–121.
- Epstein, W, Glenberg, AM & Bradley, MM, 1984, 'Coactivation and comprehension: Contribution of text variables to the illusion of knowing', *Memory & Cognition*, vol. 12, no. 4, pp. 355–360.
- Field, J, 2008, *Listening in the language classroom*, Cambridge University Press, Cambridge.
- Field, J, 2009, 'The cognitive validity of the lecture-based question in the IELTS Listening paper', *IELTS Research Reports*, vol. 9, pp. 17–65. IELTS Australia, Canberra and British Council, London.
- Field, J, 2013, Cognitive validity. In A Geranpayeh & L Taylor (eds), *Examining listening: Research and practice in assessing second language listening*, Cambridge University Press, Cambridge.
- Flavell, JH, 1971, 'First discussant's comments: What is memory development the development of?' *Human Development*, vol. 14, no. 4, pp. 272–278.
- Flavell, JH, 1992, Metacognition and cognitive monitoring: A new area of cognitive-developmental inquiry. In TO Nelson (ed.), *Metacognition: Core reading*, Allyn and Bacon, Boston.

- Forest, J & Altbach, PG, (eds) 2006, *International handbook of higher education, Vol.1*, Springer, Dordrecht.
- Fulcher, G, 2010, *Practical language testing*, Hodder Education, London.
- Gagné, ED, Yekovich, CW & Yekovich, FR, 1993, *The cognitive psychology of school learning*, HarperCollins College Publishers, New York, NY.
- Gass, S & Mackey, A, 2007, *Data elicitation for second and foreign language research*, Lawrence Erlbaum, Mahwah, NJ.
- Gigerenzer, G, Hoffrage, U & Kleinbölting, H, 1991, 'Probabilistic mental models. A Brunswikian theory of confidence', *Psychological Review*, vol. 98, no. 4, pp. 506–528.
- Glenberg, AM & Epstein, W, 1985, 'Calibration of comprehension', *Journal of Experimental Psychology: Learning, Memory, and Cognition*, vol. 11, no. 4, pp. 702–718.
- Glenberg, AM, Sanocki, T, Epstein, W & Morris, C, 1987, 'Enhancing calibration of comprehension', *Journal of Experimental Psychology: General*, vol. 116, no. 2, pp. 119–136.
- Goh, C, 2008, 'Metacognitive Instruction for Second Language Listening Development Theory, Practice and Research Implications', *RELC journal*, vol. 39, no. 2, pp. 188–213.
- Gutierrez, AP & Schraw, G, 2014, 'Effects of strategy training and incentives on students' performance, confidence, and calibration', *The Journal of Experimental Education*, vol. 7, no. 1, pp. 1–19.
- Hacker, DJ, Dunlosky, J & Graesser, AC, (eds) 2009, *Handbook of metacognition in education*. Routledge, New York.
- Hadwin, AF & Webster, EA, 2013, 'Calibration in goal setting: Examining the nature of judgments of confidence', *Learning and Instruction*, vol. 24, no. 1, pp. 37–47.
- Hattie, J, 2013. 'Calibration and confidence: Where to next?' *Learning and Instruction*, vol. 24, no. 1, pp. 62–66.
- Hertzog, C & Nesselroade, JR, 1987, 'Beyond autogressive models: Some implications of the trait-state distinction for the structural modeling of developmental changes', *Child Development*, vol. 58, no. 1, pp. 93–109.
- Hymes, D, 1972, On communicative competence. In JB Pride & J Holmes (eds), *Sociolinguistics: Selected readings*, Penguin, Harmondsworth, Middlesex.
- Jackson, SA & Kleitman, S, 2014, 'Individual differences in decision-making and confidence: Capturing decision tendencies in a fictitious medical test', *Metacognition Learning*, vol. 9, no. 1, pp. 25–49.
- Jiang, Y & Kleitman, S, 2015, 'Metacognition and Motivation: Links between Confidence, Self-protection and Self-enhancement', *Learning and Individual Differences*, vol. 37, no. 2, pp. 222–230.
- Johnson, JE & Bruce, AC, 2001, 'Calibration of subjective probability judgments in a naturalistic setting', *Organizational Behavior and Human Decision Processes*, vol. 85, no. 2, pp. 265–290.
- Juslin, P, 1994, 'The overconfidence phenomenon as a consequence of informal experimenter guided selection of almanac items', *Organizational Behavior and Human Decision Processes*, vol. 57, no. 2, pp. 226–246.
- Juslin, P, Winman, A & Persson, T, 1995, 'Can overconfidence be used as an indicator of reconstructive rather than retrieval processes?', *Cognition*, vol. 54, no. 1, pp. 99–130.
- Klassen, R, 2002. 'A question of calibration: A review of the self-efficacy beliefs of students with learning disabilities', *Learning Disability Quarterly*, vol. 25, no. 1, pp. 88–102.
- Kleitman, S & Costa, DSJ, 2014, 'The role of a novel formative assessment tool (Stats-mIQ) and individual differences in real-life academic performance', *Learning and Individual Differences*, vol. 29, pp. 150–161.
- Kleitman, S & Stankov, L, 2001, 'Ecological and person-oriented aspects of metacognitive processing in test taking', *Applied Cognitive Psychology*, vol. 15, no. 3, pp. 321–341.
- Kleitman, S & Stankov, L, 2007, 'Self-confidence and metacognitive processes', *Learning and Individual Differences*, vol. 17, no. 2, pp. 161–173.
- Kline, RB, 2011, *Principles and practice of structural equation modeling*, 3rd edn, Guilford Press, New York and London.
- Koriat, A, 2011, 'Subjective confidence in perceptual judgments: a test of the self-consistency model', *Journal of Experimental Psychology: General*, vol. 140, no. 1, pp. 117–139.

- Koriat, A, Lichtenstein, S & Fischhoff, B, 1980, 'Reasons for confidence', *Journal of Experimental Psychology: Human Learning and Memory*, vol. 6, no. 2, pp. 107–118.
- Kostons, D, van Gog, T & Paas, F, 2012, 'Training self-assessment and task selection skills: a cognitive approach to improving self-regulated learning', *Learning and Instruction*, vol. 22, no. 2, pp. 121–132.
- Kruger, J & Dunning, D, 1999, 'Unskilled and unaware of it: How difficulties in recognising one's own incompetence lead to inflated self-assessments', *Journal of Personality and Social Psychology*, vol. 77, no. 6, pp. 1121–1134.
- Kunnan, AJ, 1995, *Test taker characteristics and test performance: A structural modeling approach*, Cambridge University Press, Cambridge.
- Labuhn, AS, Zimmerman, BJ & Hasselhorn, M, 2010, 'Enhancing students' self-regulation and mathematics performance: The influence of feedback and self-evaluative standards', *Metacognition and Learning*, vol. 5, no. 2, pp. 173–194.
- Lin, L-M & Zabucky, KM, 1998, 'Calibration of comprehension: Research and implications for education and instruction', *Contemporary Educational Psychology*, vol. 23, no. 4, pp. 345–391.
- Little, TD, Cunningham, WA, Shahar, G & Widaman, KF, 2002, 'To parcel or not to parcel: Exploring the question, weighing the merits', *Structural Equation Modeling*, vol. 9, no. 2, pp. 151–173.
- Maki, RH & Serra, M, 1992, 'Role of practice tests in the accuracy of test predictions on text material', *Journal of Educational Psychology*, vol. 84, no. 2, pp. 200–210.
- Matsuno, S, 2009, 'Self-, peer-, and teacher-assessment in Japanese university EFL writing classrooms', *Language Testing*, vol. 26, no. 1, pp. 75–100.
- McNamara, T, 1996, *Measuring second language performance*, Longman, London and New York.
- Moore, DA & Healy, PJ, 2008, 'The trouble with overconfidence', *Psychological Review*, vol. 115, no. 2, pp. 502–517.
- Morony, S, Kleitman, S, Lee, YP & Stankov, L, 2013, 'Predicting achievement: Confidence versus self-efficacy, anxiety, and self-concept in Confucian and European countries', *International Journal of Education Research*, vol. 58, no. 1, pp. 79–96.
- Murphy, AH & Brown, BG, 1984, 'A comparative evaluation of objective and subjective weather forecasts in the United States', *Journal of Forecasting*, vol. 3, pp. 369–393.
- Murphy, AH & Winkler, RL, 1984, 'Probability forecasting in meteorology', *Journal of the American Statistical Association*, vol. 79, no. 387, pp. 489–500.
- Nelson, TO, 1994, Why investigate metacognition?. In J Metcalfe & A Shimamura (eds), *Metacognition: Knowing about knowing*, A Bradford Book, Cambridge, Massachusetts.
- Nelson, TO & Narens, L, 1990, Metamemory: A theoretical framework and new findings. In G Bower (ed.), *The psychology of learning and motivation*, Vol. 26, Academic, New York.
- Nietfield, JL & Schraw, G, 2002, 'The effect of knowledge and strategy training on monitoring accuracy', *Journal of Educational Research*, vol. 95, no. 3, pp. 131–142.
- Ockey, GJ, 2014, Exploratory factor analysis and structural equation modeling. In AJ Kunnan (ed.), *Companion to language assessment*, Wiley-Blackwell, London.
- Organization for Economic Cooperation and Development (OECD) 2012, *Education at a Glance 2012: OECD Indicators*, viewed 24 November, 2012, <<http://www.oecd.org/edu/eag2012.htm>>.
- Oscarson, M, 1997, Self-assessment of foreign and second language proficiency. In C Clapham & D Corson (eds), *Encyclopedia of language and education, Volume 7: Language testing and assessment*, Kluwer Academic Publishers, Dordrecht, Netherlands.
- Oscarson, M, 2014, Self-assessment in the classroom. In AJ Kunnan (ed.), *Companion to language assessment*, Wiley-Blackwell, London.
- Oxford, RL, 2011, *Teaching and researching language learning strategies*, Longman, Harlow, England.
- Pajares, F, 1996, 'Self-efficacy beliefs in academic settings', *Review of Educational Research*, vol. 66, no. 4, pp. 543–578.

- Pallant, J, 2010, *SPSS survival manual: A step by step guide to data analysis using SPSS*, Open University Press, Buckingham.
- Pallier, G, 2003, 'Gender differences in the self-assessment accuracy on cognitive tasks', *Sex Roles*, vol. 48, nos. 5-6, pp. 265–276.
- Pallier, G, Wilkinson, R, Danthiir, V, Kleitman, S, Knezevic, G & Stankov, L et al., 2002, 'Individual differences in the realism of confidence judgments', *Journal of General Psychology*, vol. 129, no. 3, pp. 257–300.
- Phakiti, A, 2003a, 'A closer look at gender and strategy use in L2 reading', *Language Learning*, vol. 53, no. 4, pp. 649–702.
- Phakiti, A, 2003b, 'A closer look at the relationship of cognitive and metacognitive strategy use to EFL reading achievement test performance', *Language Testing*, vol. 20, no. 1, pp. 26–56.
- Phakiti, A, 2005, 'An empirical investigation into the nature of and factors affecting test takers' calibration within the context of an English Placement Test (EPT)', *Spain Fellow Working Papers in Second or Foreign Language Assessment*, vol. 3, pp. 27–46.
- Phakiti, A, 2007a, 'On the nature of L2 test-takers' calibration in a reading test. In C Gitsaki (ed.), *Language and languages: Global and local tensions*, Cambridge Scholars Publishing, Newcastle, UK.
- Phakiti, A, 2007b, *Strategic competence and EFL reading test performance: A structural equation modeling approach*, Peter Lang, Frankfurt am Main.
- Phakiti, A, 2008a, 'Construct validation of Bachman and Palmer's (1996) strategic competence model over time in EFL reading tests', *Language Testing*, vol. 25, no. 2, pp. 237–272.
- Phakiti, A, 2008b, 'Strategic competence as a fourth-order factor model: A structural equation modeling approach', *Language Assessment Quarterly*, vol. 5, no. 1, pp. 20–42.
- Phakiti, A, 2016 'Structural models of calibration, confidence and cognitive and metacognitive strategy use in an English placement test', *Language Assessment Quarterly*, vol. 13, no. 2, pp. 75–108.
- Purpura, JE, 1999, *Learner strategy use and performance on language tests: A structural equation modeling approach*, Cambridge University Press, Cambridge.
- Purpura, JE, 2014, 'Cognition and language assessment. In AJ Kunnan (ed.), *The companion to language assessment*, Wiley-Blackwell, Oxford, UK.
- Reise, SP & Widaman, KF, 1999, 'Assessing the fit of measurement models at the individual level: A comparison of item response theory and covariance structure approaches', *Psychological Methods*, vol. 4, no. 1, pp. 2–21.
- Roderer, T & Roebbers, CM, 2014, 'Can you see me thinking (about my answers)? Using eye-tracking to illuminate developmental differences in monitoring and control skills and their relation to performance' *Metacognition Learning*, vol. 9, no. 1, pp. 1–23.
- Ross, S, 1998, 'Self-assessment in second language testing: A meta-analysis and analysis of experiential factors', *Language Testing*, vol. 15, no. 1, pp. 1–20.
- Rost, M, 2011, *Teaching and researching listening*, 2nd edn, Longman, London.
- Schmidt, R, 1995, 'Consciousness and foreign language learning: A tutorial on the role of attention and awareness in learning. In R Schmidt (ed.), *Attention and awareness in foreign language learning*, University of Hawai'i Press, Honolulu.
- Schneider, SL, 1995, 'Item difficulty, discrimination, and the confidence-frequency effect in a categorical judgment task', *Organizational Behavior and Human Decision Processes*, vol. 61, no. 2, pp. 148–167.
- Schraw, G, 2009, 'A conceptual analysis of five measures of metacognitive monitoring', *Metacognition and Learning*, vol. 4, no. 1, pp. 33–45.
- Schraw, G, Kuch, F & Gutierrez, AP, 2013, 'Measure for measure: Calibrating ten commonly used calibration scores', *Learning and Instruction*, vol. 24, no. 1, pp. 48–57.
- Schraw, G, Kuch, F, Gutierrez, AP & Richmond, AS, 2014, 'Exploring a three-level model of calibration accuracy', *Journal of Educational Psychology*, online first publication, doi: 10.1037/a0036653.
- Schumacker, RE & Lomax, RG, 2010, *A beginner's guide to structural equation modeling*, 3rd edn, Taylor and Francis, Mahwah, NJ.
- Shaw, SD & Weir, CJ, 2007, *Examining writing: Research and practice in assessing second language writing*, Cambridge University Press, Cambridge.

