

**Université de Montréal**

Building a Validity Argument for the Listening Component of the *Test de Connaissance du français* in the Context of Quebec Immigration.

Par:  
Angel Manuel Arias De Los Santos

Département d'Administration et fondements de l'éducation, Faculté des Sciences de  
l'éducation

Thèse de doctorat présentée à la Faculté des sciences de l'éducation en vue de l'obtention du  
grade de Ph.D. en mesure et évaluation en éducation

Mars, 2018

© Angel Manuel Arias De Los Santos, 2018

## Résumé

L'évaluation linguistique est une pratique omniprésente dans les contextes d'immigration, utilisée comme une méthode de collecte de données pour évaluer la capacité des immigrants à communiquer dans la langue du pays d'accueil afin de promouvoir l'intégration sociale et économique ainsi que la productivité au travail (McNamara & Shohamy, 2008). Contrairement aux tests d'anglais, peu d'attention est accordée à l'interprétation et à l'utilisation des scores aux tests en français, ce qui incite - et demande – de la validation des scores pour justifier l'utilisation des tests. Cette étude, qui fait appel aux avancées de la théorie de la validité des tests (Kane, 2006, 2013), construit un argumentaire de validité pour la composante de la compréhension orale du *Test de connaissance du français* (TCF) dans le contexte de l'immigration au Québec. La théorie de la validité des tests a évolué considérablement depuis le modèle tripartite traditionnel de contenu, de prédiction et de construit (Cronbach et Meehl, 1955), a été conceptualisée comme un construit unitaire (Messick, 1989) et, plus récemment, a été théorisée en termes d'argumentation (Kane, 2006, 2013), empruntant des concepts de modèles d'inférence (Toulmin [1958], 2003), qui englobent des inférences de score, de généralisation, d'explication, d'extrapolation et de décision, ayant des rôles importants dans un argumentaire de validité.

Dans une approche de validité fondée sur l'argumentation, les affirmations relatives aux instruments de mesure sont composées de garanties qui doivent être étayées par des études empiriques, qui sont fondamentales pour les affirmations, mais qui appuient également les inférences qui autorisent chacune des affirmations de l'argumentaire. Plus précisément, cette étude a analysé des données empiriques pour appuyer les inférences de scores, de généralisation et d'explication, en proposant trois questions de recherche portant sur la représentativité du construit du TCF, le fonctionnement différentiel d'items et l'utilité de la technique de collecte de données. Les questions ont porté sur les sous-compétences de compréhension orale dont le TCF évalue, le fonctionnement différentiel des items (FDI) selon le genre, la langue maternelle, l'âge et l'emplacement géographique des candidats ainsi que le fonctionnement des items à choix multiples dans l'évaluation de la compréhension orale en langue seconde.

Bien que de nombreux modèles statistiques et de mesure soient couramment disponibles pour analyser les données de réponse aux tests, cette étude a privilégié l'analyse factorielle

confirmatoire (AFC) pour examiner les sous-compétences de compréhension orale opérationnalisées dans le TCF, en spécifiant des modèles suivant les suggestions d'un comité d'experts. Le modèle unidimensionnel de Rasch a permis de générer les paramètres de difficulté des items entre les sous-groupes d'intérêt pour effectuer les analyses FDI. Et le modèle à réponses nominales (MRN) a été utilisé pour modéliser les options des items à choix multiples. Les résultats issus de ces trois études ont permis d'étayer chacune des inférences retenues dans l'argumentaire de validité pour le TCF.

Selon les modèles d'AFC recommandés par le comité d'experts, les résultats suggèrent que les deux versions analysées dans cette étude évaluent principalement la compréhension de l'information explicitement énoncée dans le discours oral, sous représentant ainsi le construit. Quelques items visaient l'habileté à inférer des idées implicites et la compréhension du sujet général ou de l'idée principale, mais cette dernière sous-compétence ne se retrouvait que dans une seule version du test, ce qui suggère que les versions ne sont pas comparables. L'analyse FDI a identifié de nombreux items dans les versions du test et entre les sous-groupes d'intérêt, mais très peu ont été associés à un biais potentiel, qui comprenait la perception de la voix, le genre littéraire et la familiarité du vocabulaire. Par conséquent, étant donné que de nombreux items signalés pour fonctionnement différentiel ne pouvaient pas être associés à un biais potentiel, la réponse à cette question est partiellement élaborée et atténuée l'argumentaire de validité. Les résultats du MRN suggèrent que la plupart des items fonctionnaient bien, tandis que d'autres avaient potentiellement deux bonnes réponses.

L'approche de la validité fondée sur l'argumentation s'avère utile pour regrouper des études empiriques dans un ensemble cohérent permettant d'étayer et de justifier l'interprétation et les utilisations du TCF dans un contexte d'immigration, qui peut à son tour servir à remédier les points faibles constatés, en fournissant un moyen d'atténuer les réfutations potentielles qui menacent la validité de l'argument. Certaines mises en garde dans le cadre de validation sont également soulignées et concernent l'accessibilité des données pour aborder les inférences d'extrapolation et de décision dans des contextes d'immigration, mais comme Newton et Shaw (2014, p. 142) le soulignent: « l'approche de validité fondée sur l'argumentation sous-tend le fait que la validation n'est pas simplement une étude isolée, mais un programme: potentiellement un programme très intensif ». Et ce programme peut inclure des parties

prenantes importantes comme les représentants gouvernementaux qui peuvent aider à compléter l'argumentaire de validité du TCF en matière d'immigration au Québec.

**Mots-clés** : La validité fondée sur l'argumentation; la compréhension orale en langues secondes; l'évaluation linguistique pour l'immigration; l'analyse factorielle confirmatoire; le fonctionnement différentiel d'items; le modèle à réponses nominales.

## Abstract

Language testing is a ubiquitous practice in immigration contexts used as a data collection procedure to assess immigrants' ability to communicate in the language of the host country to promote social as well as economic integration and productivity in the workplace (McNamara & Shohamy, 2008). Unlike English tests, little attention has been directed to the interpretation and uses of scores from French proficiency tests, which prompts – indeed, requires – validation research to justify test use. Drawing on advances in test validity theory (Kane, 2006, 2013), this study builds a validity argument for the listening component of the *Test de connaissance du français* (TCF) in the context of Quebec immigration. Test validity theory has evolved considerably since the traditional tripartite model of content, predictive and construct components (Cronbach & Meehl, 1955), have been conceptualized as a unitary construct (Messick, 1989) and more recently have been theorized in terms of argumentation (Kane, 2006, 2013), borrowing concepts from models of inference (Toulmin [1958], 2003), which include scoring, generalization, explanation, extrapolation and decision inferences that play key roles in a validity argument.

In an argument-based approach to validity, claims about testing instruments are composed of warrants that must be supported by backings in the form of empirical studies, which are foundational for the claims, but also support the inferences that authorize each of the claims in the argument. More specifically, this study gathered empirical evidence to support the scoring, generalization and explanation inferences, proposing three research questions that addressed construct representation, potential bias and test method usefulness. The questions were concerned with the listening subskills that the TCF assesses, differential item functioning (DIF) across gender, first language, age, and geographical location as well as the option functioning of multiple choice (MC) items in the assessment of second language listening comprehension. Although multiple statistical and measurement models are readily available to analyze test response data, this study privileged confirmatory factor analysis (CFA) to examine the listening subskills operationalized in the TCF, specifying the models following suggestions from a panel of experts. The unidimensional Rasch model was used to generate the difficulty parameters across subgroups of interest to perform the DIF analyses. And the nominal response model

(NRM) was used to model the response options of the MC items. The results from these three studies yielded backings for each of the selected inferences in the validity argument for the TCF. Based on the CFA models recommended by the panel of experts, the results suggested that the TCF test forms under study primarily assess examinees' understanding of explicitly stated information in aural discourse, thereby underrepresenting the listening construct. A few items were found to target the ability to infer implicit ideas and understanding of the general topic or main idea, however, this latter subskill was only found in one test form, suggesting that the forms are not equivalent. The DIF analysis flagged multiple items across test forms and between the subgroups of interest, but very few were associated to potential bias, which included speech perception, literary genre and vocabulary familiarity. Thus, given that many items flagged for DIF could not be associated to a potential bias, this question was partially answered and attenuates the validity argument. The results from the NRM suggested that most items functioned well while others were potentially double keyed.

The argument-based approach to validity proved helpful in putting together empirical evidence into a coherent whole to support and build a case for the interpretation and uses of the TCF in the context of immigration, which in turn can be used to address the identified weaknesses, providing a means to attenuate the potential rebuttals that threaten the validity of the argument. Some caveats in the validation framework were also outlined and relate to the accessibility of data to address the extrapolation and decision inferences in immigration contexts, but as Newton and Shaw (2014, p. 142) advocated "the argument-based approach underlies the fact that validation is not simply a one-off-study but a program: potentially a very intensive program". And this program can include key stakeholders such as government officials that help complete the validity argument for the TCF in Quebec immigration.

**Keywords:** Argument-based validity; second language listening; language assessment for immigration; confirmatory factor analysis; differential item functioning; nominal response model.

## Table of Contents

<b>Résumé .....</b>	<b>i</b>
<b>Abstract .....</b>	<b>iv</b>
<b>List of Tables .....</b>	<b>xi</b>
<b>List of Figures.....</b>	<b>xiv</b>
<b>Dedication.....</b>	<b>xvii</b>
<b>Acknowledgements.....</b>	<b>xviii</b>
<b>Introduction.....</b>	<b>1</b>
<b>Chapter 1: Research Problem .....</b>	<b>5</b>
<b>1.1 Issues in Language Assessment for Immigration .....</b>	<b>5</b>
1.1.1 The power of tests in modern immigration .....	7
1.1.2 Language policy and ideologies in the context of immigration .....	9
1.1.3 Language assessment for immigration in Canada .....	11
1.1.3.1 Canadian immigration through the province of Quebec .....	12
1.1.3.2 Language benchmarks in Quebec .....	14
1.1.3.3 Quebec’s immigration points-based system.....	16
1.1.3.4 Waves of Quebec prospective immigrants registering for the TCF .....	17
1.1.3.5 The immigrants’ journey in the province of Québec .....	18
<b>1.2 Test Fairness in Language Assessment.....</b>	<b>22</b>
<b>1.4 Significance of the Study .....</b>	<b>25</b>
1.4.1 Social relevance .....	25
1.4.2 Scientific significance .....	26
<b>1.5 Objectives of the Study.....</b>	<b>26</b>
1.5.1 Research questions.....	27
<b>Chapter 2: Theoretical Framework .....</b>	<b>28</b>
<b>2.1 Validity.....</b>	<b>29</b>
<b>2.2 The Current Dominant View of Validity .....</b>	<b>41</b>
2.2.1 Scoring inference .....	45

2.2.2	Generalization inference .....	45
2.2.3	Explanation inference .....	46
2.2.4	Extrapolation inference .....	47
2.2.5	Decision inference .....	47
<b>2.3</b>	<b>Other Views on the Concept of Validity in the 21<sup>st</sup> Century.....</b>	<b>48</b>
<b>2.4</b>	<b>Listening Comprehension: Issues and Historical Perspectives.....</b>	<b>50</b>
2.4.1	The bottom up view in listening comprehension .....	54
2.4.1.1	Acoustic lexical recognition in listening comprehension.....	56
2.4.2	The top down view in listening comprehension.....	58
2.4.3	Metacognition in L2 listening comprehension.....	61
2.4.4	The interactive view in listening comprehension .....	64
<b>2.5</b>	<b>Assessing L2 Listening Comprehension .....</b>	<b>66</b>
2.5.1	Construct definition .....	66
2.5.2	Issues in listening assessment.....	68
2.5.3	Measurement models in L2 listening assessments .....	69
<b>2.6</b>	<b>Validity in Language Assessment.....</b>	<b>72</b>
2.6.1	Theoretical impact on language assessment.....	72
2.6.2	Validation research in language assessment .....	74
2.6.3	Confirmatory factor analysis in language assessment .....	79
2.6.4	Differential item functioning in language assessment .....	82
2.6.5	Distractor analysis of MC items in language assessment .....	84
2.6.6	Validation of language tests in the context of immigration .....	85
	<b>Chapter 3: Methodology .....</b>	<b>88</b>
<b>3.1</b>	<b>Test de connaissance du français .....</b>	<b>88</b>
3.1.1	Test administration procedures.....	89
<b>3.2</b>	<b>Participants.....</b>	<b>90</b>
<b>3.2</b>	<b>Panel of experts.....</b>	<b>91</b>
<b>3.2</b>	<b>Evaluation of Claim One Using Confirmatory Factor Analysis .....</b>	<b>92</b>
3.2.1	Confirmatory factor analysis .....	92
3.2.1.1	CFA model specification.....	93

3.2.1.2 CFA model identification .....	94
3.2.1.3 Guidelines for model identification .....	95
3.2.1.4 CFA model estimation .....	95
<b>3.3 Evaluation of Claim Two Using Rasch-based DIF analysis and the Standardization DIF Method .....</b>	<b>97</b>
3.3.1 The Rasch model .....	99
3.3.1.1 Dimensionality .....	100
3.3.1.2 Local independence.....	101
3.3.1.3 Rasch fit statistics .....	102
3.3.1.4 Rasch-based differential item functioning .....	104
3.3.2 The Standardization index.....	106
<b>3.4 Evaluation of Claim Three Using the Nominal Response Model.....</b>	<b>107</b>
3.4.1 The nominal response model (NRM) .....	107
<b>3.5 Software used in this study.....</b>	<b>109</b>
<b>Chapter 4: Results.....</b>	<b>110</b>
<b>4.1 Research Question One: TCF Listening Subskills .....</b>	<b>112</b>
4.1.1 Descriptive statistics .....	112
4.1.2 Item Coding.....	117
4.1.2 Confirmatory Factor Analysis .....	119
4.1.2.1 The Correlated two-factor model for Form A .....	122
4.1.2.2 The Correlated two-factor model for Form B.....	126
4.1.2.3 Similarities and differences between the modified CFA models .....	129
<b>4.2 Research Question Two: TCF Fairness .....</b>	<b>130</b>
4.2.1 Rasch fit analyses .....	131
4.2.1 Assessing dimensionality and local independence .....	137
4.2.1.1 Examining dimensionality through principal component analysis of standardized Rasch residuals .....	138
4.2.1.2 Examining local independence .....	143
4.2.2 Rasch modeling of the TCF listening Test.....	145
4.2.3 Differential item functioning across gender groups .....	148