- Soll, JB, 1996, 'Determinants of overconfidence and miscalibration: The role of random error and ecological structure', *Organizational Behavior and Human Decision Processes*, vol. 65, no. 2, pp. 117–137.
- Song, X, 2004, *Language learning strategy use and language performance for Chinese learners of English*, Unpublished master's thesis, Queen's University, Kingston, Ontario, Canada.
- Stankov, L & Crawford, JD, 1996, 'Confidence judgments in studies of individual differences', *Personality and Individual Differences*, vol. 21, no. 6, pp. 971–986.
- Stankov, L & Crawford, JD, 1997, 'Self-confidence and performance on tests of cognitive activities', *Intelligence*, vol. 25, no. 1, pp. 93–109.
- Stankov, L & Lee, J 2008, 'Confidence and cognitive performance', *Journal of Educational Psychology*, vol. 100, no. 4, pp. 961–976.
- Stankov, L & Lee, J, 2014a, 'Overconfidence across world regions', *Journal of Cross-Cultural Psychology*, vol. 45, no. 5, pp. 821–837.
- Stankov, L & Lee, J, 2014b, 'Quest for the best non-cognitive predictor of academic achievement', *Educational Psychology: An International Journal of Experimental Educational Psychology*, vol. 34, no. 1, pp. 1–8.
- Stankov, L, Lee, J & Paek, I, 2009, 'Realism of confidence judgments', *European Journal of Psychological Assessment*, vol. 25, no. 2, pp. 123–130.
- Stankov, L, Lee, J, Luo, W & Hogan, DJ, 2012, 'Confidence: A better predictor of academic achievement than self-efficacy, self-concept and anxiety?', *Learning and Individual Differences*, vol. 22, no. 6, pp. 747–758.
- Stankov, L, Pallier, G, Danthiir, V & Morony, S, 2012, 'Perceptual underconfidence: A conceptual illusion?', *European Journal of Psychological Assessment*, vol. 28, no. 3, pp. 190–200.
- Stone, NJ, 2000, 'Exploring the relationship between calibration and self-regulated learning', *Educational Psychology Review*, vol. 12, no. 4, pp. 437–475.
- Suantak, L, Bolger, F & Ferrell, WR, 1996, 'The hard-easy effect in subjective probability calibration', *Organizational Behavior and Human Decision Processes*, vol. 67, no. 2, pp. 201–221.
- Tobias, S & Everson, H, 2009, The importance of knowing what you know: A knowledge monitoring framework for studying metacognition in education. In DJ Hacker, J Dunlosky & AC Graesser (eds), *Handbook of metacognition in education*, New York, Routledge.
- Trofimovich, P, Isaacs, T, Kennedy, S, Saito, K & Crowther, D, 2016, 'Flawed self-assessment: Investigating self- and other-perception of second language speech', *Bilingual: Language and Cognition*, vol. 19, no.1, pp. 122–140.
- van Loon, M, de Bruin, ABH, van Gog, T, & van Merriënboer, J, 2013, 'Activation of inaccurate prior knowledge affects primary-school students' metacognitive judgments and calibration', *Learning and Instruction*, vol. 24, no. 1, pp. 15–25.
- Vandergrift, L & Baker, S, 2015, 'Learner Variables in Second Language Listening Comprehension: An Exploratory Path Analysis', *Language Learning*, vol. 65, no. 2, pp. 390–416.
- Vandergrift, L & Goh, C, 2012, *Teaching and learning second language listening: metacognition in action*, Routledge, New York.
- Vandergrift, L, 2015, Researching listening. In B Paltridge & A Phakiti (eds), *Research methods in applied linguistics: A practical resource*, London, Bloomsbury.
- VanPatten, B, 1994, 'Evaluating the role of consciousness in second language acquisition: Terms, linguistic features and research methodology', *AILA Review*, vol. 11, pp. 27–36.
- Weaver, CA III & Bryant, DS, 1995, 'Monitoring of comprehension: The role of text difficulty in metamemory for narrative and expository text', *Memory & Cognition*, vol. 23, no. 1, pp. 12–22.
- Weir, CJ, 2005, *Language testing and validation: An evidence-based approach*, Palgrave Macmillan, Hampshire.
- Winke, P & Lim, H, 2014, 'The effects of testwiseness and test-taking anxiety on L2 listening test performance: A visual (eye-tracking) and attentional investigation', *IELTS Research Report Series*, vol. 3, pp. 1–30, IELTS Australia Pty Limited, Canberra.
- Winke, P, 2014, 'Testing hypotheses about language learning using structural equation modeling', *Annual Review of Applied Linguistics*, vol. 34, pp. 102–122.

Yates, JF, Lee, J & Shinotsuka, H, 1996, 'Beliefs about overconfidence, including its cross-national variation', *Organizational Behavior and Human Decision Processes*, vol. 65, no. 2, pp. 138–147.

Zhang, L & Zhang, LJ, 2013, 'Relationships between Chinese college test takers' strategy use and EFL reading test performance: A structural equation modeling approach', *RELC Journal*, vol. 44, no. 1, pp. 35–57.

Zhang, L, Gao, C & Kunnan, AJ, 2014, 'Analysis of test takers' metacognitive and cognitive strategy use and EFL reading test performance: A multi-sample SEM approach', *Language Assessment Quarterly*, vol. 11, no. 1, pp. 76–102.

Zimmerman, BJ, 1994, Dimensions of academic self-regulation: A conceptual framework for education. In DH Schunk & BJ Zimmerman (eds), *Self-regulation of learning and performance: Issues and educational applications*, Lawrence Erlbaum Associates, Hillsdale, NJ.

APPENDIX 1: RESEARCH INSTRUMENTS

A1.1 General instructions

Overall Instructions

1. Answer the pre-listening questionnaire in **IELTS Listening tests**.
2. Take an IELTS listening test. This test measures your ability to comprehend spoken English. There are 4 sections in this test, each with ten questions (Total of 40 questions). You will hear the recording ONCE only and answer the questions as you listen.
Section 1: A conversation between two people in social language use
Section 2: A talk about social issues
Section 3: A conversation between up to four people in situations related to educational or training contexts
Section 4: A talk in situations related to educational or training contexts
3. Immediately after you answer each question, indicate the level of your confidence in the correctness of your answer in percentage (0%, 25%, 50%, 75%, 90% or 100%) in your test.
Choose your first impression about your confidence.
4. You will be given about 3 minutes after each test section to transfer your answers and confidence from your test into the given answer sheet.
5. At the end of this test, please answer the post-listening questionnaire in **this IELTS Listening test**.

A1.2 Background questionnaire

Name: _____ Gender: Male Female

Age: _____ Nationality: _____ Native Language: _____

Have you ever taken an official IELTS before: Yes No

What is your latest overall IELTS score band (if any)? _____

What is your latest IELTS Listening score (if any)? _____

Degree (if applicable): Undergraduate Postgraduate

Faculty (if applicable): _____

A1.3 Trait strategy use and IELTS listening difficulty questionnaire

Pre-Questionnaire (IELTS Listening in general)

Directions: Read each statement and indicate how you **generally** think in an **IELTS Listening test**. Choose **1** (never), **2** (rarely), **3** (sometimes), **4** (often), **5** (usually) or **6** (always) on each statement that best describes how you think. Cross your answer (**X**).

No	Your thinking	1	2	3	4	5	6
1.	I know what I have to do in an IELTS Listening test.	1	2	3	4	5	6
2.	I make sure I clarify what the test tasks require me to do.	1	2	3	4	5	6
3.	I look up all test sections to see what I will have to complete.	1	2	3	4	5	6
4.	I try to understand test tasks that I know I am not good at when possible.	1	2	3	4	5	6
5.	I think ahead what I will hear next while I listen and answer questions.	1	2	3	4	5	6
6.	I try to figure out what the speaker(s) means or tries to say.	1	2	3	4	5	6
7.	I look up questions or tasks as I listen.	1	2	3	4	5	6
8.	I try to retain what I hear in my memory including taking notes.	1	2	3	4	5	6
9.	I use my prior knowledge or experience to help me listen.	1	2	3	4	5	6
10.	I reread test questions or tasks as required.	1	2	3	4	5	6
11.	I know which information is more or less important.	1	2	3	4	5	6
12.	I guess meanings of unknown words.	1	2	3	4	5	6
13.	I know how much time I should spend to answer questions.	1	2	3	4	5	6
14.	I tell myself to concentrate on the test tasks.	1	2	3	4	5	6
15.	I check my answers against test questions.	1	2	3	4	5	6
16.	I know the time limitation and constraint in the test.	1	2	3	4	5	6
17.	I notice when I am not sure I understand what I hear.	1	2	3	4	5	6
18.	I evaluate my performance as I move along the test tasks.	1	2	3	4	5	6
19.	I am aware of how well I am doing in the test.	1	2	3	4	5	6
20.	I immediately correct mistakes or answers when found.	1	2	3	4	5	6
21.	I check whether my answers are spelt correctly.	1	2	3	4	5	6
22.	I double-check my test performance.	1	2	3	4	5	6
	1 = Not at all true 2 = Not true 3 = Neither 4 = True 5 = Absolutely true	1	2	3	4	5	
23.	I find it difficult to find correct answers to questions.	1	2	3	4	5	
24.	I find it difficult to understand what I hear.	1	2	3	4	5	
25.	I find it difficult to remember or recall answers to questions.	1	2	3	4	5	
26.	I find it difficult to spell answers correctly.	1	2	3	4	5	
27.	I find it difficult to concentrate in IELTS Listening tests.	1	2	3	4	5	

A1.4 Practice IELTS Listening test questions with appraisal confidence rating

PRACTICE QUESTIONS FOR CONFIDENCE RATING

Questions 1-5

Complete the notes below.

Write **NO MORE THAN THREE WORDS** for each answer.

Rate your **CONFIDENCE** as soon as you answer each.

Transport from Airport to Milton

Example

Answer

Distance 147 miles

0%	25%	50%	75%	90%	100%
----	-----	-----	-----	-----	-----------------

Options:

- Car hire
- Don't want to drive

• **1** _____

0%	25%	50%	75%	90%	100%
----	-----	-----	-----	-----	------

- -expensive

• Greyhound bus

- \$15 single, \$27.50 return

- direct to the **2** _____

0%	25%	50%	75%	90%	100%
----	-----	-----	-----	-----	------

- long **3** _____

0%	25%	50%	75%	90%	100%
----	-----	-----	-----	-----	------

• Airport Shuttle

- **4** _____ service

0%	25%	50%	75%	90%	100%
----	-----	-----	-----	-----	------

- every 2 hours

- \$35 single, \$65 return

- need to **5** _____

0%	25%	50%	75%	90%	100%
----	-----	-----	-----	-----	------

A1.5 The IELTS Listening test

SECTION 1

Questions 1-10

Questions 1-6

Complete the form below.

Write **NO MORE THAN THREE WORDS AND/OR NUMBER** for each answer. Rate your **CONFIDENCE (X)** as soon as you have answered each question.

HOMESTAY APPLICATION	
<i>Example</i>	<i>Answer</i>
Surname:	<u>Yuichini</u> 0% 25% 50% 75% 90% 100%
First name:	1 _____ 0% 25% 50% 75% 90% 100%
Sex:	<u>female</u> Nationality: <u>Japanese</u>
Passport number:	2 _____ 0% 25% 50% 75% 90% 100%
Age:	<u>28 years</u>
Present address:	<u>Room 21C, Willow College</u>
Length of homestay:	<u>approx 3</u> _____ 0% 25% 50% 75% 90% 100%
Course enrolled in:	4 _____ 0% 25% 50% 75% 90% 100%
Family preferences:	<u>no 5</u> _____ 0% 25% 50% 75% 90% 100%
	<u>no objection to 6</u> _____ 0% 25% 50% 75% 90% 100%

Questions 7-10

Answer the questions below.

Which **NO MORE THAN TWO WORDS** for each answer. Rate your **CONFIDENCE (X)** as soon as you have answered each question.

7. What does the student particularly like to eat? _____
0% 25% 50% 75% 90% 100%

8. What sport does the student play? _____
0% 25% 50% 75% 90% 100%

9. What mode of transport does the student prefer? _____
0% 25% 50% 75% 90% 100%

10. When will the student find out her homestay address? _____
0% 25% 50% 75% 90% 100%

Answer Sheet: Transfer your **answers** and **confidence (X)**. You may change your confidence at this stage.

1		<input type="checkbox"/> 0% <input type="checkbox"/> 25% <input type="checkbox"/> 50% <input type="checkbox"/> 75% <input type="checkbox"/> 90% <input type="checkbox"/> 100%
2		<input type="checkbox"/> 0% <input type="checkbox"/> 25% <input type="checkbox"/> 50% <input type="checkbox"/> 75% <input type="checkbox"/> 90% <input type="checkbox"/> 100%
3		<input type="checkbox"/> 0% <input type="checkbox"/> 25% <input type="checkbox"/> 50% <input type="checkbox"/> 75% <input type="checkbox"/> 90% <input type="checkbox"/> 100%
4		<input type="checkbox"/> 0% <input type="checkbox"/> 25% <input type="checkbox"/> 50% <input type="checkbox"/> 75% <input type="checkbox"/> 90% <input type="checkbox"/> 100%
5		<input type="checkbox"/> 0% <input type="checkbox"/> 25% <input type="checkbox"/> 50% <input type="checkbox"/> 75% <input type="checkbox"/> 90% <input type="checkbox"/> 100%
6		<input type="checkbox"/> 0% <input type="checkbox"/> 25% <input type="checkbox"/> 50% <input type="checkbox"/> 75% <input type="checkbox"/> 90% <input type="checkbox"/> 100%
7		<input type="checkbox"/> 0% <input type="checkbox"/> 25% <input type="checkbox"/> 50% <input type="checkbox"/> 75% <input type="checkbox"/> 90% <input type="checkbox"/> 100%
8		<input type="checkbox"/> 0% <input type="checkbox"/> 25% <input type="checkbox"/> 50% <input type="checkbox"/> 75% <input type="checkbox"/> 90% <input type="checkbox"/> 100%
9		<input type="checkbox"/> 0% <input type="checkbox"/> 25% <input type="checkbox"/> 50% <input type="checkbox"/> 75% <input type="checkbox"/> 90% <input type="checkbox"/> 100%
10		<input type="checkbox"/> 0% <input type="checkbox"/> 25% <input type="checkbox"/> 50% <input type="checkbox"/> 75% <input type="checkbox"/> 90% <input type="checkbox"/> 100%

Your overall confidence in this section: 0% 25% 50% 75% 90% 100%

SECTION 2

Questions 11-20

Questions 11-13

Choose the correct letter, **A**, **B** or **C**.

Rate your **CONFIDENCE (X)** as soon as you have answered each question.

THE HISTORY OF ROSEWOOD HOUSE

11 When the writer Sebastian George first saw Rosewood House, he

A thought he might rent it.

B felt it was too expensive for him.

C was unsure whether to buy it.

0%	25%	50%	75%	90%	100%
----	-----	-----	-----	-----	------

12 Before buying the house, George had

A experienced severe family problems.

B struggled to become a successful author.

C suffered a serious illness.

0%	25%	50%	75%	90%	100%
----	-----	-----	-----	-----	------

13 According to the speaker, George viewed Rosewood House as

A a rich source of material for his books.

B a way to escape from his work.

C a typical building of the region

0%	25%	50%	75%	90%	100%
----	-----	-----	-----	-----	------

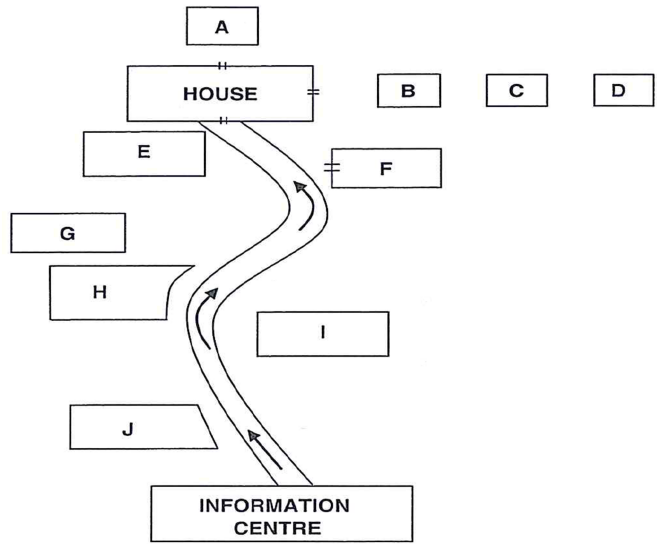
Questions 14 -17

Label the map below.

Write the correct letter, **A-J**, next to questions 14-17.

Rate your **CONFIDENCE (X)** as soon as you have answered each question.

ROSEWOOD HOUSE AND GARDENS



- 14 Pear Alley _____
- 15 Mulberry Garden _____
- 16 Shop _____
- 17 Tea Room _____

0%	25%	50%	75%	90%	100%
0%	25%	50%	75%	90%	100%
0%	25%	50%	75%	90%	100%
0%	25%	50%	75%	90%	100%

Questions 18 - 20

Complete the sentence below

Write **ONE WORD ONLY** for each answer. Rate your **CONFIDENCE (X)** as soon as you have answered each question.

- 18 You can walk through the _____ that goes along the river bank.

0%	25%	50%	75%	90%	100%
----	-----	-----	-----	-----	------

- 19 You can go over the _____ and then into a wooded area.

0%	25%	50%	75%	90%	100%
----	-----	-----	-----	-----	------

- 20 On your way back, you could also go up to the _____.

0%	25%	50%	75%	90%	100%
----	-----	-----	-----	-----	------

Answer Sheet: Transfer your **answers** and **confidence (X)**. You may change your confidence at this stage.

11		<input type="checkbox"/> 0% <input type="checkbox"/> 25% <input type="checkbox"/> 50% <input type="checkbox"/> 75% <input type="checkbox"/> 90% <input type="checkbox"/> 100%
12		<input type="checkbox"/> 0% <input type="checkbox"/> 25% <input type="checkbox"/> 50% <input type="checkbox"/> 75% <input type="checkbox"/> 90% <input type="checkbox"/> 100%
13		<input type="checkbox"/> 0% <input type="checkbox"/> 25% <input type="checkbox"/> 50% <input type="checkbox"/> 75% <input type="checkbox"/> 90% <input type="checkbox"/> 100%
14		<input type="checkbox"/> 0% <input type="checkbox"/> 25% <input type="checkbox"/> 50% <input type="checkbox"/> 75% <input type="checkbox"/> 90% <input type="checkbox"/> 100%
15		<input type="checkbox"/> 0% <input type="checkbox"/> 25% <input type="checkbox"/> 50% <input type="checkbox"/> 75% <input type="checkbox"/> 90% <input type="checkbox"/> 100%
16		<input type="checkbox"/> 0% <input type="checkbox"/> 25% <input type="checkbox"/> 50% <input type="checkbox"/> 75% <input type="checkbox"/> 90% <input type="checkbox"/> 100%
17		<input type="checkbox"/> 0% <input type="checkbox"/> 25% <input type="checkbox"/> 50% <input type="checkbox"/> 75% <input type="checkbox"/> 90% <input type="checkbox"/> 100%
18		<input type="checkbox"/> 0% <input type="checkbox"/> 25% <input type="checkbox"/> 50% <input type="checkbox"/> 75% <input type="checkbox"/> 90% <input type="checkbox"/> 100%
19		<input type="checkbox"/> 0% <input type="checkbox"/> 25% <input type="checkbox"/> 50% <input type="checkbox"/> 75% <input type="checkbox"/> 90% <input type="checkbox"/> 100%
20		<input type="checkbox"/> 0% <input type="checkbox"/> 25% <input type="checkbox"/> 50% <input type="checkbox"/> 75% <input type="checkbox"/> 90% <input type="checkbox"/> 100%

Your overall confidence in this section: 0% 25% 50% 75% 90% 100%

SECTION 3

Questions 21-30

Complete the notes below.

Write **NO MORE THAN TWO WORDS AND/OR A NUMBER** for each answer.

Rate your **CONFIDENCE (X)** as soon as you have answered each question.

Study Skills Tutorial – Caroline Benning

- Dissertation topic:** the 21 _____ 0% 25% 50% 75% 90% 100%
- Strength:**
- 22 _____ 0% 25% 50% 75% 90% 100%
 - computer modeling
- Weaknesses:**
- lack of background information
 - poor 23 _____ 0% 25% 50% 75% 90% 100%

Possible Strategy	Benefits	Problems
peer group discussion	increases 24 _____ <input type="checkbox"/> 0% <input type="checkbox"/> 25% <input type="checkbox"/> 50% <input type="checkbox"/> 75% <input type="checkbox"/> 90% <input type="checkbox"/> 100%	dissertations tend to contain the same 25 _____ <input type="checkbox"/> 0% <input type="checkbox"/> 25% <input type="checkbox"/> 50% <input type="checkbox"/> 75% <input type="checkbox"/> 90% <input type="checkbox"/> 100%
use the 26 _____ service <input type="checkbox"/> 0% <input type="checkbox"/> 25% <input type="checkbox"/> 50% <input type="checkbox"/> 75% <input type="checkbox"/> 90% <input type="checkbox"/> 100%	provides structured programme	limited 27 _____ <input type="checkbox"/> 0% <input type="checkbox"/> 25% <input type="checkbox"/> 50% <input type="checkbox"/> 75% <input type="checkbox"/> 90% <input type="checkbox"/> 100%
consult study skills books	are a good source of reference	can be too 28 _____ <input type="checkbox"/> 0% <input type="checkbox"/> 25% <input type="checkbox"/> 50% <input type="checkbox"/> 75% <input type="checkbox"/> 90% <input type="checkbox"/> 100%

- Recommendations:**
- use a card index
 - read all notes 29 _____ 0% 25% 50% 75% 90% 100%
- Next tutorial date:** 30 _____ January 0% 25% 50% 75% 90% 100%

Answer Sheet: Transfer your **answers** and **confidence (X)**. You may change your confidence at this stage.

21		<input type="checkbox"/> 0% <input type="checkbox"/> 25% <input type="checkbox"/> 50% <input type="checkbox"/> 75% <input type="checkbox"/> 90% <input type="checkbox"/> 100%
22		<input type="checkbox"/> 0% <input type="checkbox"/> 25% <input type="checkbox"/> 50% <input type="checkbox"/> 75% <input type="checkbox"/> 90% <input type="checkbox"/> 100%
23		<input type="checkbox"/> 0% <input type="checkbox"/> 25% <input type="checkbox"/> 50% <input type="checkbox"/> 75% <input type="checkbox"/> 90% <input type="checkbox"/> 100%
24		<input type="checkbox"/> 0% <input type="checkbox"/> 25% <input type="checkbox"/> 50% <input type="checkbox"/> 75% <input type="checkbox"/> 90% <input type="checkbox"/> 100%
25		<input type="checkbox"/> 0% <input type="checkbox"/> 25% <input type="checkbox"/> 50% <input type="checkbox"/> 75% <input type="checkbox"/> 90% <input type="checkbox"/> 100%
26		<input type="checkbox"/> 0% <input type="checkbox"/> 25% <input type="checkbox"/> 50% <input type="checkbox"/> 75% <input type="checkbox"/> 90% <input type="checkbox"/> 100%
27		<input type="checkbox"/> 0% <input type="checkbox"/> 25% <input type="checkbox"/> 50% <input type="checkbox"/> 75% <input type="checkbox"/> 90% <input type="checkbox"/> 100%
28		<input type="checkbox"/> 0% <input type="checkbox"/> 25% <input type="checkbox"/> 50% <input type="checkbox"/> 75% <input type="checkbox"/> 90% <input type="checkbox"/> 100%
29		<input type="checkbox"/> 0% <input type="checkbox"/> 25% <input type="checkbox"/> 50% <input type="checkbox"/> 75% <input type="checkbox"/> 90% <input type="checkbox"/> 100%
30		<input type="checkbox"/> 0% <input type="checkbox"/> 25% <input type="checkbox"/> 50% <input type="checkbox"/> 75% <input type="checkbox"/> 90% <input type="checkbox"/> 100%

Your overall confidence in this section: 0% 25% 50% 75% 90% 100%

SECTION 4

Questions 31-40

Questions 31-36

Choose the correct letter, A, B or C.

Rate your **CONFIDENCE (X)** as soon as you have answered each question.

Wildlife in City Gardens

- 31** What led the group to choose their topic?
- A** They were concerned about the decline of one species.
 - B** They were interested in the effects of city growth.
 - C** They wanted to investigate a recent phenomenon.
- 32** The exact proportion of land devoted to private gardens was confirmed by
- A** consulting some official documents.
 - B** taking large-scale photos.
 - C** discussing with town surveyors.
- 33** The group asked garden owners to
- A** take part in formal interviews.
 - B** keep a record of animals they saw.
 - C** get in contact when they saw a rare species.
- 34** The group made their observations in gardens
- A** which has a large number of animal species.
 - B** which they considered to be representative.
 - C** which had stable populations of rare animals.
- 35** The group did extensive reading on
- A** wildlife problems in rural areas.
 - B** urban animal populations.
 - C** current gardening practices.
- 36** The speaker focuses on three animal species because
- A** a lot of data been obtained about them.
 - B** the group were most interested in them.
 - C** the best indicated general trends.

0%	25%	50%	75%	90%	100%
----	-----	-----	-----	-----	------

0%	25%	50%	75%	90%	100%
----	-----	-----	-----	-----	------

0%	25%	50%	75%	90%	100%
----	-----	-----	-----	-----	------

0%	25%	50%	75%	90%	100%
----	-----	-----	-----	-----	------

0%	25%	50%	75%	90%	100%
----	-----	-----	-----	-----	------

0%	25%	50%	75%	90%	100%
----	-----	-----	-----	-----	------

Questions 37-40

Complete the table below.

Write **ONE WORD ONLY** for each answer. Rate your **CONFIDENCE (X)** as soon as you answer each.

Animals	Reason for population increase in gardens	Comments
37 _____ <input type="checkbox"/> 0% <input type="checkbox"/> 25% <input type="checkbox"/> 50% <input type="checkbox"/> 75% <input type="checkbox"/> 90% <input type="checkbox"/> 100%	Suitable stretches of water	Massive increase in urban population
Hedgehogs	Safer from 38 _____ when in cities <input type="checkbox"/> 0% <input type="checkbox"/> 25% <input type="checkbox"/> 50% <input type="checkbox"/> 75% <input type="checkbox"/> 90% <input type="checkbox"/> 100%	Easy to 39 _____ Them accurately <input type="checkbox"/> 0% <input type="checkbox"/> 25% <input type="checkbox"/> 50% <input type="checkbox"/> 75% <input type="checkbox"/> 90% <input type="checkbox"/> 100%
Song thrushes	- a variety of 40 _____ to eat <input type="checkbox"/> 0% <input type="checkbox"/> 25% <input type="checkbox"/> 50% <input type="checkbox"/> 75% <input type="checkbox"/> 90% <input type="checkbox"/> 100% - More nesting places available	Large survey starting soon

Answer Sheet: Transfer your **answers** and **confidence (X)**. You may change your confidence at this stage.

31	<input type="checkbox"/> 0% <input type="checkbox"/> 25% <input type="checkbox"/> 50% <input type="checkbox"/> 75% <input type="checkbox"/> 90% <input type="checkbox"/> 100%
32	<input type="checkbox"/> 0% <input type="checkbox"/> 25% <input type="checkbox"/> 50% <input type="checkbox"/> 75% <input type="checkbox"/> 90% <input type="checkbox"/> 100%
33	<input type="checkbox"/> 0% <input type="checkbox"/> 25% <input type="checkbox"/> 50% <input type="checkbox"/> 75% <input type="checkbox"/> 90% <input type="checkbox"/> 100%
34	<input type="checkbox"/> 0% <input type="checkbox"/> 25% <input type="checkbox"/> 50% <input type="checkbox"/> 75% <input type="checkbox"/> 90% <input type="checkbox"/> 100%
35	<input type="checkbox"/> 0% <input type="checkbox"/> 25% <input type="checkbox"/> 50% <input type="checkbox"/> 75% <input type="checkbox"/> 90% <input type="checkbox"/> 100%
36	<input type="checkbox"/> 0% <input type="checkbox"/> 25% <input type="checkbox"/> 50% <input type="checkbox"/> 75% <input type="checkbox"/> 90% <input type="checkbox"/> 100%
37	<input type="checkbox"/> 0% <input type="checkbox"/> 25% <input type="checkbox"/> 50% <input type="checkbox"/> 75% <input type="checkbox"/> 90% <input type="checkbox"/> 100%
38	<input type="checkbox"/> 0% <input type="checkbox"/> 25% <input type="checkbox"/> 50% <input type="checkbox"/> 75% <input type="checkbox"/> 90% <input type="checkbox"/> 100%
39	<input type="checkbox"/> 0% <input type="checkbox"/> 25% <input type="checkbox"/> 50% <input type="checkbox"/> 75% <input type="checkbox"/> 90% <input type="checkbox"/> 100%
40	<input type="checkbox"/> 0% <input type="checkbox"/> 25% <input type="checkbox"/> 50% <input type="checkbox"/> 75% <input type="checkbox"/> 90% <input type="checkbox"/> 100%

Your overall confidence in this section: 0% 25% 50% 75% 90% 100%

A1.6 State strategy use and IELTS listening difficulty questionnaire

Post-Questionnaire (This IELTS Listening test)

Directions: Read each statement and indicate how you generally thought in **this IELTS Listening test**. Choose **1** (never), **2** (rarely), **3** (sometimes), **4** (often), **5** (usually) or **6** (always) on each statement that best describes how you think. Cross your answer (**X**).

No	Your thinking	1	2	3	4	5	6
1.	I knew what I had to do in this IELTS Listening test.	1	2	3	4	5	6
2.	I made sure I clarified what the test tasks required me to do.	1	2	3	4	5	6
3.	I looked up all test sections to see what I would have to complete.	1	2	3	4	5	6
4.	I tried to understand test tasks that I knew I was not good at when possible.	1	2	3	4	5	6
5.	I thought ahead what I would hear next while I listened and answered questions.	1	2	3	4	5	6
6.	I tried to figure out what the speaker(s) meant or tried to say.	1	2	3	4	5	6
7.	I looked up questions or tasks as I listened.	1	2	3	4	5	6
8.	I tried to retain what I heard in my memory including taking notes.	1	2	3	4	5	6
9.	I used my prior knowledge or experience to help me listen.	1	2	3	4	5	6
10.	I reread test questions or tasks as required.	1	2	3	4	5	6
11.	I knew which information was more or less important.	1	2	3	4	5	6
12.	I guessed meanings of unknown words.	1	2	3	4	5	6
13.	I knew how much time I should spend to answer questions.	1	2	3	4	5	6
14.	I told myself to concentrate on the test tasks.	1	2	3	4	5	6
15.	I checked my answers against test questions.	1	2	3	4	5	6
16.	I knew the time limitation and constraint in the test.	1	2	3	4	5	6
17.	I noticed when I was not sure I understood what I heard.	1	2	3	4	5	6
18.	I evaluated my performance as I moved along the test tasks.	1	2	3	4	5	6
19.	I was aware of how well I was doing in the test.	1	2	3	4	5	6
20.	I immediately corrected mistakes or answers when found.	1	2	3	4	5	6
21.	I checked whether my answers are spelt correctly.	1	2	3	4	5	6
22.	I double-checked my test performance.	1	2	3	4	5	6
	1 = Not at all true 2 = Not true 3 = Neither 4 = True 5 = Absolutely true	1	2	3	4	5	
23.	I found it difficult to find correct answers to questions.	1	2	3	4	5	
24.	I found it difficult to understand what I heard.	1	2	3	4	5	
25.	I found it difficult to remember or recall answers to questions.	1	2	3	4	5	
26.	I found it difficult to spell answers correctly.	1	2	3	4	5	
27.	I found it difficult to concentrate in this IELTS Listening test.	1	2	3	4	5	

Thank you for your cooperation and contribution to this study.

A1.7 Answer keys

IELTS Listening Test

Answer Sheet: Transfer your **answers** and **confidence (X)**. You may change your confidence at this stage.