4.2.4 Differential item functioning across first language .....	150
4.2.5 Differential item functioning across age for immigration classes .....	163
4.2.6 Differential item functioning across age with similar sample sizes .....	166
4.2.7 Differential item functioning across North and West African groups.....	168
4.2.8 DIF across first language using the Standardization Index.....	174
4.2.9 Summary of DIF analysis across grouping variables .....	176
<b>4.3 Research Question Three: Usefulness of Multiple Choice Items .....</b>	<b>176</b>
4.3.1 Descriptive statistics .....	177
4.3.2 Data modeling with the nominal response model .....	179
<b>Chapter 5: Discussion .....</b>	<b>184</b>
<b>5.1 Implications of CFA Results Substantiating Research Question One .....</b>	<b>184</b>
<b>5.2 Complementarity of Confirmatory Factor Analysis and Rasch Modeling .....</b>	<b>186</b>
<b>5.3 Implications of DIF Results for Research Question Two.....</b>	<b>187</b>
5.3.1 Unidimensionality and local independence in Rasch-based DIF .....	188
5.3.2 DIF investigation implications for gender .....	189
5.3.3 DIF investigation implications for first language (L1) .....	189
5.3.4 DIF investigation implications for age groups .....	190
5.3.5 DIF investigation implications for geographical location.....	191
5.3.6 Implications for the interpretation of DIF analyses .....	192
<b>5.4 Implications of NRM Results for Research Question Three.....</b>	<b>193</b>
5.4.1 Implications of the option frequency of MC items.....	193
5.4.2 Implications of the NRM modelling.....	194
<b>5.4 Validity Argument of the Test de connaissance du français .....</b>	<b>195</b>
5.4.1 Scoring inference .....	197
5.4.2 Generalization inference .....	199
5.4.3 Explanation inference .....	201
5.4.3 Remarks of the extrapolation and decision inferences .....	203
<b>Chapter 6: Conclusion .....</b>	<b>204</b>
<b>6.1 Summing up of the Validity Argument.....</b>	<b>204</b>
<b>6.2 Overview and Appraisal of the Research Questions .....</b>	<b>204</b>

**6.3 Validity Argument of the Test de connaissance du français ..... 206**  
**6.4 Limitations of the Study ..... 207**  
**6.5 Recommendation for Future Research..... 208**  
**References..... 210**  
**Appendix A: Descriptive Statistics for Nationalities and First Languages ..... 237**  
**Appendix B: Correlation of Residuals, DIF Analyses and Graphics ..... 243**  
**Appendix C: Nominal Response Model Parameters and Graphics ..... 267**  
**Appendix D: R codes..... 273**

## List of Tables

Table I Linguistic Points-based System for Quebec Immigration.....	16
Table II Registration for the <i>Test de connaissance du français</i> between 2006 - 2013 .....	17
Table III Research phases to build a validity argument for the <i>Test de connaissance du français</i> .....	27
Table IV Facets of Validity from Messick (1989).....	40
Table V Research Questions and Statements for the Interpretation/Use Argument of the TCF L2 Listening Test.....	87
Table VI Nationalities of Test Takers of the TCF for Immigration to Quebec .....	90
Table VII First Language and Age Groups of Test Takers Taking the TCF for Immigration to Quebec .....	90
Table VIII Wu and Adams (2013) guidelines for mean square fit statistics .....	103
Table IX TCF Listening Validation Studies.....	111
Table X Descriptive Statistics for Subgroups of Examinees in Form A.....	112
Table XI Descriptive Statistics for Subgroups of Examinees in Form B.....	114
Table XII Item Level Analysis for Form A.....	116
Table XIII Item Level Analysis for Form B.....	116
Table XIV Distribution of Experts and Agreement Category for Form A.....	118
Table XV AC <sub>1</sub> Coefficient and Expert Agreement for Form A .....	118
Table XVI Distribution of Experts and Agreement Category for Form B.....	119
Table XVII AC <sub>1</sub> Coefficient and Expert Agreement for Form B.....	119
Table XVIII Summary Results for CFA models in Form A and Form B Listening Sections	121
Table XIX Standardized Factor Loadings and Associated Critical Ratios, Explained and Residual Variances for Form A.....	124
Table XX Standardized Factor Loadings and Associated Critical Ratios, Explained and Residual Variances for Form B .....	127
Table XXI Difficulty Measures, Fit Statistics, Fit Criteria and Listening Subskills for Form A .....	131

Table XXII Difficulty Measures, Fit Statistics, Fit Criteria and Listening Subskills for Form B .....	133
Table XXIII Difficulty Measures, Fit Statistics, Fit Criteria and Subskills for Form A after Item Deletion.....	135
Table XXIV Difficulty Measures, Fit Statistics, Fit Criteria and Subskills for Form B after Item Deletion.....	136
Table XXV Standardized Residual Variance in Eigenvalue Units for Form A .....	139
Table XXVI Standardized Residual Variance in Eigenvalue Units for Form A without Noisy Items .....	141
Table XXVII Standardized Residual Variance in Eigenvalue Units for Form B.....	141
Table XXVIII Standardized Residual Variance in Eigenvalue Units for Form B without Noisy Items .....	143
Table XXIX Largest Residual Correlations for both Test Forms.....	144
Table XXX Largest Residual Correlations for both Test Forms without Noisy Items.....	144
Table XXXI Rasch-based Summary Statistics for both Test Forms after Item Deletion .....	147
Table XXXII Option Frequencies for Items in Form A.....	177
Table XXXIII Option Frequencies for Items in Form B.....	178
Table XXXIV Descriptive Statistics for Nationalities in Form A .....	237
Table XXXV Descriptive Statistics for First Language in Form A.....	239
Table XXXVI Descriptive Statistics for Nationalities in Form B .....	240
Table XXXVII Descriptive Statistics for First Language in Form B .....	242
Table XXXVIII Correlation of Standardized Rasch Residuals in Form A .....	243
Table XXXIX Correlation of Standardized Rasch Residuals in Form B.....	243
Table XL DIF Contrasts for Gender Subgroups in Form A .....	244
Table XLI DIF Contrasts for L1 Subgroups in Form A.....	246
Table XLII DIF Contrasts for Age Subgroups by Immigration Policy in Form A.....	248
Table XLIII DIF Contrasts for Age Subgroups by Similar Sample Sizes in Form A .....	250
Table XLIV DIF Contrasts for Geographical Location Subgroups in Form A.....	253
Table XLV DIF Contrasts for Gender Subgroups in Form B .....	255
Table XLVI DIF Contrasts for L1 Subgroups in Form B .....	257
Table XLVII DIF Contrasts for Age Subgroups by Immigration Policy in Form B .....	260

Table XLVIII DIF Contrasts for Age Subgroups by Similar Sample Sizes in Form B.....	262
Table XLIX DIF Contrasts for Geographical Location Subgroups in Form B .....	265
Table L Intercept and Slope NRM Parameters for Males and Females in Form A.....	267
Table LI Intercept and Slope NRM Parameters for Males and Females in Form B.....	268
Table LII Intercept and Slope NRM Parameters with Standard Errors for the Parsimonious Model of Form A.....	269
Table LIII Intercept and Slope NRM Parameters with Standard Errors for the Parsimonious Model of Form B.....	271

## List of Figures

Figure 1 Immigrant's Journey in the Province of Quebec (adapted from Saville, 2009).....	20
Figure 2 Example structure of an interpretation/use argument for test fairness claim of the TCF listening component in the context of Quebec immigration.....	44
Figure 3 Example of a Reflective Measurement Model in L2 Listening Comprehension.....	72
Figure 4 Validity arguments using Rasch measurement (adapted from Aryadoust, 2009) .....	79
Figure 5 Illustration of a two-factor model with correlated factors.....	94
Figure 6 Score Distribution of Test Form A.....	113
Figure 7 Score Distribution of Test Form B.....	115
Figure 8 Correlated Two-Factor Model for Form A.....	123
Figure 9 Modified Correlated Two-Factor Model for Form A.....	125
Figure 10 Modified Correlated Three-Factor Model for Form B.....	126
Figure 11 Modified Three-Factor Correlated Model for Form B.....	129
Figure 12 Standardized Residual Plot for the First Contrast in Form A.....	140
Figure 13 Standardized Residual Plot for the First Contrast in Form B .....	142
Figure 14 Wright Maps for both Test Forms.....	146
Figure 15 Item Characteristic Curves for Items 4, 8, 16 and 23 in Form A.....	149
Figure 16 Item Characteristic Curves for Items 9 and 17 in Form B.....	150
Figure 17 Item Characteristic Curves for Items 1, 2, 3, 4, 6 and 7 in Form A.....	152
Figure 18 Item Characteristic Curves for Items 8 through 13 in Form A.....	153
Figure 19 Item Characteristic Curves for Items 14 through 19 in Form A.....	155
Figure 20 Item Characteristic Curves for Items 20 through 25 in Form A.....	156
Figure 21 Item Characteristic Curves for Items 26 and 28 in Form A .....	157
Figure 22 Item Characteristic Curves for Items 1, 4, 5, 6, 7 and 8 in Form B .....	158
Figure 23 Item Characteristic Curves for Items 9 through 14 in Form B .....	159
Figure 24 Item Characteristic Curves for Items 15 through 20 in Form B .....	161
Figure 25 Item Characteristic Curves for Items 21, 23, 24, 25, and 27 in Form B .....	162
Figure 26 Item Characteristic Curves for Items 1, 11, 17 and 24 in Form A.....	164
Figure 27 Item Characteristic Curves for Item 8 in Form B .....	165

Figure 28 Item Characteristic Curves for Items 1, 2, 11, 17, 24 and 28 in Form A .....	167
Figure 29 Item Characteristic Curves for Item 6 in Form B .....	168
Figure 30 Geographical Location for North and West African Countries in the DIF Study ..	169
Figure 31 Item Characteristic Curves for Items 1, 2, 3, 4, 8 and 13 in Form A .....	170
Figure 32 Item Characteristic Curves for Items 15, 16, 22, 24, 25 and 26 in Form A .....	171
Figure 33 Item Characteristic Curves for Items 1, 2, 3, 4, 5 and 7 in Form B .....	172
Figure 34 Item Characteristic Curves for Items 9, 10, 11, 21, 23 and 25 in Form B.....	173
Figure 35 Standardization Indices for Chinese and Russian L1 Speakers .....	175
Figure 36 Standardization Indices for Chinese and Portuguese L1 Speakers .....	175
Figure 37 NRM Item Category Response Curves for Items 1, 14, 26, and 30 in Form A.....	181
Figure 38 NRM Item Category Response Curves for Items 1, 21, 27, and 30 in Form B.....	182
Figure 39 Validity Argument of the <i>Test de connaissance du français</i> Listening Test.....	197
Figure 40 Scoring Inference of the TCF Listening Test with Warrants, Assumptions, Rebuttals, and Backings .....	198
Figure 41 Generalization Inference of the TCF Listening Test with Warrants, Assumptions, Rebuttals, and Backings.....	200
Figure 42 Explanation Inference of the TCF Listening Test with Warrants, Assumptions, Rebuttals, and Backings.....	202
Figure 43 Item Characteristic Curves for Gender DIF Subgroups in Form A .....	245
Figure 44 Item Characteristic Curves for L1 DIF Subgroups in Form A .....	247
Figure 45 Item Characteristic Curves for Age DIF Subgroups by Immigration Policy in Form A .....	249
Figure 46 Item Characteristic Curves for Age DIF Subgroups by Similar Sample Sizes in Form A .....	252
Figure 47 Item Characteristic Curves for Geographical Location DIF Subgroups in Form A	254
Figure 48 Item Characteristic Curves for Gender DIF in Form B.....	256
Figure 49 Item Characteristic Curves for L1 DIF Subgroups in Form B .....	259
Figure 50 Item Characteristic Curves for Age DIF Subgroups by Immigration Policy in Form B .....	261
Figure 51 Item Characteristic Curves for Age DIF Subgroups by Sample Sizes in Form B ..	264
Figure 52 Item Characteristic Curves for Geographical Location DIF Subgroups in Form B	266

Figure 53 NRM Item Characteristic Response Curves for Form A.....270  
Figure 54 NRM Item Characteristic Response Curves for Form B.....272

### **Dedication**

To my daughter Jadah Arias Fernandez who enriches my life enormously. Also, I would like to present this work as a token of appreciation to the youth of my hometown, Serie 081, Rio San Juan, Dominican Republic, hoping to motivate our younger generations to aim high, while remembering a powerful quote from Aristotle, “You are what you repeatedly do”.

## Acknowledgements

A doctoral degree trains you to be a scholar and a researcher in your chosen field (Jerrard & Jerrard, 1998), but at the end of the journey, you also realize that the scholar and researcher you have become has been highly moulded by the community of practice and by the people who somehow were part of that journey. I would like to thank my supervisor, Professor Jean-Guy Blais, who has enormously influenced my thinking on educational measurement, challenging me to be a critical thinker and consumer of research. During my doctoral studies, Jean-Guy also organized a private seminar on test validity theory and another on measurement. He invited several professors and fellow PhD students to engage in discussions on these two topics, which I consider are at the heart of our field.