1	Keiko	<input type="checkbox"/> 0% <input type="checkbox"/> 25% <input type="checkbox"/> 50% <input type="checkbox"/> 75% <input type="checkbox"/> 90% <input type="checkbox"/> 100%
2	J06337	<input type="checkbox"/> 0% <input type="checkbox"/> 25% <input type="checkbox"/> 50% <input type="checkbox"/> 75% <input type="checkbox"/> 90% <input type="checkbox"/> 100%
3	4/four months	<input type="checkbox"/> 0% <input type="checkbox"/> 25% <input type="checkbox"/> 50% <input type="checkbox"/> 75% <input type="checkbox"/> 90% <input type="checkbox"/> 100%
4	(Advanced) English (Studies)	<input type="checkbox"/> 0% <input type="checkbox"/> 25% <input type="checkbox"/> 50% <input type="checkbox"/> 75% <input type="checkbox"/> 90% <input type="checkbox"/> 100%
5	(young) children/kids	<input type="checkbox"/> 0% <input type="checkbox"/> 25% <input type="checkbox"/> 50% <input type="checkbox"/> 75% <input type="checkbox"/> 90% <input type="checkbox"/> 100%
6	pets	<input type="checkbox"/> 0% <input type="checkbox"/> 25% <input type="checkbox"/> 50% <input type="checkbox"/> 75% <input type="checkbox"/> 90% <input type="checkbox"/> 100%
7	seafood	<input type="checkbox"/> 0% <input type="checkbox"/> 25% <input type="checkbox"/> 50% <input type="checkbox"/> 75% <input type="checkbox"/> 90% <input type="checkbox"/> 100%
8	tennis	<input type="checkbox"/> 0% <input type="checkbox"/> 25% <input type="checkbox"/> 50% <input type="checkbox"/> 75% <input type="checkbox"/> 90% <input type="checkbox"/> 100%
9	train/(the) train	<input type="checkbox"/> 0% <input type="checkbox"/> 25% <input type="checkbox"/> 50% <input type="checkbox"/> 75% <input type="checkbox"/> 90% <input type="checkbox"/> 100%
10	this/that afternoon	<input type="checkbox"/> 0% <input type="checkbox"/> 25% <input type="checkbox"/> 50% <input type="checkbox"/> 75% <input type="checkbox"/> 90% <input type="checkbox"/> 100%

Answer Sheet: Transfer your **answers** and **confidence (X)**. You may change your confidence at this stage.

11	C	<input type="checkbox"/> 0% <input type="checkbox"/> 25% <input type="checkbox"/> 50% <input type="checkbox"/> 75% <input type="checkbox"/> 90% <input type="checkbox"/> 100%
12	A	<input type="checkbox"/> 0% <input type="checkbox"/> 25% <input type="checkbox"/> 50% <input type="checkbox"/> 75% <input type="checkbox"/> 90% <input type="checkbox"/> 100%
13	C	<input type="checkbox"/> 0% <input type="checkbox"/> 25% <input type="checkbox"/> 50% <input type="checkbox"/> 75% <input type="checkbox"/> 90% <input type="checkbox"/> 100%
14	H	<input type="checkbox"/> 0% <input type="checkbox"/> 25% <input type="checkbox"/> 50% <input type="checkbox"/> 75% <input type="checkbox"/> 90% <input type="checkbox"/> 100%
15	F	<input type="checkbox"/> 0% <input type="checkbox"/> 25% <input type="checkbox"/> 50% <input type="checkbox"/> 75% <input type="checkbox"/> 90% <input type="checkbox"/> 100%
16	B	<input type="checkbox"/> 0% <input type="checkbox"/> 25% <input type="checkbox"/> 50% <input type="checkbox"/> 75% <input type="checkbox"/> 90% <input type="checkbox"/> 100%
17	D	<input type="checkbox"/> 0% <input type="checkbox"/> 25% <input type="checkbox"/> 50% <input type="checkbox"/> 75% <input type="checkbox"/> 90% <input type="checkbox"/> 100%
18	field	<input type="checkbox"/> 0% <input type="checkbox"/> 25% <input type="checkbox"/> 50% <input type="checkbox"/> 75% <input type="checkbox"/> 90% <input type="checkbox"/> 100%
19	footbridge	<input type="checkbox"/> 0% <input type="checkbox"/> 25% <input type="checkbox"/> 50% <input type="checkbox"/> 75% <input type="checkbox"/> 90% <input type="checkbox"/> 100%
20	viewpoint	<input type="checkbox"/> 0% <input type="checkbox"/> 25% <input type="checkbox"/> 50% <input type="checkbox"/> 75% <input type="checkbox"/> 90% <input type="checkbox"/> 100%

Answer Sheet: Transfer your **answers** and **confidence (X)**. You may change your confidence at this stage.

21	fishing industry	<input type="checkbox"/> 0% <input type="checkbox"/> 25% <input type="checkbox"/> 50% <input type="checkbox"/> 75% <input type="checkbox"/> 90% <input type="checkbox"/> 100%
22	statistics	<input type="checkbox"/> 0% <input type="checkbox"/> 25% <input type="checkbox"/> 50% <input type="checkbox"/> 75% <input type="checkbox"/> 90% <input type="checkbox"/> 100%
23	note-taking/ note taking	<input type="checkbox"/> 0% <input type="checkbox"/> 25% <input type="checkbox"/> 50% <input type="checkbox"/> 75% <input type="checkbox"/> 90% <input type="checkbox"/> 100%
24	confidence	<input type="checkbox"/> 0% <input type="checkbox"/> 25% <input type="checkbox"/> 50% <input type="checkbox"/> 75% <input type="checkbox"/> 90% <input type="checkbox"/> 100%
25	ideas	<input type="checkbox"/> 0% <input type="checkbox"/> 25% <input type="checkbox"/> 50% <input type="checkbox"/> 75% <input type="checkbox"/> 90% <input type="checkbox"/> 100%
26	student support	<input type="checkbox"/> 0% <input type="checkbox"/> 25% <input type="checkbox"/> 50% <input type="checkbox"/> 75% <input type="checkbox"/> 90% <input type="checkbox"/> 100%
27	places	<input type="checkbox"/> 0% <input type="checkbox"/> 25% <input type="checkbox"/> 50% <input type="checkbox"/> 75% <input type="checkbox"/> 90% <input type="checkbox"/> 100%
28	general	<input type="checkbox"/> 0% <input type="checkbox"/> 25% <input type="checkbox"/> 50% <input type="checkbox"/> 75% <input type="checkbox"/> 90% <input type="checkbox"/> 100%
29	3/three times	<input type="checkbox"/> 0% <input type="checkbox"/> 25% <input type="checkbox"/> 50% <input type="checkbox"/> 75% <input type="checkbox"/> 90% <input type="checkbox"/> 100%
30	25th/25/twenty five (of)	<input type="checkbox"/> 0% <input type="checkbox"/> 25% <input type="checkbox"/> 50% <input type="checkbox"/> 75% <input type="checkbox"/> 90% <input type="checkbox"/> 100%

Answer Sheet: Transfer your **answers** and **confidence (X)**. You may change your confidence at this stage.

31	C	<input type="checkbox"/> 0% <input type="checkbox"/> 25% <input type="checkbox"/> 50% <input type="checkbox"/> 75% <input type="checkbox"/> 90% <input type="checkbox"/> 100%
32	A	<input type="checkbox"/> 0% <input type="checkbox"/> 25% <input type="checkbox"/> 50% <input type="checkbox"/> 75% <input type="checkbox"/> 90% <input type="checkbox"/> 100%
33	B	<input type="checkbox"/> 0% <input type="checkbox"/> 25% <input type="checkbox"/> 50% <input type="checkbox"/> 75% <input type="checkbox"/> 90% <input type="checkbox"/> 100%
34	B	<input type="checkbox"/> 0% <input type="checkbox"/> 25% <input type="checkbox"/> 50% <input type="checkbox"/> 75% <input type="checkbox"/> 90% <input type="checkbox"/> 100%
35	A	<input type="checkbox"/> 0% <input type="checkbox"/> 25% <input type="checkbox"/> 50% <input type="checkbox"/> 75% <input type="checkbox"/> 90% <input type="checkbox"/> 100%
36	C	<input type="checkbox"/> 0% <input type="checkbox"/> 25% <input type="checkbox"/> 50% <input type="checkbox"/> 75% <input type="checkbox"/> 90% <input type="checkbox"/> 100%
37	frog/frogs	<input type="checkbox"/> 0% <input type="checkbox"/> 25% <input type="checkbox"/> 50% <input type="checkbox"/> 75% <input type="checkbox"/> 90% <input type="checkbox"/> 100%
38	predators	<input type="checkbox"/> 0% <input type="checkbox"/> 25% <input type="checkbox"/> 50% <input type="checkbox"/> 75% <input type="checkbox"/> 90% <input type="checkbox"/> 100%
39	count	<input type="checkbox"/> 0% <input type="checkbox"/> 25% <input type="checkbox"/> 50% <input type="checkbox"/> 75% <input type="checkbox"/> 90% <input type="checkbox"/> 100%
40	seed/seeds	<input type="checkbox"/> 0% <input type="checkbox"/> 25% <input type="checkbox"/> 50% <input type="checkbox"/> 75% <input type="checkbox"/> 90% <input type="checkbox"/> 100%

A1.8 IELTS Listening tapescripts

IELTS listening Transcripts

Section 1	Example
OFFICER: Yes, what can I do for you? STUDENT: My friend is in homestay...and she really enjoys it...so I'd like to join a family as well.	Q1
OFFICER: Okay, so let me get some details. What's your name? STUDENT: My name is Keiko Yuichini.	
OFFICER: Could you spell your family name for me? STUDENT: It's...Yuichini, that's <u>Y-U-I-C-H-I-N-I</u> .	Example
OFFICER: And your first name? STUDENT: It's Keiko. <u>K-E-I-K-O</u> . OFFICER: That's Keiko Yuichini...okay...and you're female. And your nationality?	Q1
STUDENT: I'm Japanese. OFFICER: Right and could I see your passport, please? STUDENT: Here it is...	
OFFICER: Okay...your passport number is <u>JO 6337</u> ...And you're how old?	Q2
STUDENT: I'm twenty-eight years old. OFFICER: Okay...you live at one of the colleges...which one?	
STUDENT: Willow College, umm...Room 21C OFFICER: Right, 21C Willow College, and how long are you planning on staying with homestay?	
STUDENT: <u>About four months</u> ... longer if I like it... OFFICER: And what course are you enrolled in? STUDENT: Well, I've enrolled for twenty weeks in the...um...Advanced English Studies because I need help with my writing...and I'm nearly at the end of my first five week course.	Q3
OFFICER: Okay...Do you have any preference for a family with children or without children?	
STUDENT: I prefer...I mean I like young children, <u>but I'd like to be with older people</u> ...you know...adults...someone around my age.	Q4
OFFICER: Okay, and <u>what about pets</u> ?	Q5
STUDENT: <u>I am a veterinarian so that's fine</u> ...the more the better.	Q6
OFFICER: All right, now what about you? Are you a vegetarian or do you have any special food requirements?	
STUNDET: No, I am not a vegetarian...but I don't eat a lot of meat... <u>I really like seafood</u> .	Q7
OFFICER: And what are your hobbies?	
STUDENT: I like reading and going to the movies.	
OFFICER: Do you play any sports?	
STUDENT: Yes, I joined the handball team, but I didn't like that...so I stopped playing. <u>Now I play tennis</u> on the weekend with my friends...	Q8
OFFICER: All right, let's see, name, age, now the location. Are you familiar with the public transport system?	

<p>STUDENT: No...I'm not really because I have been living on campus...I've been to the city a few times on the bus, but they are always late.</p> <p>OFFICER: What about the trains?</p> <p>STUDENT: <u>I like catching the train...they are much faster...</u></p> <p>OFFICER: Now, let me go check on the computer and see who I've got...Listen, leave it with me...<u>I'll check my records and I'll give you details this afternoon.</u></p> <p>STUDENT: Thank you for helping me...</p> <p>OFFICER: It's a pleasure. Bye.</p> <p>STUDENT: Bye.</p>	<p>Q9</p> <p>Q10</p>
<p>SECTION 2</p> <p>Welcome, everybody, to the lovely house and gardens of Rosewood, once the home of the famous writer, Sebastian George. He bought the house in 1902 although he had first seen it two years earlier. At the time the owners let it out to tenant because <u>George was to slow making up his mind to buy it.</u> When it came back on the market, there was no hesitation and he bought it immediately, for £9,300, even though the house had no bathroom, no running water upstairs, and no electricity.</p> <p>When he came here, he'd been married for ten years. During that time, he'd become one of the most famous writers in the English-speaking world. His professional success was enormous, <u>but his personal life wasn't as successful. He was no longer on speaking terms with his brother and had been devastated by the death at the age of seven of his elder daughter, Josephine.</u></p> <p>Moving to Rosewood allowed the family to start a new life. <u>George regarded Rosewood as a pure example of a traditional country house of this part of England</u> and did some of his most successful writing here. The house and its grounds became the family haven and their escape to privacy and quiet. The walls, and the mullioned windows were built of the local sandstone, the tiles on the roofs and the bricks of the chimney stack were backed from local clay, and the wooden structures inside came from oak trees which grow around here.</p> <p>Now, please look at the map I've given you of the house and gardens. We're here at the Information Centre. Follow the path marked with the arrow and the first area you come to is the orchard on your left.</p> <p>As you go further down the path, there's kitchen garden on the right and <u>as you go round the first sharp corner you will find, to your left, an area where different types of pear trees have been planted as well as some lovely flowers, and this is known as Pear Alley</u> - designed by George himself.</p> <p>Next to this is the greenhouse where some exotic plants and fruits are grown. <u>Follow the path round the second corner and on your right you will see the entrance to the Mulberry Garden</u> with</p>	<p>Q11</p> <p>Q12</p> <p>Q13</p> <p>Q14</p> <p>Q15</p>

<p>its 500-year-old tree. Past the Mulberry Garden, follow the path until you reach the front of the house. I suggest you spend a good hour wandering around this lovely building. A guide takes visitor groups round every two hours.</p> <p>If you would like to purchase any of George's books or other souvenirs, then <u>leave the house by the side entrance, where you will find out shop, which is situated between the house and the garage</u> which contains the magnificent old Rolls-Royce car which used to belong to George. I expect by this time you may also be in need of a rest and some refreshment. Most visitors are, so why don't you visit <u>the tea room on the far side of the garage?</u></p> <p>If you have time, there is a lovely walk down towards the River Dudwell. For me, this is the best part of the estate. This isn't on the map but it is all clearly signposted. You cross the <u>field</u> which spreads along the banks of the river. In spring, this area is well worth a visit. Spend a minute or two watching the water pass by underneath as you cross the <u>footbridge</u>, trees along this path provide the electricity for the house-only about four hours every evening-in George's time. And, finally, for those of you who would like to see stunning views of the surrounding countryside and who are a little bit more energetic, when you return from the mill take the first turning on your left and climb up to the <u>viewpoint</u>. You won't regret it.</p>	<p>Q16</p> <p>Q17</p> <p>Q18</p> <p>Q19</p> <p>Q20</p>
<p>Section 3</p> <p>Tutor: Ah Caroline...come on in. Sit down.</p> <p>CAROLINE: Thanks</p> <p>TUTOR: So how's the dissertation planning going?</p> <p>CAROLINE: Well Dr Schulmann, I'm still having a lot of trouble deciding on the title.</p> <p>TUTOR: Well, that's perfectly normal at this stage. And this is what your tutorials will help you to do.</p> <p>CAROLINE: Right.</p> <p>TUTOR: What we'll do is jot down some points that might help your decision. First of all, you have chosen your general topic area, haven't you?</p> <p>CAROLINE: Yes, it's <u>the fishing industry</u>.</p> <p>TUTOR: Oh yes. That was one of the areas you mentioned. Now, what aspects of the course are you good at?</p> <p>CAROLINE: Well, <u>I think I'm coping well with statistics</u>, and I'm never bored by it.</p> <p>TUTOR: Good. Anything else?</p> <p>CAROLINE: Well, I found computer modelling fascinating - -I have no problem following what's being taught, whereas quite a few of my classmates find it difficult.</p> <p>TUTOR: Well, that's very good. Do you think these might be areas you could bring into your dissertation?</p> <p>CAROLINE: Oh yes, if possible. It's just that I'm having difficulty thinking how I can do that. You see I feel I don't have sufficient background information.</p> <p>TUTOR: I see. Well, do you take notes?</p>	<p>Q21</p> <p>Q22</p>

CAROLINE: <u>I'm very weak at note-taking.</u> My teachers always used to say that.	Q23
TUTOR: Well, I think you really need to work on these weaknesses before you go any further.	
CAROLINE: What do you suggest?	
TUTOR: Well, I can go through the possible strategies with you and let you decide where to go from there.	
CAROLINE: Okay, thanks.	
TUTOR: Well, some people find it helpful to organize peer-group discussions - you know, each week a different topic and shares it with the group.	
CAROLINE: Oh right.	
TUTOR: <u>It really helps build confidence,</u> you know, having to present something to others.	Q24
CAROLINE: I can see that.	
TUTOR: <u>The drawback is that everyone in the group seems to share the same ideas...they keep being repeated in all the dissertations.</u>	Q25
CAROLINE: Okay.	
TUTOR: <u>You could also try a service called 'Student Support'.</u> It's designed to give you a structured programme over a number of weeks to develop your skills.	Q26
CAROLINE: Sounds good.	
TUTOR: Yes, <u>unfortunately there are only a few places.</u> But it's worth looking into.	Q27
CAROLINE: Yes, of course. I know I've got to work on my study skills.	
TUTOR: And then there are several skills books you can consult.	
CAROLINE: Right.	
Tutor: They'll be a good source of reference but <u>the problem is they are sometimes too general.</u>	Q28
CAROLINE: Yes, that's what I've found.	
TUTOR: Other than that I would strongly advise quite simple ideas like using a card index.	
CAROLINE: Well, yes. I've never done that before.	
TUTOR: It's simple, but it really works because you have to get points down in a small space. Another thing I always advise is don't just take your notes and forget about them. <u>Read everything three times</u> - that'll really fix them in your mind.	Q29
CAROLINE: Yes, I can see it'd take discipline but...	
TUTOR: Well, if you establish good study skills at this stage they'll be with you all your life.	
CAROLINE: Oh yes, I completely agree. It's just that I don't seem to be able to discipline myself. I need to talk things over.	
TUTOR: Well, we'll be continuing these tutorials of course. Let's arrange next month's now. Let's see, I can see you virtually any time during the week starting 22 nd January.	
CAROLINE: What about the 24 th ? I'm free in the afternoon.	
TUTOR: Sorry, I'm booked then. What about the following day?	

<p>CAROLINE: Thursday? I can make the morning. TUTOR: Fine, <u>we'll go for the 25th then.</u> CAROLINE: That's great, thanks.</p>	<p>Q30</p>
<p>Section 4 Good morning. Today I'd like to present the findings of our year 2 project on wildlife found in gardens throughout our city. I'll start by saying something about the background to the project, then talk a little bit about our research techniques, and then indicate some of our interim findings.</p> <p>First of all, how did we choose our topic? Well, there are four of us in the group and one day while we were discussing a possible focus, <u>two of the group mentioned that they had seen yet more sparrow-hawks - one of Britain's most interesting birds of prey - in their own city centre gardens and wondered why they were turning up in these gardens in great numbers.</u> We were all very engaged by the idea of why wild animals would choose to inhabit a city garden. Why is it so popular with wildlife when the countryside itself is becoming less so?</p> <p>The first thing we did was to establish what population of the urban land is taken up by private gardens. We estimated that it was about one fifth, and <u>this was endorsed by looking at large-scale usage maps in the town land survey office - 24% to be precise.</u> Our own informal discussions with neighbours and friends led us to believe that many garden owners had interesting experiences to relate regarding wild animal sightings so we decided to <u>survey garden owners from different areas of the city. Just over 100 of them completed a survey once every two weeks for twelve months - ticking off species they had seen from a pro forma list-and adding the names of any rarer ones.</u> Meanwhile, we were doing our own observations in selected gardens throughout the city. <u>We deliberately chose smaller ones because they were by far the most typical in the city. The whole point of the project was to look at the norm not the exception.</u> Alongside this primary research on urban gardens, <u>we were studying a lot of books about the decline of wild animals in the countryside and thinking of possible causes for this.</u></p> <p>So what did we find? Well, so much that I just won't have time to tell you about here. If you're interested in reading our more comprehensive findings, we've produced detailed graphic representations on the college web-site and of course any of the group would be happy to talk to you about them. Just email us.</p> <p><u>What we've decided to present today is information about just three species - because we felt these gave a good indication of the processes at work in rural and urban settings as a whole.</u></p> <p><u>The first species to generate a lot of interesting information was frogs.</u> And there was a clear pattern here-they proliferate where there is suitable water. Garden ponds are on the increase, rural ponds are disappearing, leading to massive migration to the towns.</p>	<p>Q31</p> <p>Q32</p> <p>Q33</p> <p>Q34</p> <p>Q35</p> <p>Q36</p> <p>Q37</p>

<p>Hedgehogs are also finding it easier to live in urban areas - this time because <u>their predators are not finding it quite so attractive to leave their rural environment, so hedgehogs have a better survival rate in cities</u>. We had lots of sightings, so all in all <u>we had no difficulties with our efforts to court their numbers precisely</u>.</p>	Q38
<p>Our final species is the finest of bird singers, the song thrush. On the decline in the countryside, they are experiencing a resurgence in urban gardens because these days gardeners are buying lots of different plants which means <u>there's an extensive range of seeds around, which is what they feed on</u>. Another factor is the provision of nesting places - which is actually better in gardens than the countryside. Hard to believe it, but it's true. Incidentally, we discovered that a massive new survey on song thrushes is about to be launched, so you should keep an eye open for that.</p>	Q39
<p>Now, I'd be happy to answer any questions you may have...</p>	Q40

A1.9 Example of feedback to students

IELTS and Realism Feedback

ID: XXX

Name:XXX

Email: XXX

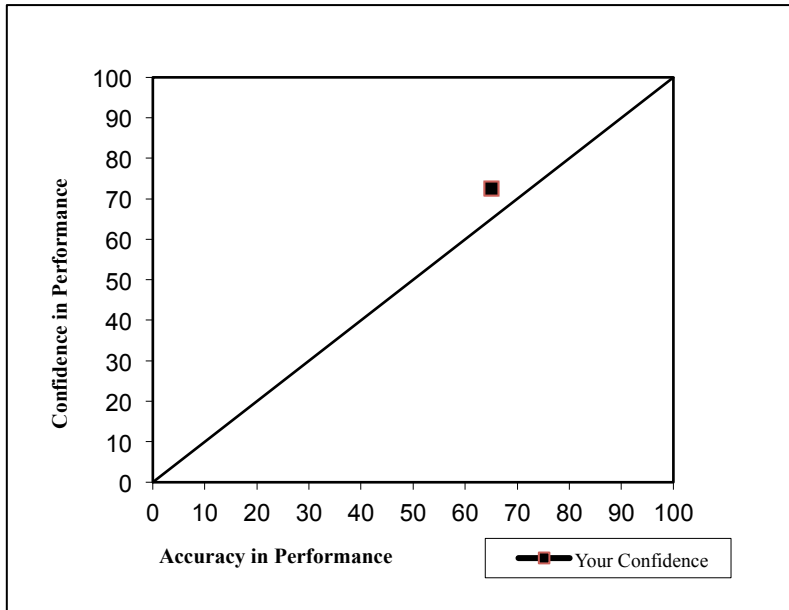
Raw Test Score: 26

Test Score in Percentage: 65%

Confidence Score: 73%

Calibration Score: + 8%

Your Calibration Diagram:



Your current appraisal calibration:

- Realistic (within $\pm 5\%$) Just overconfident ($6 < 10\%$) Generally overconfident (11-24%)
 Extremely overconfident ($> 25\%$) Just underconfident ($-6 < -10\%$) Generally underconfident (-11- -24%)
 Extremely underconfident ($> -25\%$)

You can generally approximate your current performance success. However, you have a tendency to be overconfident in your performance, especially when you are faced with difficult questions or tasks. Your calibration score is within the top 10 percent. You should engage more practice in estimating your listening performance success and observe the match between your confidence in your performance and the actual performance. This practice will help you become more realistic about your listening performance.

APPENDIX 2: IRT ANALYSIS

A2.1 Calculating fit statistics

```
>=====<
Standardized Residuals N(0,1) Mean: .00 S.D.: .99
Time for estimation: 0:0:0.281
Processing Table 0
Listening Test File for Rasch.sav
```

PERSON	388 INPUT	388 MEASURED	INFIT		OUTFIT			
	TOTAL	COUNT	MEASURE	REALSE	IMNSQ	ZSTD	OMNSQ	ZSTD
MEAN	20.1	40.0	-.01	.41	1.01	.0	.99	.0
S.D.	7.3	.0	1.11	.06	.18	1.0	.34	.8
REAL RMSE	.41	TRUE SD	1.03	SEPARATION	2.49	PERSON	RELIABILITY	.87

ITEM	40 INPUT	40 MEASURED	INFIT		OUTFIT			
	TOTAL	COUNT	MEASURE	REALSE	IMNSQ	ZSTD	OMNSQ	ZSTD
MEAN	194.7	388.0	.00	.13	.99	-.1	.99	.0
S.D.	85.2	.0	1.34	.03	.14	2.5	.30	2.6
REAL RMSE	.14	TRUE SD	1.33	SEPARATION	9.85	ITEM	RELIABILITY	.99

A2.2 Item fit graph: Misfit order

ENTRY	MEASURE	INFIT	MEAN-SQUARE	OUTFIT	MEAN-SQUARE	ITEM			
NUMBER	-	+	0.0	1	2	0.0	1	2	ITEM
9	*	:	.	*	:	.	.	*	Q9
32	*	:	.	*	:	.	.	*	Q32
13	*	:	.	*	:	.	*	.	Q13
33	*	:	.	*	:	.	*	.	Q33
35	*	:	.	*	:	.	*	.	Q35
34	*	:	.	*	:	.	*	.	Q34
31	*	:	.	*	:	.	*	.	Q31
7	*	:	.	*	:	.	*	.	Q7
4	*	:	.	*	:	.	*	.	Q4
3	*	:	.	*	:	.	*	.	Q3
25	*	:	.	*	:	.	*	.	Q25
2	*	:	.	*	:	.	*	.	Q2
6	*	:	.	*	:	.	*	.	Q6
16	*	:	.	*	:	.	*	.	Q16
1	*	:	.	*	:	.	*	.	Q1
8	*	:	.	*	:	.	*	.	Q8
36	*	:	.	*	:	.	*	.	Q36
15	*	:	.	*	:	.	*	.	Q15
10	*	:	.	*	:	.	*	.	Q10
17	*	:	.	*	:	.	*	.	Q17
5	*	:	.	*	:	.	*	.	Q5
19	*	:	.	*	:	.	*	.	Q19
37	*	:	.	*	:	.	*	.	Q37
30	*	:	.	*	:	.	*	.	Q30
11	*	:	.	*	:	.	*	.	Q11
12	*	:	.	*	:	.	*	.	Q12
14	*	:	.	*	:	.	*	.	Q14
18	*	:	.	*	:	.	*	.	Q18
20	*	:	.	*	:	.	*	.	Q20
39	*	:	.	*	:	.	*	.	Q39
26	*	:	.	*	:	.	*	.	Q26
29	*	:	.	*	:	.	*	.	Q29
38	*	:	.	*	:	.	*	.	Q38
22	*	:	.	*	:	.	*	.	Q22
21	*	:	.	*	:	.	*	.	Q21
40	*	:	.	*	:	.	*	.	Q40
23	*	:	.	*	:	.	*	.	Q23
24	*	:	.	*	:	.	*	.	Q24
27	*	:	.	*	:	.	*	.	Q27
28	*	:	.	*	:	.	*	.	Q28

A2.3 Item statistics: Measure order

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL		INFIT		OUTFIT		PTMEASURE-A		EXACT MATCH		ITEM
				S.E.	MNSQ	ZSTD	MNSQ	ZSTD	CORR.	EXP.	OBS%	EXP%		
19	17	388	3.64	.26	.96	-.1	.49	-1.5	.31	.24	95.9	95.6	Q19	
38	36	388	2.74	.19	.88	-.9	.55	-2.0	.44	.31	91.0	91.0	Q38	
20	63	388	2.00	.15	.93	-.7	.75	-1.5	.45	.37	84.3	84.8	Q20	
39	70	388	1.85	.14	.90	-1.2	.75	-1.7	.48	.38	85.1	83.3	Q39	
40	76	388	1.72	.14	.85	-2.0	.71	-2.2	.52	.39	84.5	82.1	Q40	
22	97	388	1.34	.13	.86	-2.2	.75	-2.3	.54	.42	80.2	78.2	Q22	
32	97	388	1.34	.13	1.29	4.0	1.86	6.0	.13	.42	75.0	78.2	Q32	
4	99	388	1.31	.13	1.04	.6	1.15	1.3	.37	.42	79.1	77.8	Q4	
18	101	388	1.28	.13	.93	-1.1	.79	-2.0	.50	.42	77.1	77.5	Q18	
31	138	388	.72	.12	1.03	.6	1.25	2.9	.39	.44	72.9	72.6	Q31	
26	157	388	.46	.12	.88	-2.6	.83	-2.5	.55	.45	75.8	70.7	Q26	
23	158	388	.44	.12	.84	-3.6	.82	-2.7	.57	.45	78.1	70.7	Q23	
33	159	388	.43	.12	1.26	5.1	1.33	4.2	.23	.45	61.3	70.6	Q33	
37	159	388	.43	.12	.95	-1.1	.96	-.5	.48	.45	73.2	70.6	Q37	
13	161	388	.40	.12	1.28	5.5	1.47	5.9	.20	.45	63.9	70.4	Q13	
10	169	388	.30	.11	.97	-.7	.90	-1.4	.48	.45	67.8	69.9	Q10	
34	174	388	.23	.11	1.18	3.7	1.29	3.9	.29	.45	63.9	69.6	Q34	
21	175	388	.22	.11	.85	-3.5	.81	-3.0	.57	.45	74.5	69.5	Q21	
35	175	388	.22	.11	1.24	4.9	1.32	4.3	.25	.45	58.5	69.5	Q35	
16	193	388	-.01	.11	1.03	.7	1.01	.2	.42	.44	68.8	69.1	Q16	
36	193	388	-.01	.11	.99	-.1	.97	-.4	.45	.44	68.3	69.1	Q36	
27	216	388	-.31	.11	.83	-4.0	.75	-3.8	.58	.44	76.5	69.2	Q27	
17	220	388	-.36	.11	.97	-.6	.93	-.9	.46	.44	67.8	69.3	Q17	
14	225	388	-.43	.11	.94	-1.4	.89	-1.5	.49	.43	71.6	69.5	Q14	
2	229	388	-.48	.11	1.04	1.0	1.07	.9	.39	.43	68.0	69.7	Q2	
12	231	388	-.51	.12	.94	-1.4	.89	-1.4	.49	.43	71.1	69.8	Q12	
5	233	388	-.53	.12	.96	-.8	.92	-1.0	.46	.43	73.7	70.0	Q5	
6	238	388	-.60	.12	1.04	.8	1.03	.4	.40	.43	71.1	70.3	Q6	
7	250	388	-.76	.12	1.13	2.6	1.19	2.1	.30	.42	67.8	71.3	Q7	
25	252	388	-.79	.12	.98	-.4	1.09	1.0	.42	.42	75.5	71.5	Q25	
11	259	388	-.89	.12	.94	-1.2	.89	-1.2	.46	.41	72.7	72.3	Q11	
24	268	388	-1.02	.12	.84	-3.2	.74	-2.8	.54	.40	78.4	73.5	Q24	
9	283	388	-1.24	.13	1.35	5.4	2.01	6.7	.03	.39	69.3	75.8	Q9	
15	290	388	-1.36	.13	.98	-.4	.94	-.5	.41	.38	76.3	76.9	Q15	
28	294	388	-1.42	.13	.81	-3.1	.65	-3.0	.54	.37	81.4	77.6	Q28	
3	305	388	-1.61	.13	1.03	.4	1.10	.7	.32	.36	80.7	79.7	Q3	
29	319	388	-1.88	.14	.88	-1.5	.74	-1.6	.44	.34	85.1	82.8	Q29	
8	324	388	-1.98	.15	1.01	.1	1.00	.0	.32	.33	84.3	83.9	Q8	
30	333	388	-2.19	.15	.95	-.5	.89	-.5	.35	.31	86.6	86.0	Q30	
1	351	388	-2.69	.18	.99	.0	1.01	.1	.27	.26	90.2	90.5	Q1	
MEAN	194.7	388.0	.00	.13	.99	-.1	.99	.0			75.7	75.5		
S.D.	85.2	.0	1.34	.03	.14	2.5	.30	2.6			8.3	7.0		