I am also indebted to my co-supervisor, Professor Michel Laurier, who has greatly challenged my thinking on language testing and assessment, encouraging me to explore thoroughly the concepts of ethicality and consequences in this field. I wish to express my appreciation to Professor Nathalie Loye who always supported my doctoral journey and who organized a second private seminar on test validity theory with the collaboration of Professor Sébastien Béland.

Many thanks go to the *Fonds Québécois de recherche sur la société et la culture* (FQRSC), the *Group de recherche interuniversitaire sur l'évaluation et la mesure en éducation à l'aide des TIC* (GRIÉMÉtic, [under the direction of Professor Jean-Guy Blais]) and the *Faculté des sciences de l'éducation* at the University of Montréal, which provided funding for my doctoral studies on many different levels.

Also, I am sincerely grateful to the *Centre international d'études pédagogiques* (CIEP) for providing access to the data, namely, Dr. Sébastien Georges, who was the liaison between the CIEP and I. Similarly, my thanks and appreciation go to Professor Beverly Baker for her collaboration in the recruitment of participants for part of the study.

Finally, I would like to thank my fellow PhD colleague students, who were always supportive, especially Maxim Morin, who challenged me to learn R to present elegant graphics parsimoniously, enhancing data visualization.

## Introduction

The assessment of language proficiency for immigration purposes is a complex and consequential practice challenging governments, policy makers, and language assessment professionals in the 21<sup>st</sup> century. Governments are increasingly monitoring immigration waves to ensure the social integration of newcomers and the decisions made on behalf of prospective immigrants can be life-changing events. An important claim advocating the use of language assessment in immigration contexts is the idea of social and economic integration as well as access to different levels of society (e.g., labor market, education, etc.). Therefore, when one considers the various consequences attached to language assessment in the context of immigration, it is difficult to imagine the use of invalid test scores as a factor to grant or deny permanent residence status to potential immigrants. Thus, it stands to say that validation research plays an integral and a pivotal role in this context.

To immigrate to Canada through the province of Quebec, proof of language proficiency is a requirement under certain immigration programs (e.g., Quebec's skilled temporary foreign worker). As such, applicants often provide test scores from designated language tests as a requirement to apply for the Quebec selection certificate (*Certificat de sélection du Québec*), an official document issued by the provincial government of Quebec and required prior to applying for Canadian permanent residence. To ensure that test scores are used ethically, the practice of language testing for immigration purposes calls for the collaboration of various key stakeholders from different disciplines, including language policy makers and language assessment professionals, to comply with testing standards (Saville, 2009).

This study<sup>1</sup> draws on Kane's (2001, 2006, 2013) validation framework to build a validity argument for the interpretation and uses of the test scores obtained on listening component of the *Test de connaissance du français*<sup>2</sup> (TCF), one of the several tests used by the government of Quebec for immigration purposes and currently accepted by the federal government of Canada for citizenship applications (Government of Canada, 2018). The study is organized in six

---

<sup>1</sup> The term *study* is used as an overarching concept to encompass the three empirical studies that were conducted in this thesis.

<sup>2</sup> The *Test de connaissance du français* is owned by the Centre International d'Études Pédagogiques (CIEP), a French testing firm based in Paris, France. The TCF is used for immigration purposes by the Québec Government and for citizenship by the Federal Government of Canada.

chapters: research problem, theoretical framework, methodology, results, discussion and conclusion. The paragraphs below provide a roadmap and a brief introduction to the chapters that follow.

Chapter 1 provides an overview of the issues of language assessment for immigration including the power of tests, language ideologies in immigration contexts, and the responsibility of key stakeholders as participants in the selection process of immigrants. The chapter also draws on the current *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association & National Council on Measurement on Education [AERA, APA & NCME], 2014, hereafter *Standards*) to emphasize the importance of test fairness in the context of immigration. Although the Standards can be considered as an American product, “they can play a key role in the testing and assessment community worldwide” (Zumbo, 2014, p. 32), and provide guidance for test score validation. The chapter additionally presents a general overview of immigration through the province of Quebec and the policies that are in place to award points to permanent residence applications. Drawing on the work of Saville (2009), the immigrant’s journey in the province of Quebec is mapped out describing the unique features and challenges of the immigration process in this Canadian Province. The chapter ends with the research questions of the study and the methods that were used to address the questions.

Chapter 2 presents the theories underlining the theoretical framework of the study, which includes validity theory and second language (L2) listening theory. In the section of validity theory, the chapter provides an account of the evolution of the concept of validity in the field of educational and psychological measurement and its semantics from the inception of the term (e.g., Cureton, 1951; Garret, 1937; Guildford, 1946) until recent developments (e.g., Kane, 2006, 2013). This is achieved by presenting dialectic views on validity theory and the expansions that have resulted in the current dominant view for test score validation (i.e., argument-based approaches to validity). This section concludes with Kane’s (2006, 2013) proposed two-fold model for the validation of the interpretation and uses of test scores, consisting of an interpretation use argument (IUA) and a validity argument (VA). Although Kane (2001, 2006, 2013), Chapelle, Enright and Jamieson (2008), and Shaw, Crisp and Johnson (2012) proposed several inferences to be included in the validation of test scores (e.g., scoring, generalization, explanation, extrapolation and decision), this study only presents empirical

evidence to support the *scoring*, *generalization*, and *explanation* inferences for the validity argument made on behalf of the TCF.

In the section of second language (L2) listening theory, the chapter reviews the intricacies of this rather complex language skill (e.g., Buck, 2001; Field, 2008a, 2013). This review draws on the different processes (i.e., bottom up, top down, and interactive processes) that take place when decoding aural discourse from a stream of sounds. It also includes a discussion on metacognition and its relation to regulatory strategies that listeners might draw upon to make sense of aural discourse to achieve comprehension. In addition, the chapter provides a definition of reflective measurement models in educational measurement (Edwards, 2011) and relates reflective measurement models to how constructs of L2 listening comprehension are conceptualized. Moreover, this section reviews validation research in language assessment concerning the factor structure of tests, conceptual articles on validity in language assessment, and the investigation of differential item functioning in language assessment contexts. Finally, the research questions are recast in the form of claims that are included in the interpretation use argument of the TCF.

Chapter 3 describes the participants of the study, provides an overview of the TCF, and spells out the analytical tools that were used to accumulate the empirical evidence to support the scoring, generalization, and explanation inferences of the validity argument for the listening test scores of the TCF. The chapter elaborates on the measurement models that were chosen to analyze the data and the assumptions underlying each of the models. The models include: confirmatory factor analyses, the unidimensional Rasch model and the nominal response model. Chapter 4 provides the results of the empirical studies (i.e., the backings) that were conducted to substantiate each of the warrants supporting the scoring, generalization and explanation inferences in the validity argument. The results stemmed from a panel of experts, item level data and test taker demographics of two test forms of the TCF (hereafter, Form A and Form B).

Chapter 5 discusses the results and connects them to findings from previous validation research of L2 listening assessments and language assessment in general. The results from the nominal response model (Bock, 1972) are discussed in relation to a small body of research from educational and psychological measurement because the model has been underutilized and have received relatively limited application to real-world measures. Furthermore, this chapter presents the scoring, generalization and explanation inferences in the form of diagrams that

follow Toulmin's (1958, 2003) model of inference. This approach is widely adopted in validation research for language assessments (Aryadoust, 2013; Chapelle, Enright & Jamieson, 2008) and in validity theory (e.g., Kane, 2006, 2013; Mislevy, Steinberg & Almond, 2003) to illustrate how test score validation and development can be framed in terms of arguments.

Chapter 6 provides the summary and the conclusion of the study, outlines the limitations, and presents how future validation research can provide the backings for the extrapolation and decision inferences in the validity argument for the TCF. Finally, this chapter also reflects on the importance of collaboration among key stakeholders and how such collaboration can potentially improve language testing practices in the province of Quebec to ultimately help with the social and economic integration of immigrants.

## **Chapter 1: Research Problem**

The assessment of second language proficiency for immigration purposes has been a lingering concern in applied linguistics, namely in the fields of language testing and assessment, and language policy. Often, heated debates on the topic of language assessment for immigration stem from the social complexity of language (Fulcher, 2010; McNamara & Roever, 2006) and the agendas of policymakers (e.g., government officials) who advocate political, social, and economic ideologies for the prosperity and benefit of society (Shohamy, 2001, 2006; McNamara, 2009). In this arena, language assessment instruments have become integral to immigration policies and the quality of the data yielded by those instruments is an essential concern to language assessment professionals, test users and consumers of test information. That concern is heightened whenever test results inform important decisions regarding immigration as the use of test scores in high-stakes contexts often have meaningful social consequences.

The role of test developers and test users of language assessments is, or should be, to provide evidence that test scores are being interpreted reliably and validly to justify interpretation and use (Chapelle, Enright & Jamieson, 2008). This practice is consistent with the *Standards* and the code of ethics for language assessment (ILTA, 2000), as validation is seen as a joint responsibility of the test developer and the test user (i.e., key stakeholders). In this vein, the test developer is responsible to provide relevant evidence and a rationale in support of any score interpretations for specified uses intended by the developer, whereas the test user is ultimately responsible for evaluating the evidence in the setting in which the test is to be used (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education [AERA, APA, & NCME], 2014). Following this line of reasoning, the purpose of this study is to provide a validity argument for the interpretation and uses of the test scores obtained on the listening component of the TCF in the context of Quebec immigration.

### **1.1 Issues in Language Assessment for Immigration**

Language tests are viewed not only as pedagogical tools, but especially as social and political instruments that have a strong impact on education, immigration, and can determine social order (Shohamy, 2006). The complexity of the testing enterprise can be traced to, for

example, notions introduced by Shohamy (2007) who elaborated on the term *use oriented testing* referring to a view of testing that is concerned with what happens to test takers, to the knowledge that is created through tests and to key stakeholders involved in the process of teaching, learning and testing, as well as the consequences of testing. This view of testing is also consistent with the idea that tests should not be viewed in isolation [or a vacuum] but rather as connected to psychological, social, and political variables that have effects on curricula, ethicality, social classes, bureaucracy, politics, knowledge, inclusion and exclusion (Messick, 1981, 1994, 1996). The view of *use oriented testing* is operationalized in immigration contexts when the validation of proposed interpretations and uses of test scores are important to key stakeholders, thus ensuring responsible testing practices that are not detrimental to potential newcomers.

Concerns in language testing for immigration purposes have led to thorough reflections and scholarly dialogues regarding its justice, fairness, and validity. This has gained popularity among the community of language testing professionals as the topic is hotly debated across several countries where language policies are problematic and questionable. (See Blackledge, 2009; Cooke, 2009; Kunnan, 2009; May, 2008; McNamara, 2009; McNamara & Shohamy, 2008; Shohamy, 2001; Shohamy & Kanza, 2009; Van Avermaet & Rocca, 2013). In addition, an indication that there has been an purposeful effort to address the problems related to language assessment for immigration is the scholarly attention the topic has received in a series of publications in peer-reviewed journals (for example, *Language Assessment Quarterly*, special issue on language tests for citizenship, immigration and asylum in 2009), funded research programs (e.g., discourses on language and integration funded by the UK Arts and Humanities Research Council), as well as in high-profile colloquia in widely recognized conference venues such as the American Association for Applied Linguistics (AAAL, 2007, 2008), the Language Testing Research Colloquium (LTRC, 2007), and the Language Association of Language Testers in Europe (ALTE, 2006, 2008, 2011, 2014).

Furthermore, ALTE (2011, 2014) has organized in various venues forums on Language Assessment for Migration and Integration (LAMI) to address questions regarding implications for access, policymaking, fairness, and practical solutions in the arena of language assessment for immigration purposes. Remarkably, the dialogue within the community of language testers often questions the fairness and ethicality of imposing knowledge of a national language as an

immigration requirement in some countries (e.g., Australia, the United Kingdom and the Netherlands). However, little attention have been paid to the insightful discussions and proposed ideas in these venues.

### **1.1.1 The power of tests in modern immigration**

A growing number of countries require that prospective immigrants provide language test results from designated testing firms as a condition for obtaining the right to enter the country in the first place (e.g., the Netherlands), and to obtain permanent residency and eventually citizenship (McNamara & Shohamy, 2008). Why are language tests used in the reinforcement of such policies? There are many reasons, but one crucial factor is the control tests render to policy makers (Shohamy, 2001). In most societies, tests results have been constructed as symbols of success, achievement and mobility, and reinforced by dominant social and educational institutions as major criteria of success, quality and value (Shohamy, 2001; Spolsky, 1995). As such, “tests have been associated with standards, objectivity and merit, and in the context of immigration, are associated with productivity in the workplace and in society” (McNamara & Shohamy, 2008, p. 89).

Modern migration involves political, economic, social, cultural and educational issues that need to be addressed drawing on multidisciplinary approaches or shared frameworks that call for a variety of expertise and coordinated actions led by key stakeholders (Bolli, 2013), and third parties such as government-funded language training programs and designated language testing firms. Saville (2009, p. 25) referred to this initiative as a framework of “inter- or transdisciplinary collaboration on the theme of migration”. The shared framework proposed by Saville (2009) and the view of use oriented testing evoked by Shohamy (2007) calls for the collaboration of key stakeholders from different disciplines to put in place research agendas to ensure that the decisions being made based on test scores are reliable, fair, and valid. However, despite the recommendations made by experts in different relevant disciplines akin to immigration policy and language assessment, there seems to be a lack of collaborative efforts among key stakeholders and research on this issue is scanty.