A2.4 Person statistics: Measure order

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	ZSTD	OUTFIT MNSQ	ZSTD	PTMEASURE-A CORR.	EXP.	OBS%	EXP%	EXACT MATCH	PERSON
351	38	40	3.76	.79	1.23	.6	.54	.1	.26	.31	95.0	95.0	USFP111	USFP111
45	37	40	3.24	.66	.80	-.3	.34	-.4	.49	.35	95.0	93.0	CET45	CET45
357	37	40	3.24	.66	.66	-.6	.28	-.6	.55	.35	95.0	93.0	USFP117	USFP117
193	36	40	2.85	.59	.85	-.2	.68	-.1	.46	.39	92.5	90.9	TELP47	TELP47
350	36	40	2.85	.59	.94	.0	.63	-.1	.44	.39	92.5	90.9	USFP110	USFP110
26	35	40	2.53	.54	1.13	.5	.73	-.1	.38	.41	87.5	89.0	CET26	CET26
347	35	40	2.53	.54	1.06	.3	1.60	.9	.33	.41	87.5	89.0	USFP107	USFP107
376	35	40	2.53	.54	1.21	.7	1.78	1.1	.25	.41	87.5	89.0	USFP136	USFP136
336	34	40	2.26	.50	1.20	.7	1.16	.5	.32	.43	85.0	87.1	USFP96	USFP96
338	34	40	2.26	.50	1.55	1.6	2.79	2.0	.02	.43	80.0	87.1	USFP98	USFP98
361	34	40	2.26	.50	1.00	.1	.67	-.3	.47	.43	85.0	87.1	USFP121	USFP121
371	34	40	2.26	.50	1.28	.9	1.11	.4	.29	.43	85.0	87.1	USFP131	USFP131
380	34	40	2.26	.50	1.35	1.1	1.18	.5	.26	.43	80.0	87.1	USFP140	USFP140
250	33	40	2.02	.47	1.37	1.3	2.26	1.8	.14	.45	82.5	85.0	USFP10	USFP10
317	33	40	2.02	.47	1.06	.3	.84	-.1	.44	.45	82.5	85.0	USFP77	USFP77
328	33	40	2.02	.47	.96	.0	.67	-.4	.50	.45	87.5	85.0	USFP88	USFP88
341	33	40	2.02	.47	1.26	.9	.90	.0	.34	.45	82.5	85.0	USFP101	USFP101
369	33	40	2.02	.47	1.17	.7	1.09	.4	.35	.45	82.5	85.0	USFP129	USFP129
70	32	40	1.81	.45	.97	.0	1.23	.6	.44	.46	80.0	83.2	CET70	CET70
180	32	40	1.81	.45	1.08	.4	1.15	.5	.41	.46	80.0	83.2	TELP34	TELP34
248	32	40	1.81	.45	1.07	.4	.93	.0	.43	.46	85.0	83.2	USFP08	USFP08
325	32	40	1.81	.45	1.24	.9	1.09	.3	.34	.46	80.0	83.2	USFP85	USFP85
334	32	40	1.81	.45	.82	-.6	.79	-.2	.56	.46	85.0	83.2	USFP94	USFP94
345	32	40	1.81	.45	.90	-.3	.75	-.3	.52	.46	90.0	83.2	USFP105	USFP105
28	31	40	1.61	.44	1.03	.2	.84	-.2	.47	.47	80.0	81.6	CET28	CET28
175	31	40	1.61	.44	.70	-1.3	.48	-1.2	.66	.47	90.0	81.6	TELP29	TELP29
245	31	40	1.61	.44	1.10	.5	.78	-.3	.45	.47	75.0	81.6	USFP05	USFP05
252	31	40	1.61	.44	.84	-.6	.93	.0	.55	.47	90.0	81.6	USFP12	USFP12
292	31	40	1.61	.44	.84	-.6	.58	-.9	.59	.47	80.0	81.6	USFP52	USFP52
314	31	40	1.61	.44	.97	.0	.69	-.6	.52	.47	80.0	81.6	USFP74	USFP74
319	31	40	1.61	.44	1.13	.6	1.41	.9	.35	.47	80.0	81.6	USFP79	USFP79
329	31	40	1.61	.44	.79	-.9	.57	-.9	.61	.47	85.0	81.6	USFP89	USFP89
344	31	40	1.61	.44	1.10	.5	1.10	.4	.41	.47	80.0	81.6	USFP104	USFP104
348	31	40	1.61	.44	.79	-.8	.55	-1.0	.62	.47	80.0	81.6	USFP108	USFP108
356	31	40	1.61	.44	.67	-1.5	.47	-1.2	.68	.47	90.0	81.6	USFP116	USFP116
365	31	40	1.61	.44	1.02	.2	1.04	.3	.45	.47	85.0	81.6	USFP125	USFP125
381	31	40	1.61	.44	1.50	1.9	3.35	3.3	.06	.47	70.0	81.6	USFP141	USFP141
190	30	40	1.43	.42	.94	-.2	.91	-.1	.51	.48	77.5	80.0	TELP44	TELP44
240	30	40	1.43	.42	1.16	.8	1.78	1.6	.34	.48	77.5	80.0	TELP94	TELP94
251	30	40	1.43	.42	1.01	.1	1.06	.3	.46	.48	82.5	80.0	USFP11	USFP11
335	30	40	1.43	.42	1.41	1.7	1.31	.8	.25	.48	67.5	80.0	USFP95	USFP95
346	30	40	1.43	.42	1.12	.6	1.18	.5	.40	.48	77.5	80.0	USFP106	USFP106
358	30	40	1.43	.42	1.22	1.0	1.03	.2	.37	.48	72.5	80.0	USFP118	USFP118
379	30	40	1.43	.42	1.29	1.3	1.24	.7	.31	.48	72.5	80.0	USFP139	USFP139
38	29	40	1.26	.41	1.32	1.5	2.02	2.1	.23	.48	70.0	78.6	CET38	CET38
44	29	40	1.26	.41	1.05	.3	.97	.1	.46	.48	75.0	78.6	CET44	CET44
46	29	40	1.26	.41	.82	-.8	.81	-.4	.59	.48	85.0	78.6	CET46	CET46
60	29	40	1.26	.41	1.15	.8	1.46	1.2	.36	.48	75.0	78.6	CET60	CET60
129	29	40	1.26	.41	1.01	.1	.87	-.2	.50	.48	80.0	78.6	CET129	CET129
201	29	40	1.26	.41	.82	-.8	.80	-.4	.59	.48	85.0	78.6	TELP55	TELP55
214	29	40	1.26	.41	.98	.0	.73	-.6	.53	.48	75.0	78.6	TELP68	TELP68
264	29	40	1.26	.41	1.03	.2	.99	.1	.47	.48	80.0	78.6	USFP24	USFP24
297	29	40	1.26	.41	1.03	.2	1.34	.9	.44	.48	80.0	78.6	USFP57	USFP57
320	29	40	1.26	.41	1.12	.6	.96	.0	.43	.48	70.0	78.6	USFP80	USFP80
327	29	40	1.26	.41	1.06	.4	1.09	.4	.44	.48	80.0	78.6	USFP87	USFP87
333	29	40	1.26	.41	1.30	1.4	1.46	1.2	.28	.48	75.0	78.6	USFP93	USFP93
372	29	40	1.26	.41	.88	-.5	.69	-.7	.58	.48	80.0	78.6	USFP132	USFP132
385	29	40	1.26	.41	1.26	1.2	1.23	.7	.34	.48	70.0	78.6	USFP145	USFP145
34	28	40	1.10	.40	1.12	.7	1.11	.4	.41	.49	72.5	77.6	CET34	CET34
49	28	40	1.10	.40	1.14	.8	1.02	.2	.42	.49	72.5	77.6	CET49	CET49
92	28	40	1.10	.40	.80	-1.0	.65	-1.0	.62	.49	82.5	77.6	CET92	CET92
99	28	40	1.10	.40	1.30	1.5	1.49	1.3	.28	.49	67.5	77.6	CET99	CET99
116	28	40	1.10	.40	1.53	2.4	1.80	1.9	.13	.49	62.5	77.6	CET116	CET116
149	28	40	1.10	.40	.81	-1.0	.76	-.6	.61	.49	82.5	77.6	TELP03	TELP03
225	28	40	1.10	.40	1.18	.9	1.41	1.1	.35	.49	72.5	77.6	TELP79	TELP79
242	28	40	1.10	.40	1.00	.1	.95	.0	.49	.49	77.5	77.6	USFP02	USFP02
253	28	40	1.10	.40	1.08	.4	1.32	.9	.42	.49	77.5	77.6	USFP13	USFP13
254	28	40	1.10	.40	1.11	.6	1.57	1.5	.37	.49	77.5	77.6	USFP14	USFP14
331	28	40	1.10	.40	.95	-.2	.78	-.5	.54	.49	77.5	77.6	USFP91	USFP91
343	28	40	1.10	.40	1.30	1.5	1.31	.9	.30	.49	67.5	77.6	USFP103	USFP103
349	28	40	1.10	.40	1.15	.8	1.34	1.0	.37	.49	77.5	77.6	USFP109	USFP109
378	28	40	1.10	.40	.87	-.6	.73	-.7	.58	.49	82.5	77.6	USFP138	USFP138
382	28	40	1.10	.40	1.34	1.6	1.16	.5	.31	.49	62.5	77.6	USFP142	USFP142
30	27	40	.94	.39	.82	-1.0	.66	-1.0	.62	.49	80.0	76.5	CET30	CET30
108	27	40	.94	.39	1.26	1.3	1.58	1.6	.30	.49	75.0	76.5	CET108	CET108
109	27	40	.94	.39	.95	-.2	.81	-.5	.54	.49	80.0	76.5	CET109	CET109
138	27	40	.94	.39	.96	-.1	.85	-.3	.52	.49	80.0	76.5	CET138	CET138
261	27	40	.94	.39	.93	-.3	1.66	1.8	.48	.49	80.0	76.5	USFP21	USFP21

PHAKITI: TEST-TAKERS' PERFORMANCE APPRAISALS, APPRAISAL CALIBRATION, STATE-TRAIT STRATEGY USE,
AND STATE-TRAIT IELTS LISTENING DIFFICULTY IN A SIMULATED IELTS LISTENING TEST

277	27	40	.94	.39	1.00	.1	1.18	.6	.47	.49	80.0	76.5	USFP37
280	27	40	.94	.39	1.12	.7	1.11	.4	.42	.49	75.0	76.5	USFP40
296	27	40	.94	.39	1.03	.2	.99	.1	.48	.49	75.0	76.5	USFP56
308	27	40	.94	.39	.72	-1.6	.64	-1.1	.67	.49	85.0	76.5	USFP68
332	27	40	.94	.39	1.47	2.3	1.38	1.1	.22	.49	60.0	76.5	USFP92
366	27	40	.94	.39	1.21	1.1	1.17	.6	.37	.49	70.0	76.5	USFP126
377	27	40	.94	.39	1.07	.4	1.04	.2	.45	.49	75.0	76.5	USFP137
1	26	40	.79	.38	.78	-1.3	.66	-1.1	.64	.49	82.5	75.3	CET01
47	26	40	.79	.38	.93	-.3	.99	.1	.52	.49	82.5	75.3	CET47
139	26	40	.79	.38	1.04	.3	1.19	.7	.46	.49	72.5	75.3	CET139
142	26	40	.79	.38	1.20	1.1	1.40	1.3	.35	.49	72.5	75.3	CET142
204	26	40	.79	.38	.97	-.1	.89	-.2	.52	.49	77.5	75.3	TELP58
209	26	40	.79	.38	.96	-.2	.89	-.3	.52	.49	77.5	75.3	TELP63
230	26	40	.79	.38	.89	-.6	.76	-.7	.58	.49	77.5	75.3	TELP84
235	26	40	.79	.38	1.43	2.2	1.34	1.1	.25	.49	62.5	75.3	TELP89
257	26	40	.79	.38	1.01	.1	.96	.0	.49	.49	72.5	75.3	USFP17
281	26	40	.79	.38	1.24	1.3	1.18	.6	.35	.49	67.5	75.3	USFP41
322	26	40	.79	.38	.98	-.1	.96	.0	.51	.49	72.5	75.3	USFP82
339	26	40	.79	.38	1.09	.5	1.01	.1	.45	.49	72.5	75.3	USFP99
352	26	40	.79	.38	.94	-.3	.76	-.7	.56	.49	72.5	75.3	USFP112
359	26	40	.79	.38	1.23	1.3	1.30	1.0	.35	.49	67.5	75.3	USFP119
130	25	40	.64	.38	.98	-.1	1.28	1.0	.48	.50	72.5	74.1	CET130
141	25	40	.64	.38	.77	-1.4	.69	-1.1	.64	.50	77.5	74.1	CET141
154	25	40	.64	.38	.90	-.6	1.00	.1	.55	.50	82.5	74.1	TELP08
159	25	40	.64	.38	.90	-.5	.74	-.9	.58	.50	72.5	74.1	TELP13
227	25	40	.64	.38	1.09	.6	.96	.0	.46	.50	72.5	74.1	TELP81
266	25	40	.64	.38	1.15	.9	1.04	.2	.42	.50	67.5	74.1	USFP26
269	25	40	.64	.38	.98	-.1	.85	-.4	.53	.50	72.5	74.1	USFP29
276	25	40	.64	.38	.83	-1.0	.70	-1.0	.62	.50	77.5	74.1	USFP36
290	25	40	.64	.38	1.14	.8	1.22	.8	.40	.50	72.5	74.1	USFP50
337	25	40	.64	.38	1.06	.4	.97	.0	.47	.50	72.5	74.1	USFP97
354	25	40	.64	.38	.97	-.1	.83	-.5	.53	.50	72.5	74.1	USFP114
25	24	40	.50	.37	.81	-1.2	.72	-1.0	.62	.50	80.0	72.8	CET25
127	24	40	.50	.37	1.00	.1	1.10	.4	.49	.50	70.0	72.8	CET127
178	24	40	.50	.37	1.05	.3	1.09	.4	.46	.50	70.0	72.8	TELP32
189	24	40	.50	.37	1.11	.7	1.19	.7	.42	.50	70.0	72.8	TELP43
199	24	40	.50	.37	.87	-.8	.93	-.2	.57	.50	80.0	72.8	TELP53
215	24	40	.50	.37	1.02	.2	.98	.0	.49	.50	70.0	72.8	TELP69
229	24	40	.50	.37	1.15	.9	1.52	1.7	.38	.50	70.0	72.8	TELP83
287	24	40	.50	.37	.91	-.5	.81	-.6	.56	.50	75.0	72.8	USFP47
315	24	40	.50	.37	1.05	.4	1.05	.3	.46	.50	70.0	72.8	USFP75
326	24	40	.50	.37	1.23	1.4	1.13	.5	.37	.50	65.0	72.8	USFP86
330	24	40	.50	.37	.83	-1.0	.69	-1.1	.62	.50	80.0	72.8	USFP90
355	24	40	.50	.37	1.01	.1	.98	.0	.49	.50	75.0	72.8	USFP115
368	24	40	.50	.37	.95	-.2	.87	-.4	.53	.50	75.0	72.8	USFP128
374	24	40	.50	.37	.94	-.3	.82	-.6	.55	.50	75.0	72.8	USFP134
375	24	40	.50	.37	1.07	.5	1.05	.3	.46	.50	75.0	72.8	USFP135
52	23	40	.37	.37	.98	-.1	1.07	.3	.50	.50	70.0	71.9	CET52
96	23	40	.37	.37	1.27	1.6	1.38	1.3	.32	.50	60.0	71.9	CET96
126	23	40	.37	.37	1.06	.4	1.19	.7	.44	.50	70.0	71.9	CET126
155	23	40	.37	.37	.89	-.7	.86	-.4	.56	.50	80.0	71.9	TELP09
179	23	40	.37	.37	.72	-1.9	.60	-1.6	.68	.50	85.0	71.9	TELP33
210	23	40	.37	.37	.99	.0	.93	-.2	.51	.50	75.0	71.9	TELP64
211	23	40	.37	.37	1.03	.3	.96	.0	.48	.50	75.0	71.9	TELP65
237	23	40	.37	.37	1.02	.2	1.00	.1	.49	.50	70.0	71.9	TELP91
243	23	40	.37	.37	.73	-1.8	.62	-1.5	.67	.50	80.0	71.9	USFP03
249	23	40	.37	.37	.95	-.3	.92	-.2	.53	.50	75.0	71.9	USFP09
272	23	40	.37	.37	.75	-1.7	.62	-1.5	.67	.50	75.0	71.9	USFP32
283	23	40	.37	.37	.82	-1.1	.72	-1.0	.62	.50	80.0	71.9	USFP43
289	23	40	.37	.37	.84	-1.0	.72	-1.1	.61	.50	80.0	71.9	USFP49
299	23	40	.37	.37	.84	-1.0	.90	-.3	.58	.50	85.0	71.9	USFP59
318	23	40	.37	.37	.88	-.7	.98	.0	.56	.50	70.0	71.9	USFP78
323	23	40	.37	.37	.75	-1.7	.64	-1.4	.66	.50	85.0	71.9	USFP83
324	23	40	.37	.37	1.00	.0	.91	-.2	.51	.50	70.0	71.9	USFP84
342	23	40	.37	.37	1.35	2.1	1.98	2.9	.22	.50	65.0	71.9	USFP102
362	23	40	.37	.37	.83	-1.1	.71	-1.1	.62	.50	70.0	71.9	USFP122
370	23	40	.37	.37	.99	.0	.87	-.4	.52	.50	70.0	71.9	USFP130
373	23	40	.37	.37	1.03	.3	.98	.0	.48	.50	75.0	71.9	USFP133
2	22	40	.23	.37	1.60	3.4	2.00	3.0	.08	.50	52.5	71.4	CET02
3	22	40	.23	.37	.98	-.1	1.19	.8	.49	.50	72.5	71.4	CET03
23	22	40	.23	.37	1.03	.3	.99	.1	.48	.50	72.5	71.4	CET23
37	22	40	.23	.37	1.39	2.3	1.46	1.6	.25	.50	52.5	71.4	CET37
105	22	40	.23	.37	1.06	.4	.98	.0	.47	.50	67.5	71.4	CET105
131	22	40	.23	.37	1.18	1.2	1.14	.6	.39	.50	67.5	71.4	CET131
136	22	40	.23	.37	.90	-.6	.78	-.8	.57	.50	67.5	71.4	CET136
146	22	40	.23	.37	.72	-1.9	.60	-1.6	.68	.50	77.5	71.4	CET146
160	22	40	.23	.37	.97	-.1	.89	-.3	.52	.50	67.5	71.4	TELP14
173	22	40	.23	.37	.90	-.6	.83	-.6	.56	.50	77.5	71.4	TELP27
183	22	40	.23	.37	1.01	.1	.92	-.2	.50	.50	67.5	71.4	TELP37
207	22	40	.23	.37	.87	-.8	.84	-.5	.57	.50	82.5	71.4	TELP61
234	22	40	.23	.37	.79	-1.4	.84	-.5	.62	.50	77.5	71.4	TELP88
244	22	40	.23	.37	1.05	.4	1.09	.4	.46	.50	72.5	71.4	USFP04
307	22	40	.23	.37	.95	-.3	.88	-.4	.53	.50	72.5	71.4	USFP67

PHAKITI: TEST-TAKERS' PERFORMANCE APPRAISALS, APPRAISAL CALIBRATION, STATE-TRAIT STRATEGY USE,
AND STATE-TRAIT IELTS LISTENING DIFFICULTY IN A SIMULATED IELTS LISTENING TEST

316	22	40	.23	.37	.86	-.9	.82	-.6	.58	.50	82.5	71.4	USFP76
321	22	40	.23	.37	.86	-.9	.77	-.8	.59	.50	77.5	71.4	USFP81
363	22	40	.23	.37	.88	-.8	.84	-.5	.57	.50	72.5	71.4	USFP123
367	22	40	.23	.37	1.26	1.6	1.23	.9	.34	.50	57.5	71.4	USFP127
384	22	40	.23	.37	.84	-1.0	.74	-1.0	.60	.50	77.5	71.4	USFP144
9	21	40	.09	.37	.81	-1.3	.85	-.5	.60	.49	80.0	71.4	CET09
12	21	40	.09	.37	1.04	.3	1.00	.1	.47	.49	75.0	71.4	CET12
117	21	40	.09	.37	.88	-.8	.83	-.6	.57	.49	75.0	71.4	CET117
150	21	40	.09	.37	.81	-1.3	.69	-1.2	.63	.49	75.0	71.4	TELP04
169	21	40	.09	.37	.95	-.3	.92	-.2	.53	.49	75.0	71.4	TELP23
184	21	40	.09	.37	1.08	.6	1.04	.2	.45	.49	70.0	71.4	TELP38
202	21	40	.09	.37	.96	-.2	.86	-.5	.53	.49	70.0	71.4	TELP56
213	21	40	.09	.37	.94	-.3	.85	-.5	.54	.49	75.0	71.4	TELP67
265	21	40	.09	.37	.70	-2.2	.60	-1.7	.69	.49	85.0	71.4	USFP25
268	21	40	.09	.37	1.00	.0	.99	.1	.49	.49	75.0	71.4	USFP28
273	21	40	.09	.37	.81	-1.3	.71	-1.1	.62	.49	75.0	71.4	USFP33
274	21	40	.09	.37	1.12	.8	1.07	.4	.43	.49	60.0	71.4	USFP34
278	21	40	.09	.37	.76	-1.7	.63	-1.5	.66	.49	75.0	71.4	USFP38
291	21	40	.09	.37	.74	-1.8	.62	-1.5	.67	.49	80.0	71.4	USFP51
383	21	40	.09	.37	.77	-1.6	.95	-.1	.62	.49	80.0	71.4	USFP143
387	21	40	.09	.37	1.57	3.3	1.54	1.8	.14	.49	45.0	71.4	USFP147
43	20	40	-.04	.37	.96	-.2	1.07	.4	.51	.49	72.5	71.3	CET43
58	20	40	-.04	.37	.92	-.5	.84	-.5	.55	.49	67.5	71.3	CET58
81	20	40	-.04	.37	1.03	.2	1.00	.1	.48	.49	62.5	71.3	CET81
104	20	40	-.04	.37	1.07	.5	1.02	.2	.45	.49	67.5	71.3	CET104
107	20	40	-.04	.37	.95	-.3	.86	-.4	.53	.49	72.5	71.3	CET107
147	20	40	-.04	.37	.90	-.6	.79	-.7	.57	.49	72.5	71.3	TELP01
167	20	40	-.04	.37	.84	-1.1	.72	-1.1	.61	.49	72.5	71.3	TELP21
168	20	40	-.04	.37	.95	-.3	.82	-.6	.54	.49	72.5	71.3	TELP22
172	20	40	-.04	.37	.73	-1.9	.70	-1.1	.66	.49	87.5	71.3	TELP26
185	20	40	-.04	.37	.86	-.9	.77	-.8	.59	.49	77.5	71.3	TELP39
222	20	40	-.04	.37	.82	-1.2	.75	-.9	.61	.49	77.5	71.3	TELP76
231	20	40	-.04	.37	.93	-.4	.86	-.4	.54	.49	77.5	71.3	TELP85
256	20	40	-.04	.37	.93	-.4	.83	-.6	.55	.49	72.5	71.3	USFP16
259	20	40	-.04	.37	.72	-2.1	.60	-1.6	.68	.49	82.5	71.3	USFP19
263	20	40	-.04	.37	.83	-1.1	.82	-.6	.59	.49	77.5	71.3	USFP23
294	20	40	-.04	.37	.89	-.7	.79	-.7	.57	.49	72.5	71.3	USFP54
340	20	40	-.04	.37	1.27	1.7	1.40	1.4	.32	.49	52.5	71.3	USFP100
353	20	40	-.04	.37	.66	-2.5	.56	-1.9	.71	.49	87.5	71.3	USFP113
71	19	40	-.17	.37	1.37	2.3	1.36	1.3	.26	.49	50.0	71.4	CET71
90	19	40	-.17	.37	1.16	1.1	1.19	.7	.38	.49	70.0	71.4	CET90
97	19	40	-.17	.37	.88	-.8	.79	-.7	.57	.49	75.0	71.4	CET97
103	19	40	-.17	.37	1.21	1.4	1.37	1.3	.34	.49	65.0	71.4	CET103
118	19	40	-.17	.37	1.09	.7	1.00	.1	.44	.49	65.0	71.4	CET118
145	19	40	-.17	.37	.85	-1.0	.77	-.8	.59	.49	80.0	71.4	CET145
171	19	40	-.17	.37	.72	-2.1	.60	-1.6	.68	.49	85.0	71.4	TELP25
195	19	40	-.17	.37	.78	-1.5	.69	-1.1	.63	.49	80.0	71.4	TELP49
203	19	40	-.17	.37	.90	-.6	.79	-.7	.56	.49	75.0	71.4	TELP57
217	19	40	-.17	.37	.85	-1.0	.83	-.5	.58	.49	85.0	71.4	TELP71
220	19	40	-.17	.37	.84	-1.1	.76	-.8	.60	.49	80.0	71.4	TELP74
284	19	40	-.17	.37	1.12	.8	1.14	.6	.41	.49	70.0	71.4	USFP44
310	19	40	-.17	.37	.88	-.8	.75	-.9	.58	.49	75.0	71.4	USFP70
31	18	40	-.31	.37	.92	-.5	.99	.1	.52	.48	77.5	71.3	CET31
40	18	40	-.31	.37	.98	-.1	1.09	.4	.48	.48	77.5	71.3	CET40
50	18	40	-.31	.37	.87	-.9	.73	-.9	.58	.48	77.5	71.3	CET50
53	18	40	-.31	.37	.92	-.5	.89	-.3	.54	.48	67.5	71.3	CET53
63	18	40	-.31	.37	.87	-.8	.74	-.9	.58	.48	72.5	71.3	CET63
111	18	40	-.31	.37	.87	-.9	.73	-.9	.58	.48	72.5	71.3	CET111
119	18	40	-.31	.37	.96	-.2	.93	-.2	.51	.48	77.5	71.3	CET119
132	18	40	-.31	.37	1.01	.1	.91	-.2	.49	.48	67.5	71.3	CET132
153	18	40	-.31	.37	1.04	.3	1.02	.2	.45	.48	72.5	71.3	TELP07
158	18	40	-.31	.37	.99	.0	1.04	.2	.48	.48	72.5	71.3	TELP12
161	18	40	-.31	.37	.69	-2.3	.57	-1.6	.69	.48	82.5	71.3	TELP15
164	18	40	-.31	.37	1.18	1.2	1.26	.9	.35	.48	72.5	71.3	TELP18
221	18	40	-.31	.37	.79	-1.4	.66	-1.2	.63	.48	77.5	71.3	TELP75
223	18	40	-.31	.37	.87	-.9	.88	-.3	.56	.48	77.5	71.3	TELP77
271	18	40	-.31	.37	.78	-1.5	.70	-1.1	.63	.48	82.5	71.3	USFP31
282	18	40	-.31	.37	.82	-1.3	.69	-1.1	.61	.48	77.5	71.3	USFP42
293	18	40	-.31	.37	.87	-.8	.73	-.9	.58	.48	72.5	71.3	USFP53
295	18	40	-.31	.37	1.01	.1	.91	-.2	.49	.48	72.5	71.3	USFP55
300	18	40	-.31	.37	.92	-.5	.83	-.5	.55	.48	72.5	71.3	USFP60
309	18	40	-.31	.37	.96	-.2	.88	-.3	.52	.48	72.5	71.3	USFP69
10	17	40	-.44	.37	1.08	.5	.99	.1	.44	.48	72.5	71.5	CET10
18	17	40	-.44	.37	.73	-1.9	.61	-1.4	.66	.48	82.5	71.5	CET18
42	17	40	-.44	.37	1.19	1.3	1.17	.6	.36	.48	67.5	71.5	CET42
55	17	40	-.44	.37	1.12	.8	1.05	.3	.41	.48	62.5	71.5	CET55
62	17	40	-.44	.37	.82	-1.2	.75	-.8	.60	.48	82.5	71.5	CET62
72	17	40	-.44	.37	1.11	.7	1.02	.2	.42	.48	62.5	71.5	CET72
79	17	40	-.44	.37	.71	-2.1	.59	-1.5	.68	.48	82.5	71.5	CET79
120	17	40	-.44	.37	.86	-.9	.78	-.7	.57	.48	77.5	71.5	CET120
140	17	40	-.44	.37	.83	-1.1	.72	-.9	.60	.48	72.5	71.5	CET140
148	17	40	-.44	.37	.84	-1.0	.71	-1.0	.59	.48	72.5	71.5	TELP02
151	17	40	-.44	.37	1.10	.7	1.08	.4	.42	.48	67.5	71.5	TELP05

PHAKITI: TEST-TAKERS' PERFORMANCE APPRAISALS, APPRAISAL CALIBRATION, STATE-TRAIT STRATEGY USE,
AND STATE-TRAIT IELTS LISTENING DIFFICULTY IN A SIMULATED IELTS LISTENING TEST