The work of Shohamy (2001) distinguishes five perspectives for the language testing community to act ethically when using tests: ethical perspective, awareness raising perspective, all consequences perspectives, perspectives of sanctioning, and perspective of shared

responsibility and open communication. These perspectives are relevant to the context of language assessment for immigration. The latter perspective, that is *perspective of shared responsibility and open communication*, also addresses the concerns raised by Bolli (2013) and Saville (2009) calling for the participation of key stakeholders (e.g., policy makers, government officials, language testers, test takers) when developing and using a test for immigration purposes. Thus, it stands to say that “if language tests become a key discriminator determining entry [to a given country], it is crucial that any test used is fair and fit the purpose so that immigrants are not unfairly denied access at any stage in their journey as migrants” (Van Avermaet & Rocca, 2013, p. 12). This is consistent with recent views in validity theory in which test validity refers to “the *utility* and *appropriateness* of a test for a *particular purpose* requiring multiple sources of evidence” (Sireci, 2007, p. 477).

Moreover, language requirements have been introduced as an important component in the context of immigration because it is thought, and widely accepted, that knowledge of the language spoken in the host country potentially promotes immigrants’ social integration (Van Avermaet & Rocca, 2013). This assessment practice is not new and has become the norm in many developed countries across the globe regardless of the controversies and consequences attached to it (Fulcher, 2010; McNamara, 2005). From a language assessment point of view, Fulcher (2010) argued that the widespread use of language tests across different arenas such as education, international mobility, and economic policy planning, makes language testing controversial, thus creating the need for empirical research that addresses key issues that threatens the validity of the interpretation and uses of test scores in immigration contexts.

Linguistic gatekeeping mechanisms are not new and can in fact be traced back to ancient times. For instance, the ubiquitously cited “National College Entrance Test” in China (the Gaokao) is normally used as a traditional example to illustrate control of the powerful (e.g., governments) over society. In keeping with China’s education testing system, each Chinese university sets its entrance scores (cut scores) and allocates seats to each province. And test scores determine which college or university each student will attend. The reasons why language tests have come about are almost always related to control and power of the nation-state over individuals and selection is at the heart of such practices.

In the context of immigration, linguistic gatekeeping mechanisms date back as far as the early 1900’s. For instance, The Immigration Restriction Act of 1901 in Australia was intended

to control immigration by administering a dictation test. Applicants had to transcribe a passage of fifty words that might have been dictated in any European language, not necessarily in English, selected at the will of the person administering the test. Failure to pass the test was used by immigration officials to stop immigrants from entering Australia based on language skills (McNamara, 2009). This example illustrates the abuses of language assessment in immigration contexts. In a sense, the way language tests are used can perpetuate the idea of gatekeeping (Spolsky, 1997) and can serve as restrictive access mechanisms.

### **1.1.2 Language policy and ideologies in the context of immigration**

From a language policy perspective, Wright (2008) argued that language tests in the context of immigration could also be tools to discourage applications from temporary immigrants (e.g., visitors, international students, etc.,) who would like to remain permanently in the host country. As noted earlier, language assessment can influence different spheres of society, makes the political discourse on immigration very complex, diverse, and occasionally, extremely sensitive and hotly debated. Current practices of language assessment for immigration purposes can also be attributed to mass migration and globalization, where involved nations seek to control immigration flows to foster an internationally integrated world economy (Cooke, 2009). In fact, the political discourse would seem to promote an economic immigration model (Wright, 2008) to bolster the declining labor market of countries with an aging population problem (e.g., Canada).

Language assessment remains a pervasive discourse in the agenda of government officials as immigration rising record levels motivate federal (e.g., Canada) and provincial (e.g., Quebec) governments to be highly involved in the selection process of newcomers (i.e., authorized immigrants). The increasing interest of government participation in the selection process of newcomers can be further explained by the multiplicity of ethnic groups seeking to immigrate to the host country. In this regard, May (2008) identified two key ideologies that call for a thorough reflection regarding the role of language as a requirement for immigration. That is, (1) whether speaking a national language should be a requirement to immigrate or stay permanently in the host country and subsequently become a citizen, or (2) whether monolingualism should be imposed over often-multilingual immigration waves. Whether a monolingual, a bilingual, or a multilingual ideology is adopted by the host country, language assessment should be supported

by validation research that justify test score interpretation and use. This can potentially warrant that language tests result in more good than harm to the individuals applying to become permanent residents of a given host country.

Given the high-stakes involved in immigration decisions, government-appointed policy makers play an important role in the development of language policies akin to immigration. For some, monolingualism ideologies serve to promote social equality, based on the premise that promoting a national language ensures immigrants' full civic and economic participation. This latter ideology is somewhat rooted in the idea that "the role of people's language as an ethnic attribute affects their earnings" (Grin, 2006, p. 79). Ricento (2006) labeled stakeholders who subscribed to this ideology as "assimilationists" who consider that the key to social equality and economic integration is the immediate shift from the immigrants' mother tongue to the language spoken by the majority of the population in the host country. In a sense, the principle of "la loi Toubon", in which legal provisions favor the use of the national language, is fully enacted by governments seeking to control diverse linguistic groups and promote an assimilationist policy (Schiffman, 2006).

May (2008) further elaborated on the concepts of monolingualism (i.e., the view adopted under the assimilation principle) and plurilingualism. He suggested that the discussions about language assessment for immigration and integration have spurred an important dialectics in which modern governments can potentially adopt one out of two positions: pluralism or civism. The former refers to the idea of maintaining the political claims and the recognition of diverse minority groups in a pluralist society, whereas the latter refers to the exclusion of the cultural characteristics (including language) of minority groups to promote social cohesion. Although nation states could integrate both pluralism and civism to help promote education, social economy and integration, they tend to adhere to a civic position (or liberal pluralism<sup>3</sup>) to promote a linguistically homogenous civic realm (May, 2008). This seems to be ubiquitous in the political discourse for immigration at a transnational level across both sides of the Atlantic (e.g., North America and Europe). It is worth noting that in Europe, a very inclusive-pluralistic political discourse is pervasive in organizations such as the Council of Europe, and UNESCO.

---

<sup>3</sup> Gordon (1981) defined liberal pluralism as the absence, even prohibition of any ethnic, religious or national group possessing separate standing before the law of government.

However, at a national level, host countries seem to encourage a retrenchment, rather stringent process, in the selection of potential immigrants. In other words, there has been a larger emphasis in many developed countries towards stricter conditions for people who want to apply for residence rights or for citizenship (See Blackledge, 2009 for an example of the United Kingdom).

Language assessment in the context of Quebec immigration that promotes trans-disciplinary collaboration among key stakeholders (Saville, 2009) can develop a more distinctive voice by relating research to questions and dilemmas created by migration. In turn, this can potentially address critical issues such as language ideologies, language policies for immigration, immigrants' integration and full participation in Quebec society. This is key to establish successful immigration programs because immigration waves tend to fluctuate and vary across time. Thus, if key stakeholders are working together addressing all the potential issues that may rise, better solutions can be put in place or tailored according to different immigration phenomena. In terms of language assessment, this entails collecting and analyzing data on designated language assessments and making appropriate reference to applied validity theory to provide empirical evidence that supports the use of language tests in such contexts. In turn, this can serve as the basis for the development of validity arguments addressing issues of validity in terms of purposes and outcomes (Saville, 2009). Providing evidence of validity seems a way forward to justify the use of language assessments in the context of immigration and can yield a better understanding of the interface between language assessment and immigration.

### **1.1.3 Language assessment for immigration in Canada**

As noted earlier, having to demonstrate “acceptable” language proficiency has increasingly become the norm in the selection process of immigrants in many countries. In Canada, there are several immigration programs (e.g., Federal Skilled Worker Program, Canadian Experience Class, Federal Skill Trades Program, etc.) through which temporary immigrants can apply to move or stay permanently in the country. As such, language requirements for immigration purposes have a significant impact on candidates' applications, determining access to the country. Applicants under these immigration programs must demonstrate an acceptable level of language proficiency (based on the Canadian Language

Benchmarks<sup>4</sup>) in either English or French to be considered as potential permanent residents. Proficiency in French and/or English in Canada has been a requirement for many years, but in most cases until 2010 (for skilled workers) and 2012 (for semi- or low skilled job) this had been subject to testing on a relatively informal and often arbitrary basis (e.g., by means of a short interview with an immigrant official with no training either in linguistic analysis or in language testing). Currently, linguistic proficiency has now emerged as one of the key conditions for granting permission to stay permanently and for naturalization in Canada, thus, more formal mechanisms for testing have been introduced. There exist three language tests that have been designated by Citizenship and Immigration Canada for this purpose (e.g., the Canadian English Language Proficiency Index Program [CELPIP], the International English Language Testing System [IELTS], and the Test d'évaluation de français [TEF]). In addition to these assessment instruments, the *Test de connaissance du français* (TCF) and other tests offered by the *Centre International d'Études Pédagogiques* and *La Chambre de Commerce d'Industrie de Paris* are designated at the provincial level in Quebec and used to apply for the Quebec selection certificate. At the time of this writing (March, 2018), the Federal Government of Canada is also accepting previous test scores used to immigrate to Quebec as proof of language proficiency for citizenship applications.

### **1.1.3.1 Canadian immigration through the province of Quebec**

The immigration journey to Canada through the province of Quebec is different from other Canadian provinces. Applicants must obtain the *Quebec selection certificate* (QSC) through the *Ministère d'Immigration, Diversité et Inclusion* before their application can be processed at the federal level (i.e., the QSC precludes the application for permanent residence). Quebec administers its own immigration programs with selection criteria that are distinct from those of federal and other provincial immigration programs. There are several programs such as the skilled worker program and the Quebec experience program through which international students, graduates, and temporary workers may apply to stay permanently in the province and eventually become Canadian citizens if desired. The province has also adopted a language policy

---

<sup>4</sup> The Canadian Language Benchmark (CLB) standard is a descriptive scale of language ability in English and French as a second language written as 12 benchmarks or reference points along a continuum from basic to advanced.

(*Charte de la langue française*, 1977, 2017) that promotes immigrants' linguistic integration through francization programs or mandatory schooling in French.

Knowledge of French is an important component of the selection criteria used to award points to applications for the QSC. The *Ministère d'Immigration, Diversité et Inclusion* (MIDI), formerly the *Ministère de l'Immigration et des Communautés Culturelles* (MICC), have designated third-party testing agencies to provide language assessment services to potential immigrants submitting test scores attesting their language proficiency in listening, speaking, reading and writing in French or English. It is assumed these standardized tests ensure that the evaluation of candidates' knowledge of French and English is accurate, consistent and fair, however, no empirical evidence is publicly available on this matter. The following tests and diplomas are recognized by the *Ministère*:

- The Test d'évaluation du français adapté pour le Québec (TEFAQ) of the Chambre de commerce et d'industrie de Paris Île-de-France (CCIP-IDF).
- The Test de connaissance du français pour le Québec (TCF-Québec) of the Centre international d'études pédagogiques (CIEP).
- The Test d'évaluation du français (TEF) of the CCIP-IDF.
- The Test d'évaluation du français pour le Canada (TEF Canada) of the CCIP-IDF.
- The Test de connaissance du français (TCF) of the CIEP.
- The Diplôme d'études en langue française (DELFF) of the CIEP.
- The Diplôme approfondi de langue française (DALF) of the CIEP.

Official attestation of French proficiency has a greater impact than English in the application for the QSC. Moreover, listening and speaking proficiency have a greater impact than reading and writing on the designation of points. And to consider the knowledge of French as a meaningful factor in the application for the QSC, applicants need to reach a minimum threshold from which points are awarded. That is, at least a level 7 (intermediate<sup>5</sup>) on the *Échelle québécoise des niveaux de compétence en français des personnes immigrantes adultes*. This is equivalent to a level B2 of the common European framework of reference for languages.

---

<sup>5</sup> Comprend sans aide, quand la situation est prévisible ou partiellement prévisible, le contenu de conversations ou de discours en français standard et au débit normal portant sur des thèmes concrets liés à des besoins courants (*when the situation is predictable, or partially predictable, the interlocutor understands without assistance the content of standard French at a normal speech rate in real-world situations*).

### 1.1.3.2 Language benchmarks in Quebec

*The Échelle québécoise des niveaux de compétence en français des personnes immigrantes adultes* is a descriptive scale of language proficiency in French as a Second (or additional) Language (FSL) written as 12 benchmarks, levels, or reference points along a continuum from basic to advanced. The Canadian Language Benchmark (CLB) inspired the language framework (i.e., the échelle québécoise) used in Québec, but this latter reflects a more realistic view of the immigrants' language needs in this Canadian province (Desbiens, Laurier & Leroux, 2011). The scale also reflects the progression of knowledge and skills that underlie basic, intermediate, and advanced ability among adult learners. Each of the benchmarks or levels contains a general description regarding what the person can do and French language instances that describe superficially the language ability of a person at a specific level. The levels describe the competency statements, which are language performance tasks that an individual should be able to demonstrate in the four major language skills: listening, speaking, reading, and writing. Similar to the *Canadian Language Benchmark and the Common European Framework of Reference for languages*, the *Échelle Québécoise*<sup>6</sup> provides a common framework of reference on French language in the province of Québec.

In general terms, the *Échelle Québécoise des niveaux de compétence en français* is :

- A series of milestones consisting of 12 levels across 3 stages comprised of 4 levels each focusing on achieving communication tasks;
- A set of descriptions illustrated by indicators, grouped by levels, which describe the language ability of non-francophone immigrants at different stages of their learning experience of French;
- A common framework of reference for learning, teaching, and assessing adult French as a second language in Québec; and
- A tool to develop tailored language programs to the need of Quebec's allophone<sup>7</sup> immigrants.

---

<sup>6</sup> <https://www.immigration-quebec.gouv.qc.ca/publications/fr/langue-francaise/Echelle-niveaux-competences.pdf>

<sup>7</sup> An immigrant whose first language is neither French nor English

Each level describes characteristics of speaking, listening, reading and writing ability, and general profile descriptions describe what a person can do in each language skill as presented in the indicators at different levels of proficiency. The 12 levels are organized into three stages: basic, intermediate and advanced. Each stage is associated with a degree of language complexity and demand. Within each stage, there are four levels or benchmarks that progress in ability. This scale indicates a learner's progression across a stage. The three stages are summarized below.

*Basic Language Ability (Levels 1–4)*

Basic language ability encompasses abilities that are required to communicate in common and predictable contexts about basic needs, common everyday activities and familiar topics of immediate personal relevance (e.g., je m'appelle...je suis [nationalité et profession], pays d'origine...je parle [langues parlées]; *Translation:* My name is...I am [nationality and occupation], country of origin...I speak [spoken languages]).

*Intermediate Language Ability (Levels 5–8)*

Intermediate language ability encompasses abilities that allow fuller participation in a wider variety of contexts. It is the range of abilities required to function independently in most familiar situations of daily social, educational, and work-related life experience, and in some less predictable contexts (e.g., hier, je me suis levé à...j'ai pris un bain chaud...avez-vous compris? *Translation:* yesterday, I woke up at...I took a warm bath...did you understand? )

*Advanced Language Ability (Levels 9–12)*

Advanced language ability encompasses abilities required to communicate effectively, appropriately, accurately, and fluently about most topics in a wide range of contexts and situations, from predictable to abstract, from general to specialized, and from specific to nuanced, in communicatively demanding contexts. Learners at this stage have a sense of purpose and audience when communicating (including distance, politeness and formality factors, appropriate register and style, suitable volume or length of communication, accuracy and coherence of discourse, vocabulary range and precision). At this stage, communicating can involve using complex and creative discourse in different contexts (e.g., social life, stakes of society, professional life, use of complex phrases, paraphrase, etc.).

### 1.1.3.3 Quebec’s immigration points-based system

The selection of prospective immigrants to Canada through the province of Québec is based on a multidimensional points-based system in which language proficiency is one of several factors (e.g., education, work experience, age, employment offer, etc.) considered in the evaluation process of applications.

Table 1 below provides the points awarded to principal applicants under economic immigration programs based on different stages of the Quebec’s French language proficiency scale in listening, speaking, reading, and writing. According to Québec’s law of immigration (Loi sur l’immigration au Québec, 2017, p. 22), no points are awarded unless a minimum level of seven (intermediate)<sup>8</sup> is reached on one of the designated language tests when applying under the skilled worker program, the Quebec experience program (Quebec’s graduate and temporary worker), the self-employed worker program, or the entrepreneur program. As noted earlier, listening and speaking have more weight than reading and writing. In this order, both listening and speaking can account from 5 to 7 points each according to the level attained (7 through 12), whereas reading and writing only account for 1 point regardless of the level attained on the test.

Table I Linguistic Points-based System for Quebec Immigration

Skill	Stage 1 ( <i>Basic</i> )				Stage 2 ( <i>Intermediate</i> )				Stage 3 ( <i>Advanced</i> )			
	1	2	3	4	5	6	7	8	9	10	11	12
Listening					5	5	6	6	7	7		
Speaking					5	5	6	6	7	7		
Reading					1	1	1	1	1	1	1	1
Writing					1	1	1	1	1	1	1	1
<b>Level</b>	1	2	3	4	5	6	7	8	9	10	11	12

A maximum of 16 points on the linguistic dimension can be awarded to principal applicants if they obtain a score equivalent to a level 7 on a designated language test. If the application includes the spouse or the common-law partner, an additional 6 points (3 points for listening and 3 points for speaking only) can be awarded based on the spouse or common-law’s language ability. When added to the other selection criteria points (i.e., education, work experience, age, family in Quebec, employment offer, children, autonomous financial capacity, and spouse or

<sup>8</sup> Description générale pour la compréhension orale : comprend sans aide, quand la situation est prévisible ou partiellement prévisible, le contenu de conversations ou de discours en français standard et au débit normal portant sur des thèmes concrets liés à des besoins courants.

common law partner profile) the applicant must reach a score of 50 (without spouse or common-law partner) or 59 (with spouse or common-law partner) points to be selected as a potential immigrant. Candidates may also choose not to use the points awarded for the knowledge of French or English or both if they indicate this in their application for the QSC.

#### 1.1.3.4 Waves of Quebec prospective immigrants registering for the TCF

The number of candidates who decide to pass the TCF is increasing exponentially each year. Table 2 below provides the country of birth, age group, and gender of candidates who have registered for the TCF to submit their test scores to immigrate to Canada through the province of Québec. The information on the table represents an approximate number of fifteen countries, but over 50,000 candidates registered for the TCF between 2006 and 2013 (*Centre International d'Études Pédagogiques*, 2013). In 2006, the number of candidates who registered for TCF is considered miniscule compared to the astonishing 20,360 candidates who registered for the test in 2013. Over an eight-year span (2006-2013), the highest registration increased was recorded in 2012. Of particular interest are the nationals from China, Colombia, Brazil, India, and Russia whose first language is not French. Note that Morocco and Tunisia recognize French as a national language, Cameroon recognizes French as an official language, and in many African countries French is the lingua franca. France is not discussed here for obvious reasons, but it is worth noting that paradoxically test results from French nationals are also accepted for QSC applications.

Table II Registration for the *Test de connaissance du français* between 2006 - 2013

Country of birth	2006	2007	2008	2009	2010	2011	2012	2013
Algeria	-	1	5	10	10	25	667	659
Brazil	-	-	34	122	70	96	688	422
Cameroon	-	3	3	43	95	172	2381	2093
China	-	1	6	15	7	15	2309	1201
Colombia	-	-	20	48	73	114	1529	1167
Ivory Coast	-	-	1	0	0	6	2239	990
Egypt	-	7	21	25	46	166	1709	986
France	-	-	2	2	10	36	1759	2065
India	-	-	0	1	13	15	211	950
Iran	-	17	36	100	318	1075	3861	2345
Lebanon	5	38	20	28	36	70	488	644
Morocco	4	2	4	3	8	26	1666	1527
Russia	-	1	41	58	84	54	214	141
Senegal	-	-	1	1	1	3	608	793
Tunisia	-	24	30	36	64	94	1452	1218
<b>Total</b>	<b>9</b>	<b>94</b>	<b>224</b>	<b>492</b>	<b>835</b>	<b>1967</b>	<b>21781</b>	<b>17201</b>

Table 2 continued

<b>Age groups</b>								
0 through 14	-	-	1	-	-	-	1	1
15 through 25	0	38	52	75	155	292	3371	3111
26 through 44	8	118	328	705	1125	2235	20943	16533
45 through 64	1	10	14	13	34	63	650	703
65 and older	-	-	-	-	-	-	11	1
<b>Total</b>	<b>9</b>	<b>166</b>	<b>395</b>	<b>793</b>	<b>1314</b>	<b>2590</b>	<b>24976</b>	<b>20349</b>
<b>Gender</b>								
Female	5	72	197	392	649	1322	11973	9471
Male	4	94	198	401	664	1268	13020	10889
	<b>9</b>	<b>166</b>	<b>395</b>	<b>793</b>	<b>1313</b>	<b>2590</b>	<b>24993</b>	<b>20360</b>

*Note.* The information on country of birth is based on the fifteen major countries whose applicants registered for the TCF. Missing data for the variables Age groups and Gender were  $n = 32$  and  $n = 5$  respectively. *Source:* Centre International d'Études Pédagogiques.

The increasing number of candidates who have registered for the TCF in the last decade underlines the need for validation research in the context of language assessment for immigration purposes. The number of examinees that have chosen to submit an attestation of language proficiency with their applications for permanent residence does not seem to abate. In fact, this number is growing at a large scale. Validation research needs to address validity issues in terms of the purpose of the test and the population of candidates (e.g., nationality, age, gender, first language) who are taking these tests to comply with immigration requirements or to enhance their chance of selection as potential immigrants.

Regardless of their nationality, age or gender, to settle permanently in the province of Quebec and subsequently become a Canadian citizen, each immigrant needs to go through a selection process or journey that involves various steppingstones until citizenship status is obtained. To some extent, the journey presented in the subsection below represents an attempt, “to preserve a public that will remain strongly intact as long as it is conceived as inherently monoglot” (Hogan-Brun, Mar-Molinero & Stevenson, 2009, p. 11) or “bilingual” in the context of Canada. Saville’s (2009) framework elaborates on the migrant’s journey in the context of Québec’s immigration.

### 1.1.3.5 The immigrants’ journey in the province of Québec

Based on the immigration stages in Great Britain, Saville (2009) proposed a frame of reference to understand the migrant’s journey when settling in a foreign country. He spelled out the bureaucratic processes that are encountered on the way to integration into a new country and

the role that language assessment might play *en route*. Drawing on Saville (2009), Figure 1 below is intended to clarify the stages of the “immigrant journey” from arrival to application for citizenship when seeking to immigrate to Canada through the province of Québec. This immigrant journey provides some points of reference so that greater clarity can be introduced into the discussion about language assessments in the province of Québec and how they relate to other considerations. It is worth noting that the bureaucratic distinctions between migrant categories and subgroups presented in the Quebec’s migrant’s journey do not capture the reality for many individuals who do not easily fit into one or other of the groups, suggesting that more has to be done to understand the needs of individuals with specific requirements (e.g., irregular immigrant whose children were born in Quebec). Moreover, Figure 1 does not suggest that the journey is linear or clearly planned. Many factors can occur to change people’s intentions and influence their choices. Therefore, this journey presents an overview of the major steps that a potential immigrant might go through in the province of Québec before obtaining Canadian citizenship.

The central shaded column illustrates the main transition stages where permissions are usually needed and where regulations must be followed, including providing language tests scores and the payment of fees and ultimately becoming a Canadian citizen. The stages are typical in most destination countries: pre-entry, arrival and entry, extension of stay, settlement indefinitely, application for naturalization (optional), granting of citizenship and issue of new passport. The length of permitted stay is always an important consideration. Some stages cannot be reached without residency requirements being satisfied. For instance, to apply for Canadian naturalization (or citizenship), applicants are not eligible unless they have lived in Canada for a minimum of five years and have been physically present in Canada for 1,095<sup>9</sup> days prior to the submission of the application. Failure to comply with these requirements results in application refusal. In a similar vein, to apply for the QSC, applicants under the Québec experience class need to have obtained an eligible Quebec diploma, or need to have held a skilled job in Québec

---

<sup>9</sup>At the time of writing, applicants may be able to use some of the time spent in Canada as a temporary resident or protected person towards their physical presence calculation. Each day spent physically in Canada as a temporary resident or protected person before becoming a permanent resident within the last 5 years counts as one half day, with a maximum of 365 days, towards physical presence in the country (Government of Canada, 2018).

for at least 12 of the last 24 months. Both requirements expect the candidates to have knowledge of spoken French (MIDI, 2017).

	“Newcomer”		
Study	Request for Entry	a) Québec Acceptance Certificate (QAC) for studies or for temporary work	Keep out! QAC, QSC, or visa denied or
Work		b) Quebec Selection Certificate (QSC) for permanent residence	Welcome! QAC, QSC, or
Family reunion		c) Application for temporary visa (Federal Government of Canada)	visa granted
Refugee/Asylum	ARRIVAL	Granted?	
Foreign students, foreign workers and visitors	Admitted Allow to extend stay	work, study, tourist Apply to extend status as student, worker, or visitor.	
	Stay permanently? A two-stage process: a) Québec selection certificate (CSQ). b) Apply to the Government of Canada for permanent residence. c) Grant unlimited right to remain (permanent resident status)	a) Québec experience program (Québec graduates & temporary workers) b) Regular selection program for skilled workers (Foreign student in Québec & temporary worker) c) Attestation of language proficiency from a designated testing agency (French assessments yield more points towards applications).	Failure to obtain the minimum threshold of points can result in the refusal of the QSC (no deportation unless a removal order has been issue due to a serious crime)
	“Settled immigrant”	One of the following: a) Approved third-party language tests (if 18 – 59 years of age) b) Proof of completion of secondary or postsecondary education in French or English c) Proof of achieving Canadian Language Benchmark/ Niveau de compétence linguistique canadien (CLB/NCLC) level 4 or higher in speaking and listening skills through certain government funded language training programs Plus Citizenship test	Accepted or refused citizenship (no deportation unless a removal order has been issue due to a serious crime)
	“Citizen”	Passport issued	

Figure 1 Immigrant's Journey in the Province of Quebec (adapted from Saville, 2009)

The right-hand side of the diagram summarizes the requirements, rights, responsibilities, immigration programs, and the kind of sanctions that are available if the rules are breached. Of particular relevance are the regulatory mechanisms that have been introduced in Quebec. That is, the QAC and the QSC for study permit and permanent residence respectively. With the increase in regulation comes the possibility of sanctions and this raises the stakes of any procedures (especially the language tests), which are used as an important component to inform decisions. It is with respect to these considerations that many of the ethical issues arise.

Language assessment plays a passive role under the family reunion and asylum seekers applications. Drawing on the concept of “hollow citizenship” (Jamal, 2007; Shohamy & Kanza, 2009), which refers to the right of citizenship to certain ethnic groups exempted to provide proof of language proficiency (e.g., Arabs living in Palestine in 1948 and immigrating to Israel), the province of Quebec may grant *hollow permanent resident status* to applicants who do not speak French or English and yet may still be admitted to the province as permanent residents. Hollow citizenship, in fact, *hollow permanent resident status*, is triggered when immigrants who do not speak French or English are still granted permanent residence status in Québec. Given the various language challenges immigrants might encounter in terms of participation in society, at school and university, family reunion applicants and asylum seekers might struggle at the beginning of their insertion into Quebec society. Knowledge of French or English plays a central role in implicit and covert ways and prevents full participation in civic life (Shohamy, 2006) to those that do not have at least a functional level of proficiency, but to address this issue, there are francization programs for newcomers to the province of Quebec who can enroll to study French.