176	17	40	-.44	.37	.70	-2.2	.59	-1.5	.68	.48	82.5	71.5	TELP30
181	17	40	-.44	.37	.87	-.8	.78	-.7	.57	.48	82.5	71.5	TELP35
186	17	40	-.44	.37	.86	-.9	.80	-.6	.57	.48	77.5	71.5	TELP40
187	17	40	-.44	.37	1.01	.1	.86	-.4	.49	.48	67.5	71.5	TELP41
194	17	40	-.44	.37	1.14	.9	1.12	.5	.39	.48	67.5	71.5	TELP48
196	17	40	-.44	.37	.96	-.2	.82	-.5	.52	.48	67.5	71.5	TELP50
216	17	40	-.44	.37	.70	-2.2	.58	-1.5	.68	.48	77.5	71.5	TELP70
224	17	40	-.44	.37	1.10	.7	1.06	.3	.42	.48	67.5	71.5	TELP78
226	17	40	-.44	.37	.66	-2.5	.55	-1.6	.70	.48	87.5	71.5	TELP80
233	17	40	-.44	.37	1.36	2.2	1.38	1.2	.25	.48	57.5	71.5	TELP87
285	17	40	-.44	.37	.84	-1.1	.73	-.9	.59	.48	77.5	71.5	USFP45
301	17	40	-.44	.37	1.23	1.5	1.21	.7	.34	.48	62.5	71.5	USFP61
306	17	40	-.44	.37	.92	-.5	.83	-.5	.54	.48	77.5	71.5	USFP66
4	16	40	-.58	.37	1.09	.6	1.13	.5	.41	.47	72.5	72.0	CET04
6	16	40	-.58	.37	1.15	1.0	1.11	.4	.38	.47	67.5	72.0	CET06
24	16	40	-.58	.37	.79	-1.4	.68	-1.0	.61	.47	82.5	72.0	CET24
54	16	40	-.58	.37	.98	.0	1.09	.4	.46	.47	77.5	72.0	CET54
57	16	40	-.58	.37	.97	-.1	.82	-.5	.51	.47	67.5	72.0	CET57
73	16	40	-.58	.37	1.14	.9	1.04	.2	.40	.47	72.5	72.0	CET73
95	16	40	-.58	.37	.89	-.7	.78	-.6	.55	.47	77.5	72.0	CET95
106	16	40	-.58	.37	.87	-.8	.75	-.7	.57	.47	77.5	72.0	CET106
239	16	40	-.58	.37	1.17	1.1	1.28	.9	.36	.47	62.5	72.0	TELP93
255	16	40	-.58	.37	.98	-.1	.90	-.2	.50	.47	72.5	72.0	USFP15
298	16	40	-.58	.37	1.22	1.4	1.08	.3	.36	.47	57.5	72.0	USFP58
303	16	40	-.58	.37	.86	-.9	.73	-.8	.57	.47	72.5	72.0	USFP63
313	16	40	-.58	.37	1.48	2.7	2.31	3.1	.09	.47	57.5	72.0	USFP73
8	15	40	-.71	.37	1.13	.9	1.15	.5	.38	.47	72.5	72.7	CET08
67	15	40	-.71	.37	1.27	1.6	1.43	1.2	.28	.47	62.5	72.7	CET67
74	15	40	-.71	.37	1.17	1.1	1.06	.3	.37	.47	72.5	72.7	CET74
85	15	40	-.71	.37	1.01	.1	.97	.0	.46	.47	72.5	72.7	CET85
91	15	40	-.71	.37	1.01	.1	.93	-.1	.47	.47	67.5	72.7	CET91
98	15	40	-.71	.37	1.18	1.1	1.13	.5	.36	.47	67.5	72.7	CET98
113	15	40	-.71	.37	.97	-.1	.94	.0	.48	.47	82.5	72.7	CET113
114	15	40	-.71	.37	.94	-.3	.89	-.2	.51	.47	77.5	72.7	CET114
152	15	40	-.71	.37	1.07	.5	1.01	.1	.43	.47	67.5	72.7	TELP06
156	15	40	-.71	.37	1.08	.6	1.04	.2	.42	.47	67.5	72.7	TELP10
162	15	40	-.71	.37	.70	-2.0	.58	-1.3	.66	.47	82.5	72.7	TELP16
170	15	40	-.71	.37	.92	-.5	.80	-.5	.53	.47	72.5	72.7	TELP24
177	15	40	-.71	.37	.74	-1.7	.61	-1.2	.64	.47	82.5	72.7	TELP31
206	15	40	-.71	.37	.99	.0	.86	-.3	.49	.47	72.5	72.7	TELP60
212	15	40	-.71	.37	1.05	.4	1.07	.3	.43	.47	67.5	72.7	TELP66
228	15	40	-.71	.37	.95	-.2	.88	-.3	.50	.47	67.5	72.7	TELP82
247	15	40	-.71	.37	.79	-1.4	.74	-.7	.60	.47	87.5	72.7	USFP07
260	15	40	-.71	.37	.87	-.8	.80	-.5	.55	.47	72.5	72.7	USFP20
279	15	40	-.71	.37	.95	-.2	.84	-.4	.51	.47	77.5	72.7	USFP39
286	15	40	-.71	.37	.74	-1.7	.62	-1.2	.64	.47	82.5	72.7	USFP46
304	15	40	-.71	.37	.97	-.2	.91	-.2	.49	.47	72.5	72.7	USFP64
305	15	40	-.71	.37	.88	-.7	.76	-.6	.55	.47	77.5	72.7	USFP65
312	15	40	-.71	.37	1.23	1.4	1.16	.5	.33	.47	62.5	72.7	USFP72
48	14	40	-.85	.38	1.03	.2	.87	-.2	.46	.46	70.0	73.6	CET48
64	14	40	-.85	.38	1.04	.3	1.04	.2	.43	.46	70.0	73.6	CET64
87	14	40	-.85	.38	1.02	.2	.90	-.2	.46	.46	70.0	73.6	CET87
94	14	40	-.85	.38	.96	-.2	.81	-.4	.50	.46	75.0	73.6	CET94
101	14	40	-.85	.38	.85	-.9	.74	-.6	.56	.46	75.0	73.6	CET101
135	14	40	-.85	.38	.96	-.2	.81	-.4	.50	.46	75.0	73.6	CET135
137	14	40	-.85	.38	1.26	1.5	1.19	.6	.30	.46	60.0	73.6	CET137
197	14	40	-.85	.38	1.08	.5	.93	-.1	.43	.46	70.0	73.6	TELP51
236	14	40	-.85	.38	.84	-.9	.81	-.4	.55	.46	80.0	73.6	TELP90
246	14	40	-.85	.38	1.34	1.9	1.47	1.2	.23	.46	65.0	73.6	USFP06
262	14	40	-.85	.38	1.08	.5	1.01	.2	.42	.46	65.0	73.6	USFP22
267	14	40	-.85	.38	1.05	.4	.88	-.2	.45	.46	70.0	73.6	USFP27
270	14	40	-.85	.38	.92	-.5	.99	.1	.49	.46	80.0	73.6	USFP30
275	14	40	-.85	.38	1.11	.7	.95	.0	.41	.46	65.0	73.6	USFP35
288	14	40	-.85	.38	1.09	.6	.93	-.1	.42	.46	65.0	73.6	USFP48
364	14	40	-.85	.38	1.16	1.0	1.03	.2	.37	.46	65.0	73.6	USFP124
7	13	40	-1.00	.38	.84	-.9	.82	-.4	.55	.45	80.0	74.7	CET07
15	13	40	-1.00	.38	.95	-.3	.81	-.4	.50	.45	75.0	74.7	CET15
20	13	40	-1.00	.38	1.08	.5	1.18	.6	.39	.45	70.0	74.7	CET20
32	13	40	-1.00	.38	1.20	1.1	1.22	.6	.32	.45	70.0	74.7	CET32
68	13	40	-1.00	.38	.91	-.5	.76	-.5	.52	.45	80.0	74.7	CET68
75	13	40	-1.00	.38	.88	-.6	.77	-.5	.53	.45	80.0	74.7	CET75
76	13	40	-1.00	.38	1.06	.4	.98	.1	.42	.45	75.0	74.7	CET76
82	13	40	-1.00	.38	1.05	.3	1.09	.4	.41	.45	80.0	74.7	CET82
93	13	40	-1.00	.38	1.15	.9	1.17	.5	.35	.45	75.0	74.7	CET93
110	13	40	-1.00	.38	.93	-.4	.76	-.5	.52	.45	75.0	74.7	CET110
122	13	40	-1.00	.38	.88	-.7	1.04	.2	.51	.45	80.0	74.7	CET122
163	13	40	-1.00	.38	.80	-1.2	.63	-.9	.59	.45	80.0	74.7	TELP17
166	13	40	-1.00	.38	.61	-2.5	.49	-1.5	.70	.45	90.0	74.7	TELP20
192	13	40	-1.00	.38	.99	.0	.98	.1	.45	.45	75.0	74.7	TELP46
258	13	40	-1.00	.38	1.34	1.8	1.37	1.0	.23	.45	65.0	74.7	USFP18
5	12	40	-1.15	.39	1.03	.2	1.37	.9	.38	.44	80.0	75.9	CET05
13	12	40	-1.15	.39	.99	.0	.82	-.3	.47	.44	75.0	75.9	CET13
27	12	40	-1.15	.39	1.18	1.0	1.24	.7	.31	.44	75.0	75.9	CET27

PHAKITI: TEST-TAKERS' PERFORMANCE APPRAISALS, APPRAISAL CALIBRATION, STATE-TRAIT STRATEGY USE,
AND STATE-TRAIT IELTS LISTENING DIFFICULTY IN A SIMULATED IELTS LISTENING TEST

59	12	40	-1.15	.39 .81	-1.1 .66	-.7 .57	.44 85.0	75.9	CET59
121	12	40	-1.15	.39 .99	.0 1.25	.7 .41	.44 80.0	75.9	CET121
133	12	40	-1.15	.39 1.10	.6 .94	.0 .40	.44 75.0	75.9	CET133
157	12	40	-1.15	.39 .97	-.1 .85	-.2 .47	.44 75.0	75.9	TELP11
188	12	40	-1.15	.39 1.12	.7 1.11	.4 .37	.44 75.0	75.9	TELP42
191	12	40	-1.15	.39 1.16	.9 1.32	.8 .32	.44 70.0	75.9	TELP45
205	12	40	-1.15	.39 .98	.0 1.10	.4 .43	.44 80.0	75.9	TELP59
16	11	40	-1.30	.40 1.33	1.6 1.33	.8 .23	.43 65.0	77.1	CET16
17	11	40	-1.30	.40 .95	-.2 .89	-.1 .46	.43 80.0	77.1	CET17
33	11	40	-1.30	.40 .84	-.8 .73	-.5 .53	.43 80.0	77.1	CET33
100	11	40	-1.30	.40 .86	-.7 .73	-.5 .52	.43 80.0	77.1	CET100
112	11	40	-1.30	.40 .78	-1.2 .66	-.7 .57	.43 85.0	77.1	CET112
115	11	40	-1.30	.40 .84	-.8 .77	-.4 .53	.43 80.0	77.1	CET115
174	11	40	-1.30	.40 1.10	.6 .96	.1 .38	.43 75.0	77.1	TELP28
182	11	40	-1.30	.40 .89	-.5 .75	-.4 .51	.43 85.0	77.1	TELP36
218	11	40	-1.30	.40 .92	-.4 .75	-.4 .50	.43 80.0	77.1	TELP72
219	11	40	-1.30	.40 .93	-.3 .77	-.4 .49	.43 80.0	77.1	TELP73
232	11	40	-1.30	.40 .90	-.5 .71	-.5 .51	.43 75.0	77.1	TELP86
238	11	40	-1.30	.40 .73	-1.5 .60	-.8 .60	.43 85.0	77.1	TELP92
241	11	40	-1.30	.40 1.12	.7 1.33	.8 .33	.43 75.0	77.1	USFP01
388	11	40	-1.30	.40 1.08	.5 .89	-.1 .41	.43 70.0	77.1	USFP148
69	10	40	-1.47	.41 1.06	.4 1.13	.4 .37	.42 82.5	78.5	CET69
89	10	40	-1.47	.41 1.04	.3 .97	.1 .39	.42 82.5	78.5	CET89
102	10	40	-1.47	.41 1.17	.8 1.10	.4 .31	.42 77.5	78.5	CET102
123	10	40	-1.47	.41 .92	-.3 .66	-.6 .50	.42 72.5	78.5	CET123
125	10	40	-1.47	.41 .82	-.9 .67	-.6 .54	.42 82.5	78.5	CET125
128	10	40	-1.47	.41 1.21	1.0 1.36	.8 .26	.42 72.5	78.5	CET128
143	10	40	-1.47	.41 1.47	2.1 1.60	1.2 .11	.42 67.5	78.5	CET143
165	10	40	-1.47	.41 .87	-.6 .69	-.5 .51	.42 82.5	78.5	TELP19
208	10	40	-1.47	.41 1.22	1.1 1.82	1.5 .22	.42 77.5	78.5	TELP62
302	10	40	-1.47	.41 .76	-1.2 .85	-.1 .55	.42 82.5	78.5	USFP62
11	9	40	-1.64	.42 .87	-.5 1.12	.4 .44	.40 82.5	80.1	CET11
14	9	40	-1.64	.42 .81	-.8 .74	-.3 .51	.40 87.5	80.1	CET14
19	9	40	-1.64	.42 1.16	.8 1.13	.4 .30	.40 77.5	80.1	CET19
22	9	40	-1.64	.42 1.09	.5 1.01	.2 .35	.40 77.5	80.1	CET22
29	9	40	-1.64	.42 1.06	.4 .94	.1 .37	.40 82.5	80.1	CET29
39	9	40	-1.64	.42 1.01	.1 .86	-.1 .41	.40 82.5	80.1	CET39
51	9	40	-1.64	.42 1.09	.5 .95	.1 .36	.40 77.5	80.1	CET51
56	9	40	-1.64	.42 .88	-.5 1.00	.2 .46	.40 87.5	80.1	CET56
66	9	40	-1.64	.42 .93	-.2 .84	-.1 .45	.40 82.5	80.1	CET66
83	9	40	-1.64	.42 1.10	.5 .95	.1 .35	.40 72.5	80.1	CET83
124	9	40	-1.64	.42 .95	-.1 .92	.0 .43	.40 82.5	80.1	CET124
144	9	40	-1.64	.42 1.04	.2 1.14	.4 .37	.40 77.5	80.1	CET144
198	9	40	-1.64	.42 1.00	.1 .76	-.3 .43	.40 77.5	80.1	TELP52
311	9	40	-1.64	.42 .75	-1.2 .51	-.9 .58	.40 82.5	80.1	USFP71
78	8	40	-1.82	.44 1.23	1.0 1.56	1.0 .21	.39 80.0	81.7	CET78
88	8	40	-1.82	.44 1.10	.5 .90	.0 .35	.39 80.0	81.7	CET88
386	8	40	-1.82	.44 1.38	1.5 1.34	.7 .15	.39 70.0	81.7	USFP146
21	7	40	-2.02	.46 .89	-.3 .94	.2 .42	.37 82.5	83.5	CET21
41	7	40	-2.02	.46 1.05	.3 .81	-.1 .36	.37 82.5	83.5	CET41
61	7	40	-2.02	.46 1.23	.9 1.69	1.1 .18	.37 77.5	83.5	CET61
77	7	40	-2.02	.46 1.33	1.2 1.52	.9 .13	.37 82.5	83.5	CET77
200	7	40	-2.02	.46 1.05	.3 .84	.0 .36	.37 82.5	83.5	TELP54
360	7	40	-2.02	.46 1.21	.8 1.79	1.2 .17	.37 82.5	83.5	USFP120
65	6	40	-2.24	.48 1.00	.1 1.13	.4 .32	.35 87.5	85.5	CET65
80	6	40	-2.24	.48 1.38	1.3 2.12	1.4 .06	.35 82.5	85.5	CET80
134	6	40	-2.24	.48 1.09	.4 1.47	.8 .27	.35 87.5	85.5	CET134
35	5	40	-2.49	.52 1.25	.8 1.33	.6 .15	.33 85.0	87.7	CET35
36	5	40	-2.49	.52 1.24	.8 1.67	1.0 .10	.33 90.0	87.7	CET36
86	5	40	-2.49	.52 1.03	.2 1.59	.9 .27	.33 85.0	87.7	CET86
84	4	40	-2.78	.56 1.02	.2 1.24	.6 .26	.30 90.0	90.0	CET84
MEAN	20.1	40.0	-.01	.40 1.01	.0 .99	.0	75.7	75.5	
S.D.	7.3	.0	1.11	.05 .18	1.0 .34	.8	7.5	4.8	