Van Avermaet (2009) posited that one could also ask the question of what motivates developed countries to have language proficiency requirements and language tests to get permanent residence status or to obtain citizenship. In Quebec, the discourse suggests that requiring immigrants to learn the standard language (i.e., French) of the province and to take a language test as a proof of language proficiency can help promoting the linguistic and social integration of migrants (Pagé, 2011). In addition, Pagé (2011) ascertained that knowledge of French can help immigrants maximize their job opportunities and their access to education to enhance their employability and their integration to the Quebec society. Finally, exploring the immigrants’ journey in the province of Quebec through the lens of Saville’s (2009) immigrant

stages, the different language assessment issues arising at each stage of the journey can be identified. Each of the stages can be useful to promote the collaboration of key stakeholders in the selection process of potential immigrants in the province of Quebec.

## **1.2 Test Fairness in Language Assessment**

Test fairness is an important component of test validation research and is integral for the interpretation and uses of test scores. The concept of fairness can be used broadly to encompass aspects of social and legal contexts (Camilli, 2006), however, in this study the concept of fairness refers to a more narrow definition – statistical or structural analysis of bias, assessed by comparing item or test performance across subgroups (e.g., first language, age, etc.) of interest (Camilli, 2006; Penfield & Camilli, 2007). Issues of test fairness have been in the best interest of the educational and psychological measurement and the assessment community for quite a long time, and the current *Standards* as well as the latest edition of *Educational Measurement* (Brennan, 2006) remind us of such importance. Similarly, language assessment professionals have also been concerned with test fairness and many studies have considered the effect of test-takers' diverse backgrounds on their performance on language tests (see Camilli, 2006; Ferne & Rupp, 2007; Geranpayeh & Kunnan, 2007; Kim, 2001; Ockey, 2007; Roever, 2007). Factors such as gender, first language, age, ethnicity, national origin, race, and religion can affect the validity of test scores in many important ways. Needlessly controversial content and highly sensitive topics related to the aforementioned factors can introduce construct irrelevant variance in the assessment of language proficiency.

There have been efforts to create codes of practice that promote the fairness facet of tests. For instance, in Canada a joint committee was appointed under the Centre for Research in Applied Measurement and Evaluation (CRAME) to create the *Principles of Fair Student Assessment for Education in Canada* to promote fair assessments in education across the country. In addition, the International Language Testing Association (ILTA) is equally concerned with test fairness and have created a code of ethics for language testing and assessment. Perhaps the most comprehensive treatment of test fairness in a code of practice is given in the most recent publication of the *Standards* (AERA, APA, NCME, 2014). The *Standards* call for the consideration of fairness issues across all stages of assessment development, and use:

A test that is fair...reflects the same construct(s) for all test takers, and scores from it have the same meaning for all individuals in the intended population...To the degree possible, characteristics of all individuals in the intended population, including those associated with race, ethnicity, gender, age...linguistic or cultural background, must be considered throughout all stages of the development administration, scoring, interpretation, and use so that barriers to fair assessment can be reduced (AERA, APA, & NCME, 2014, p. 50).

Regardless of the purpose of testing, the goal of test fairness is to maximize, to the extent possible, the opportunity for test takers to demonstrate their standing on the construct(s) the test is intended to assess. If the characteristics of the test itself are not related to the construct being targeted, test results may sometimes yield different meaning for the scores earned by different identifiable groups. This can be problematic and might threaten the validity of the interpretation and uses of scores.

Test content that confounds measurement of the target construct and differentially favors individuals from some subgroups over others is also problematic. According to the *Standards*, differential engagement and motivational value may also be factors in exacerbating construct irrelevant components of content. Material that is likely to be differentially interesting to diverse groups should be balanced to appeal broadly to the full range of the targeted testing population. Also, test content or situations that are offensive or emotionally disturbing to some groups and may impede their ability to engage with the test should not appear in the test. The examination of test fairness is normally conducted at two levels: (1) content experts review the test being used in terms of language, illustrations, graphics and other representations that might hinder the interpretation of the test by various linguistic groups and (2) differential item functioning techniques that might help flagging potential problems of individual test items or entire tests. The former calls for the expertise of content experts and have been coined as *sensitivity review* (Camilli, 2006), the latter requires psychometric approaches to detect differential item functioning (DIF) and potential item or test bias. Although DIF studies do not necessarily conclude that a test is biased, “DIF analyses are important because it is a preliminary step to validate test use, hinting at the potential sources of bias” (Kim, 2001, p. 90). Zumbo (1999, p. 12) distinguished between DIF and item bias, clarifying that “*DIF* occurs when examinees from different groups show different probabilities of success on the item after matching on the underlying ability that the item is intended to measure”, whereas *item bias* occurs when

probabilities of success between groups differ but are attributed to characteristics of the test item that are irrelevant to the test purpose.

In this regard, test developers should strive to maximize test content fairness from the early stages of test development through sensitivity review or fairness review (The *Standards*, AERA, APA & NCME, 2014) and psychometric approaches (e.g., Thissen, Steinberg & Wainer, 1988; Swaminathan & Rogers, 1990; Zumbo, 1999). This is a responsibility that transcends across all testing fields and language assessment is not an exception. For example, since the pioneer work of Chen and Henning (1985), DIF has been a subject of concern in the community of language professionals and has been addressed in various research contexts under various methodological frameworks (e.g., Aryadoust, 2012; Aryadoust, Goh, & Lee, 2011; Engelhard, Kobrin, Wind, 2014; Fidalgo, Alavi & Amirian, 2014; Zumbo, 2003, 2007). In fact, in 2007, the academic journal *Language Assessment Quarterly* dedicated an entire issue to studies that examined test items for differential item functioning.

DIF studies are scanty in language assessment for immigration purposes. This is unfortunate given the risk for test bias in this context. Perhaps one of the few DIF studies is that by Desroches, Crendal, Renaud, Casanova, Demeuse, and Artus (2006) who offered a detailed account of test bias in the context of French as a second language and elaborated on the sensitivity review in place for the Test d'évaluation de français (TEF). However, they did not document very well the methods and the procedures used in the psychometric DIF analyses of the test items. In addition, Desroches *et al.* (2006) limited their discussion to cultural bias neglecting other sources of potential bias such as nationality, age, and gender.

As diversity in immigration waves increases, test fairness becomes of great importance in test development and consequently in validation research. Although this holds true in other contexts such as language assessment for academic purposes, credentials, licenses and countless other situations, test fairness becomes highly sensitive in the context of immigration as test-takers often come from highly diverse backgrounds. Therefore, careful attention must be paid to potential issues of test bias during the test development process and the posterior validation research of language tests. To some extent, it stands to say that the responsible tester and decision maker should document the necessary evidence to ensure that sources of construct irrelevant variance (CIV) and construct underrepresentation (CUR) have been avoided or fully minimized (Messick, 1989). Extraneous, uncontrolled variables that tend to erroneously inflate

or deflate scores for some or all examinees should be closely examined to enhance the meaningfulness and accuracy of test score interpretations, the legitimacy of decisions made based on test scores, and the validity evidence for tests (Downing, 2002). For example, language tests used for immigration purposes should make efforts to avoid inappropriate content and needlessly controversial material and consider the characteristics of potential immigrants (e.g., demography, mother tongue, age groups, etc.) so that language tests are tailored according to the target population. This would avoid introducing irrelevant knowledge or offensive content to the test and can potentially enhance fairness when using a language test in immigration contexts.

#### **1.4 Significance of the Study**

The contributions of this study are twofold: social and scientific. The former refers to how the study will contribute to current practices of selection in the context of Canadian immigration through the province of Quebec. The latter refers to the application of argument-based validation frameworks in language assessment for immigration purposes.

##### **1.4.1 Social relevance**

It is or should be in the best interest of the Quebec government to gather evidence to support the decisions that are being made on behalf of potential immigrants based on language test scores. Although language proficiency is not the only factor to determine selection, it is a key component in the process to obtain the Quebec selection certificate, thus, it is difficult to overstate the importance of validation research in this context. The insights gleaned from test score validation can provide useful information for guiding the use of language tests in the selection process of prospective immigrants and serve to demonstrate to the public that efforts are made to ensure fairness and justice towards the population of applicants.

Furthermore, if candidates are granted access based on weak or poor measures of language proficiency, this can create havoc in the integration of individuals in their new society. Similarly, denying candidates access to the host country based on invalid test scores exacerbates injustice. Therefore, test score validation can contribute to attenuating such problems by providing accurate information regarding the language ability of applicants. In turn, this can feed into more

prolific policies of language assessment for immigration and create better approaches to integrate potential immigrants to the labor market and to the society in general.

### **1.4.2 Scientific significance**

Most validation research in language assessment has been largely conducted on English assessments, especially in the context of academia (e.g., TOEFL, CAEL, IELTS, PTE Academic, MELAB, etc.) and secondly in the context of immigration (e.g., CELPIP-G; IELTS general). There exists limited evidence to help our understanding of French assessments used in the context of immigration (e.g., TCF, DELF, DALF, TEF, etc.). This creates the need for research validation of language constructs across languages, cultures, and settings. This study attempts to contribute to our understanding of second language listening constructs as operationalized in French tests used for immigration as well as the applicability of validity frameworks and research methodologies to develop validity arguments in these contexts.

In terms of analytical tools, confirmatory factor analysis (CFA) has been a widely-used methodology in the validation process of English assessments. However, the application of CFA in the context of French assessment is scanty. At best, CFA has only been applied in translated versions of psychological measures that purport to assess personality or intelligence. This study will use CFA to understand the structure of the TCF and to help validate a L2 listening comprehension model relevant to immigration contexts. In a similar vein, the dichotomous Rasch model will be used in the context of differential item functioning (DIF) to investigate potential test item bias across gender, first language, age and geographical region of target examinees. As a result, the study seeks to shed light onto the Rasch-based DIF methodology framework. Finally, all the test items in the TCF use selected response format to assess listening comprehension. In practice, research validation tends neglect the information carried in the distractors to analyze the appropriateness of test method effects. To address this limitation the nominal response model (Bock, 1972) from the family of item response theory will be utilized to analyze the information in the distractors.

### **1.5 Objectives of the Study**

The objectives of this study are three-fold:

**Objective 1:** To identify the listening subskills that the TCF assesses in order to have a clear idea of the scope of the listening construct operationalized in the TCF.

**Objective 2:** To explore the fairness of the TCF in relation to the targeted test taker population to determine if the TCF reflects the same construct for all test takers and scores have the same meaning for all candidates applying for the CSQ.

**Objective 3:** To assess the efficiency of the test multiple choice questions used in the TCF to elicit L2 listening comprehension data that represents the level of proficiency of candidates.

### 1.5.1 Research questions

Three empirical studies underlie the research questions of interest to this study. These questions relate specifically to study objectives and address validation issues and factors to consider when building a validity argument for the proposed interpretation of test scores for a particular use. The research questions are spelled out as follows:

1. What listening sub-skills do the *Test de connaissance du français* assess? Do test items contribute exclusively to the assessment of the listening construct? Is there supporting evidence indicating that test items are contaminated by construct irrelevant variance or construct underrepresentation?

2. Does the *Test de connaissance du français* exhibit differential item functioning (DIF), leading to test bias against gender, first language, age, and geographical region (North and West Africa only) of examinees, threatening the validity of test score-based interpretations and uses of the test in the context of Quebec immigration?

3. What information do the distractors and the keyed option of selected-response (SR) items (multiple-choice) provide to support or attenuate the validity argument for the use of selected-response items to assess listening comprehension in a second language?

Table 3 below links the research questions to the proposed analytical tools that will be used to build a validity argument for the interpretations and uses of the TCF in the context of Quebec immigration.

The following chapter presents three sections that comprise the theoretical framework that informs this study. The first section provides a perusal account of test validity theory and explores issues of validity in the field of educational measurement. Ultimately, a validation

framework based on the current dominant view of validity theory (e.g., Kane, 2006, 2013) is selected to build a validity argument for the interpretation and uses of test scores obtained on the *Test de connaissance du français* in the context of Quebec immigration. The second section elaborates on second language listening comprehension (e.g., Buck, 2001; Geranpayeh & Taylor, 2013; Rost, 1990) and provides the conceptual basis for the proposed listening comprehension subskills identified in the literature. The third section provides an overview of validity studies in language assessment with a special interest to research that has addressed topics related to the factor structure and fairness of language assessments.

Table III Research phases to build a validity argument for the *Test de connaissance du français*

Topic	Analytical tools	Research question
Structure of the TCF and validation of a L2 listening comprehension model relevant to the context of immigration	Confirmatory factor analysis: specification of predicted variance-covariance matrices to find the best fit of the model to the data. Rasch analysis: item fit statistics	1. What listening sub-skills do the <i>Test de connaissance du français</i> assess? Do test items contribute exclusively to the assessment of the listening construct? Is there supporting evidence indicating that test items are contaminated by construct irrelevant variance or construct underrepresentation?
Fairness of the TCF across cultural groups seeking to immigrate to Canada through the province of Quebec	Rasch Analysis: dimensionality, local independence, item fit statistics, Rasch-based DIF analysis	2. Does the <i>Test de connaissance du français</i> exhibit differential item functioning (DIF), leading to test bias against gender, first language, age, and geographical region (North and West Africa only) of examinees, threatening the validity of test score-based interpretations and uses of the test in the context of Quebec immigration?
The effectiveness of selected response item in assessing L2 listening comprehension and the utility of the NRM model (Bock, 1972) in the context of language data	Nominal response model (Bock, 1972).	3. What information do the distractors and the keyed option of selected-response (SR) items (multiple-choice) provide to support the validity argument for the use of selected-response items to assess listening comprehension in a second language?

## Chapter 2: Theoretical Framework

To present a review of test validity theory, it is of paramount importance to describe prior definitions of the term outlining the evolution and the limitations of these definitions. In so doing, one can ensure that the mistakes that were committed/assumed in the past are not repeated, or at least are minimized, in the present and the future. Therefore, “to set the stage for a detailed discussion of validity in relation to current and prospective testing practice it is important to review traditional approaches to validity” (Messick, 1989, p. 16), and current validity views. The concept of validity has undergone substantial changes in recent years, and it is important to consider not only the evolution of validity theory but also the contexts in which these changes have occurred.

Conceptions of validity have changed or expanded several times since the beginning of the 20<sup>th</sup> century, but one conception, “that validity itself is pre-eminent among the various psychometric concepts, remain constant” (Angoff, 1988, p. 19). In fact, this is the only conception of validity theory that has remained constant since most validity theorists believe that validity is the flagship of educational and psychological measurement. Although definitions of validity have been changed or expanded overtime, these definitions have not always followed the same conceptualizations of previous theorists. Validity as such, has been a very controversial theory in the field of psychological and educational measurement and this makes the dialogue of validity theory highly illuminating and controversial (e.g., Markus & Borsboom, 2013; Mitchell, 2009).

Based on the idea advocated by Jones in 1999 “those who don’t study the past will repeat its errors” (p. 35), we briefly summarize the conceptions of what validity was thought to be from the early 1940’s through the 1970’s. Then, the chapter explores in more detail the concept of validity as conceptualized by Messick in 1989 and finally explores Kane’s (2006) practical and current dominant position proposed in the fourth edition of *Educational Measurement* and refined in Kane (2013). Kane’s (2006) position of test validation stems from the ideas of Messick (1989) and offers a viable model and enticing approach to validate the interpretation and uses of test scores. Kane’s (2006, 2013) model has influenced the work of several scholars in the field of language testing and assessment (e.g., Bachman & Palmer, 2010; Chapelle,

Enright, & Jamieson, 2008) and appears to dominate validation efforts in language testing and assessment (e.g., Aryadoust, 2013; Chi, 2011).

## 2.1 Validity

Given the elusive definitions that have been used to define validity – throughout its history –, it seems adequate to present a brief historical account of the concept to build a model of test validation. In so doing, the chapter looks at the different conceptualizations of validity (Cronbach, 1971; Cronbach & Meehl, 1955; Cureton, 1951; Garrett, 1937; Guilford, 1946; Messick, 1989; Kane, 2006) that have influenced and shaped our current thinking of validity theory and validation. Throughout its history, “validity has often been described as the most important concept in psychometrics, but its meaning is elusive because it has been given so many definitions over the years” (Sireci, 2009, p. 20). The definitions of validity have been spurred not only by the traditional approaches to validity, but also by the different underlying philosophies of science that have been proposed, discredited or otherwise<sup>10</sup>. The lines that follow describe the principal tenets of the different definitions validity has been given since the early 20<sup>th</sup> century.

A review of the early psychometric literature indicates that the earliest definitions of validity were largely atheoretical, defining validity primarily in terms of the correlation of test scores with some criterion. To some extent, it makes sense that validity was initially defined in this manner because the contributions of Pearson<sup>11</sup>, who had only recently published his equation for the correlation coefficient, and Spearman (1904), who had simplified the procedure, made the technique accessible to both psychologists and educationist (Newton & Shaw, 2014; Sireci, 2009). During this time, it was tempting to conceptualize validity in this sense because researchers had an elegant statistical index to relate test scores to other manifestations of the attribute tested. Thus, during the early 1900’s, tests were described as valid for anything with which they correlated. In fact, this view transcended into the middle of the 20<sup>th</sup> century as is evidenced by Guilford’s definition that “a test is valid for anything with which it correlates” (Guilford, 1946, p. 429). This was known as the criterion related-validity era. Lissitz and Samuelsen (2007) explained that those involved in psychological testing labored to quantify the

---

<sup>10</sup> Section 2.3 elaborates a bit on a view of validity which differs from mainstream enthusiasts.

<sup>11</sup> Sireci (2009) stated that Pearson published his formula for the correlation coefficient in 1896.

relationships between test scores and some criterion and that those scores were meant to have predictive power. We can notice that in Guilford (1946), it was common practice to report, for example, that there was a correlation of .60 between a score  $x$  and score  $y$ . During this period, tests, which would be called valid, were regarded as “trustworthy” or as “having diagnostic value” (Lissitz & Samuelsen, 2007, p. 438), hence the emergence of predictive validity or criterion-related validity. In other words, a score on a given test was thought to have predictive power in the sense that it would predict future performance in test related tasks in the real world.

According to Kane (2009), the statistical techniques used to derive criterion-related validity caught the keen eye of the adept quantitative researcher because “once the data were collected, they provided some quantitative answers to some questions about validity, in particular, whether it is reasonable to use test scores to draw inferences about the criterion” (Kane, 2009, p. 44). It is widely acknowledged today that criterion-related validity provides us with some parts of the puzzle in establishing the validity of test scores – if and only if one is interested in predicting future test taker performance. However, criterion-related validity does not provide the stakeholder with all the necessary information that is required to support the interpretation and uses of test scores in a particular context. One of the major limitations in the criterion model of validity is that “one cannot, from a single bivariate correlation (i.e., relationship), deduce the direction of causation” (Markus & Borsboom, 2013, p. 109), rather, one would need several tests scores depending on the same trait to construct a causal model and even then, the model would be perfectible (Markus & Borsboom, 2013).

Empirically speaking, the view of validity was also extended by the advent of factor analysis as advanced by Spearman in 1904. Early psychometrics saw factor analysis as a tool for understanding unobservable traits or constructs and Guilford (1946) was a strong proponent of using factor analysis to define validity and validate tests. His conviction was that factorial conceptions of tests gave us “the most illuminating and useful basis for drawing conclusions regarding the issues in test practice” (Guilford, 1946, p. 429). He went on to classify validity into two categories: practical validity and factorial validity. The former referred to the “correlations between test scores and relevant criteria”, whereas the latter referred to the “factors loadings of the tests on meaningful common reference factors” (Guilford, 1946, p. 428).

In addition to the view of validity based on correlation and factor analysis, a rather theoretical definition of validity was also proposed during the early 20<sup>th</sup> century. According to

Angoff (1988), before 1950, it was generally understood that it was required to demonstrate that a test was useful for a particular purpose. This requirement, known as test validity was defined as “the fidelity with which the test measures what it purports to measure” (Garrett, 1937, p. 266). Albeit credited today as an incomplete definition, “this simplistic view is still seen in textbooks and some psychometric literature because it is an important requirement to support the use of a test for a particular purpose” (Sireci, 2009, p. 22).

Most validity studies of the 1930’s and 1940’s were predictive in nature, but another type of criterion related validity developed as well – concurrent validity. This type of validity also called for the correlation of test against criterion, but the predictor scores and the criterion score were observed at the same point and time. In other words, the difference between criterion-related validity and concurrent validity had to do with when the criterion scores were collected, but both required at least two independent measures to calculate a correlation (Angoff, 1988; Lissitz & Samuelson, 2007). According to Angoff (1988, p. 21), concurrent validity data were taken as evidence that “a newly proposed test was measuring a given trait if it correlated strongly with another test already acknowledged to be a measure of that trait”. In a nutshell, both criterion-related and concurrent validity sought to predict future test taker performance. Moreover, Messick (1989) defined criterion-related validity as the degree of empirical relationship between the test scores and the criterion scores usually in terms of correlations and regressions. He went on to state that in “its pure form, criterion-related validity, is not concerned with any sorts of evidence except specific test-criterion correlations” (Messick, 1989, p. 17).

In the first edition of *Educational Measurement*, Cureton (1951) also defined validity as the correlation of observed scores on the test with true scores (i.e., the score obtained if there were no errors in measurement) on the criterion. As cited in Angoff (1988), Cureton (1951, p. 20) “distinguished the test’s validity from test’s predictive power by defining the latter as the correlation between observed scores on the test with observed scores on the criterion distinguishing both from what he called relevance”, which is the correlation between true scores on both predictor and criterion. Cureton (1951) also advanced two aspects of validity not included in criterion-related validity: relevance and reliability. The former concerned the closeness of agreement between what the test “measured” and the function that it is used to measure (content validity). The latter concerned the accuracy and consistency of the test when used in a specific population (precision).

As noted above, the early conceptions of validity not only developed conjunctively with the emergence of the Pearson correlation, but also with theories that assumed that unobservable psychological attributes existed and could be measured. This promoted the idea that validity was the degree to which a test measures what it is supposed to measure, even though it was limited to a correlation coefficient and at its best on factorial loadings. Moreover, validity was seen as a property of a test and was considered to be high, moderate or low depending on test purpose (See Cureton, 1951). Nevertheless, if we look at criterion-related validity as conceptualized in the early 1900's through the light of current developments of validity theory, one would immediately discard the idea of validity as being a form of correlation coefficient. It has been suggested that various variables could yield high coefficients when correlated with a given criterion measure and many of these variables might be inadequate to the testing context. Furthermore, correlation coefficients only tell us the existing relationship between two variables, thus, one cannot infer causality within a correlational framework. Although the idea of test criterion correlations does not hold today as the cornerstone to test validity, it did provide us with a steppingstone for future developments of test validity theory and it can still provide us with valuable information.

While the initial conception of validity seemed attractive to most scholars in that era, it did not take long for dissatisfaction to settle in and more expanded views to be proposed. To find a remedial solution, what Cureton (1951) had termed test relevance became better known as content validity. In fact, Cureton (1951) introduced the idea of content validity and established the difference between criterion-related validity and content validity. He advocated that:

We may, alternatively, ask those who know the job to list the concepts which constitute job knowledge, and to rate the relative importance of these concepts. Then when the preliminary test is finished, we may ask them to examine the test items. If they agree fairly well that the test items will evoke acts of job knowledge, and that these acts will constitute a representative sample of all such acts, we may be inclined to accept these judgements (Cureton, 1951, p. 664).

Messick (1989, p. 17) perhaps provided the clearest definition of content validity as he defined content validity as “the professional judgments about the relevance of a test content to the content of a particular behavioral domain of interest and about the representativeness with which item or task content sample that domain”. Content validity is still an important feature of

test development and validation today. For instance, in the field of language assessment, it is common practice to have content experts evaluate test items and tasks to establish the representativeness of the domain that the test purports to assess or to have them reverse engineer test items into test blueprints or test specifications. In terms of test development, if a test is used to assess the necessary language ability to enroll in academic studies, content experts are called in to define the construct operationally and then testing experts usually develop the specifications or blueprints that serve to develop test items and tasks.

In terms of validation, for example, when a new test is introduced in an educational testing program, standard setting studies are normally conducted to establish reasonable cut scores that link the new test to performance level descriptors, which are related to the curriculum. This exercise can result in cut scores that determine a pass/fail or determine different levels of proficiency (e.g., basic, intermediate, or advanced). The selection of a panel of experts is instrumental because they are first invited to familiarize themselves with the content of the performance level descriptors and estimate the examinee's probability of success on a given item, while conceptualizing the minimal competence needed to succeed on the item. Content validity does not qualify as the only information needed to support the inferences based on test scores, albeit arguable, it can indeed add to the pieces needed in validation research.

In this respect, we would disagree with Messick's (1989, p.17) assertions about content validity as he suggested that, "since test responses and test scores are not addressed in typical accounts of content validity...content validity does not qualify as validity at all". However, in our view, with the advent of testing for specific purposes in educational settings, although insufficient, content validity might count as validity evidence depending on test purpose. Content validity was an expansion of criterion-related validity that sought to contextualize criterion-related validity because scholars realized that the term as initially conceived was loosely defined. However, due to sampling error of content domains the field of educational and psychological measurement had to seek for other alternatives to better represent what validity entailed and to justify test score interpretation and use. In this regard, on one hand the field of education put more emphasis on and is more concerned with content because most educational assessments are related to curricula or a program of study. On the other hand, psychology is less concerned with content as most psychological tests are not linked or related to curricula (e.g., personality tests, intelligence tests, etc.).

By the middle of the 20<sup>th</sup> century it was realized that both criterion-related and content validity did not provide all the necessary information to conclude that the two categories exhausted the universe of testing situations (Messick, 1989). Cronbach and Meehl (1955) explained the concept of construct validity in detail in their paper published in the *Psychological Bulletin* to encompass under the umbrella of validity test of personality, which were of great importance to the psychologists of this era. Although criterion-related and content validity were the standard approach at the time, the growing dissatisfaction did not go unnoticed. In this regard, Cronbach and Meehl framed their concept of construct validity as an alternative to criterion-related and content validity:

Construct validity is ordinarily studied when the tester has no definite criterion measure of the quality with which he is concerned and must use indirect measures. Here the trait or quality underlying the test is of central importance, rather than either the test behavior or the scores on the criteria (Cronbach & Meehl, 1955, p. 282).

In educational and psychological measurement, *construct* refers to processes or entities that are not directly observed, but is assumed to explain an observable phenomenon, requiring indirect measures to be assessed (Colman, 2006). For example, intelligence and personality (constructs found in psychology) are referred to as constructs, but are only accessed through tests.

One of the major developments in validation research since the work of Cronbach and Meehl (1955) is the increasingly central role taken by construct validity. This new role has incorporated other types of evidence and validation to support the decisions made about test takers based on score interpretation. Drawing on the work of Cronbach and Meehl (1955) one can perceive a clear recognition that validity is not a mathematical property like reliability, but a matter of judgment. Therefore, the year 1955 is a turning point in terms of the conceptualization of test validity because it moves away from merely statistical coefficients to a more comprehensive research program.

While in the early 20<sup>th</sup> century validity was thought to be a property of a test, Cronbach and Meehl (1955) argued that there was no such thing as a valid test, only more or less defensible interpretations: “one does not validate a test, but only a principle for making inferences” (Cronbach and Meehl, 1955, p. 297). This view remained constant in future writing as Cronbach’s (1971) understanding of validity restated somewhat a similar idea; “One validates

not a test but an interpretation of data arising from a specific procedure” (Cronbach, 1971, p. 447). He went on to specify that “because every interpretation has its own degree of validity, one can never reach the compulsion that a particular test is valid” (Cronbach, 1971, p. 447). This position has been maintained in the *Standards* and the work of Kane (2006, 2013), Bachman and Palmer (2010) and Chapelle *et al.* (2008). Besides expanding the concept of validity to a systematic but complex approach to the validation of implicitly defined constructs, Cronbach and Meehl (1955) shifted the focus from the validation of a given test to the validity of the proposed interpretation of test scores, and this idea carried on to inform the work of Messick (1989) and consequently the work of Kane (2006, 2013).

Construct validity as proposed by Cronbach and Meehl (1955) contributed to the development of validity, but this conception also had its flaws. In fact, Cronbach and Meehl (1955) stated that “a necessary condition for a construct to be scientifically admissible is that it occur in a *nomological network*, at least some of whose laws involve observables” (Cronbach & Meehl, 1955, p. 290). A simple definition of nomological network can be conceived as a group of constructs that are theoretically interrelated and these relations are corroborated by evidence (i.e., observables). Under Cronbach and Meehl’s (1955) view, for a measure to have construct validity a nomological network needed to be developed for the measure. In other words, scores elicited by a test needed to be related to other tests’ scores claiming to measure equal constructs. However, as Messick (1989, p.23) explained “nomological networks are viewed as an illuminating way of speaking systematically about the role of constructs in psychological theory and measurement, but not as the only way”.

Cronbach and Meehl (1955) adopted these assumptions because to their knowledge, the validity of the interpretations of test scores was possible only if the evidence (i.e., the data) supported the theory. However, the idea that a test score interpretation was valid if the nomological network was corroborated by the evidence (Borsboom, Cramer, Kievit, Scholten, & Franic, 2009) was discredited for its application was deemed possible to very few specific constructs – those that could be theoretically defined in very precise forms (e.g., weight, length and temperature). Nevertheless, it is worth noting that although it is indeed difficult to establish a nomological network for constructs in the social sciences today, this might change in the future as was the case of theoretical constructs in the natural sciences.

Since a nomological span was and still is difficult to attain in psychological and educational measurement, the idea of the nomological network has spurred a lot of criticism. If the construct validity model was to hold in the domain of psychological and educational measurement, where constructs need not be defined based on a theory, the idea of a set of theoretical rules to be supported by the data resulting from test scores needed to be redefined. This theoretical flaw was acknowledged by Cronbach in 1989 and recently pointed out by Borsboom *et al.* (2009, p. 137), as they put it, “psychology simply had no nomological networks of the sort positivism required in 1955, neither vague nor clear ones, just as it has none today”. Although the general construct-based framework was conceptually rich and inclusive (Kane, 2013), it did not provide a clear guidance for the validation of particular interpretations and uses. This was later recognized by Cronbach (1989, p. 12) as he stated that “it was pretentious to dress up our immature science (construct validity) in the positivist language and that it was self-defeating to say that a construct not part of a nomological network was not scientifically admissible”.

Most of the work by Lee Cronbach prepared the floor for the substantive work of Messick (1989) who extended the view of construct validity in terms of the social implications of testing and the consequences of testing on the society. In other words, Messick’s (1989) view of construct validity considered the social consequences of measurement both the intended and unintended ones. Messick’s (1989) work envisioned construct validity as a concept that embraces almost all forms of validity evidence. In this respect, most scholars (e.g., Lissitz & Samuelsen, 2007) have labelled Messick’s work as being a *unitary conception* of construct validity that has sought to merge the different sources of evidence into a single conception underling the concept.

At this point it is worth noting Markus and Borsboom’s (2013) clarification on this matter. Validity as conceived by Messick is not a unitary conception; rather it is a *unified concept* for other sources of validity evidence are gathered independently and then are merged into a single interpretation of test scores. Thus, from a linguistic standpoint, *unitary* implies one concept whereas *unified* implies a series of factors combined to form one concept. This is the treatment that Messick gave to validity. In Markus and Borsboom’s (2013) words:

Messick formulated a validity theory that was unified, but not unitary. It was unified because, at a theoretical level, it subsumed many lines of evidence under a generalized notion of construct validity. However, the theory was not unitary because it allowed for considerable diversity in which particular lines of evidence or which types of theoretical rationales were drawn together for any given test (Markus & Borsboom, 2013, p. 12).

Messick (1998) later explained the nuance of the unified idea of construct validity as he stated that “what is singular in the unified theory of validity is the kind of validity: all validity is of one kind, namely construct validity (Messick, 1998, p. 37). Thus, Messick would agree that other types of validity whether labeled content validity or criterion-related, or any other validity nickname cannot stand alone as construct validity. Thus far, the only source of evidence not yet explicitly incorporated in the validity literature was the appraisal of social consequences (Messick, 1989) in testing practice. This will be discussed briefly in the lines below.

Eighteen years after Cronbach’s (1971) article in the second edition of *Educational Measurement*, the need for a more comprehensive view of construct validity was in place. In the third edition of *Educational Measurement*, Messick (1989) provided a theoretical definition of validity. Messick (1989, p. 13) advanced that validity was “an integrated evaluative judgement of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment”. If we look at the definition closely we can see that Messick’s (1989) was according less importance to statistical techniques advanced by previously validity models and called for an *evaluative judgement*. This takes away, to a certain extent, the burden and the uncertainty of psychometrics in educational measurement, and offers a definition close to the reality of psychological and educational measurement. That is, the adequacy and appropriateness of the inferences of test scores need to be supported by empirical evidence and this empirical evidence not only comes from psychometrics models but can also come from other qualitative types of evidence (e.g., test takers’ verbal protocols, test sensitivity review, etc.). Messick’s position embraces the nomological framework as offering a useful guide for discipline thinking about the process of validation but the nomological network does not serve as the prescriptive validation model to the exclusion of other approaches.

Although one can infer that Messick’s proposition of validity assumed that a well-developed test had the quality of measuring a given construct, the strict idea of measurement

was attenuated by assuming that one could only observe “signs” of a construct resulting from test performance. In a sense, then, the idea of measurement – if a test is indeed measuring the target construct – was also shifting to less assuming terms such as assessment or evaluation. Messick’s (1989) treatment of construct validity includes a more extensive discussion of the same conceptual and analytic issues presented by Cronbach and Meehl (1955) and Cronbach (1971). Traditional construct validity investigations focused on evidence needed to support (and challenge) the theory underlying test score interpretation (e.g., Cronbach & Meehl, 1955; Cronbach, 1971). Within this framework, the validity models advanced prior to Messick’s work are addressed as being part, of and complementary forms of evidence to be integrated into an overall judgment of construct validity. Messick’s (1989) addition of the consequential basis of test interpretation requires that we also explicitly address the value assumptions implied by the concept labels and theoretical framework selected to guide the validity investigation. For example, the same behavioral indicator has quite different interpretive meaning depending on whether it is labeled as a measure of “flexibility versus rigidity” or “confusion versus consistency” (Messick, 1989, p. 60), each of which implies a different schema for empirical evaluation. In this sense, Messick (1989) advocated that test uses are obviously derived from value positions that are amenable to political debate, as, for example, when meritocratic principles prevail over aristocratic principles as the basis for allocating educational opportunities.

In summary, Messick’s definition of validity focused on demonstrating that a test assesses as much as possible of what it should (minimize construct underrepresentation) and as little as possible of what it should not (minimize construct irrelevant variance) and this should not be divorced from social consequences of testing. Put more simply, on one hand, test developers and responsible stakeholders involved in the validation process of a test should ensure that the construct the test attempts to assess is not underrepresented by the set of items and tasks utilized to generate test scores. For example, using only selected response items tapping language knowledge<sup>12</sup> and claiming that the test assesses communicative competence does not hold according to current views of communicative competence in the field of applied linguistics.

---

<sup>12</sup> Language knowledge here refers to the discrete points of a language system such as grammar, syntax, phonetics, morphology, etc.

Thus, this testing practice would be considered an underrepresentation of the construct of communicative competence. On the other hand, extraneous factors foreign to the target construct should be minimized to the fullest so that the test does not include the testing of other abilities divorced from the testing purpose. For instance, overloading a test taker working memory in a language test attempting to assess listening comprehension would be considered a construct irrelevant variance.

On another note, a very interesting point advanced by Messick (1989) is that value assumptions shape how research questions are framed, which data are gathered, and how results are interpreted. Thus, it is the stressed on value assumptions (Shepard, 1993), construct representation, and construct relevant variance that influence scientific inquiry under the Messick's realm. This should be recognized as a validity framework that distinguishes Messick's (1989) approach from previous traditional validity frameworks. The point is to tacitly put forth these components so they influence the research design. Following this line of thought, Messick later argued that "what needs to be validated are the inferences made about score meaning, namely, the score interpretation and its implications for test use" (Messick, 1998, p. 37). Validity as conceptualized by Messick (1989) has had an impact on the positions or stands scholars adopt regarding validity and has also influenced their thinking overtime, but little to no guidance is provided to apply Messick's ideas in practice.

One important aspect of this definition is the recognition of the context in which testing practice takes place. In this regard, the validation of test scores does not need to present all the sources of evidence unless they are required or dictated by the context where the evaluation or assessment instrument is used (Bertrand & Blais, 2004, p. 238). Moreover, this definition does not leave one with the feeling that "every concern about educational testing is relevant, important, and that every concern should be addressed in testing" (Borsboom, Mellenbergh, & van Heerden, 2004, p. 1061).

Most of the dissatisfaction with Messick's unified theory of validity is based on the notion that his global view of the topic is impractical. In this regard, Lissitz and Samuelsen (2007) explained that stakeholders responsible for test score validation almost always require a concrete and detailed account on how to conduct validation activities. In this sense, the unified notion of validity is not very helpful for the responsible stakeholders. Messick (1989) did not provide a concrete framework on how to proceed to validate the interpretations and uses of test scores.

Rather, he put forth a definition that the researcher and the stakeholder responsible for test score validation would struggle with to produce a conceptual scheme to validate test score interpretations and uses. However, this constituted a steppingstone for the expansion of validity theory and test score validation.

Following traditional cell nomenclature, Messick’s (1989) conceptualization of validity and the ubiquitously cited 2 x 2 matrix (see Table 4) or the four “facets of validity” (Messick, 1989, p. 20) can be summarized as: evidential basis of test interpretation (Cell A), evidential basis of test use (Cell B), consequential basis of test interpretation (Cell C) and consequential basis of test use (Cell D). Considering the four cells in invert order, Cell D refers to “the functional worth of scores in terms of social consequences and their use” (Messick, 1989, p.84); Cell C encompasses “the proposition that value considerations underlie score meaning and are inherent at every stage of test development, administration and reporting” (Cizek, 2012, p. 33). Cell B highlights the importance that should be paid to test use and Messick indicated that this evidence also included construct validity evidence (i.e., interpretation and uses of scores) Finally, Cell A focuses on the evidential basis for test score interpretation and reflects the view that all validity is construct validity. Despite the acceptance of Messick’s writings on validity theory, his position has been criticized for the confusion it can engender. For example, according to Cizek (2012), the cell that refers to the consequential basis of test use, or consequential validity, has been a point of controversy because some scholars considered that social consequences of test use fall outside of the scope of validity (e.g., Hubley & Zumbo, 2011; Mehrens, 1997; Moss, 1998; Popham, 1997). Even though Messick succeeded in convincing the community of educational and psychological measurement that assessment instruments needed to be evaluated through the lens of scientific evaluation into score meaning, his tribulation was to reconcile ethical and scientific evaluation in testing (Newton & Shaw, 2014).  
Table IV Facets of Validity from Messick (1989)

	Test interpretation	Test use
Evidential basis	Construct validity	Construct validity + Relevance /utility
Consequential basis	Value implications	Social consequences

The following section spells out Kane's (2006, 2013) argument-based approach to validation. It is worth noticing that Kane's (2006, 2013) work has been influenced by traditional views of validity theory and test validation. The evolution of validity theory can be seen as an expansion of the theory rather than a new theory that have ignored previous conceptions and theoretical debates.

## **2.2 The Current Dominant View of Validity**

The concept of validity has evolved and has expanded substantially since the early 1900's. At the time of writing this study, the dominant position for test score validation has been advanced in the work of Michael Kane and is an enticing approach that proposes how the interpretation and uses of test scores are to be validated. This section lays out Kane's (2006, 2013) validation framework as the ground work to build a validity argument for the TCF in the context of Quebec immigration.

It is worth noting that the title of the chapter on validity in the fourth edition of *Educational Measurement* does not longer refer to validity per se but to validation. On a linguistic ground this nuance has important implications for the semantics of the concept as the view of validity shifts from a noun (validity) to a gerund or progressive tense "validation". This implies that validity would nowadays be conceived as a process and not as a definition of a term. In this regard, Kane (2006, p. 17) ascertained that "the term validation and to a lesser extent the term validity involves the development of evidence to support the proposed interpretations and uses of test scores; to validate an interpretation or use is to show that is justified". Thus, to gather evidence to support proposed interpretations and uses of test scores implies a process. If we look at this position in more detail, we can observe that this definition has aspects in common with Messick's (1989) work since both conceptualizations suggest that what is to validate is the interpretations and uses derived from test scores and both imply an evaluative judgement. In Kane's (2006) view, this is done by showing that the interpretations are plausible and justified whereas in Messick's words the interpretations should be "adequate" and "appropriate" (Messick, 1989, p. 13).

There are a few differences between Kane's (2006) and Messick's (1989) conceptualizations about validity. The latter did not provide any sort of guidance and his

definition of validity was purely theoretical and difficult to use in applied settings, the former provided a theoretical and pragmatic procedure to validate the interpretations and uses of test scores. Kane's (2006, 2013) argument-based approach to validation provides a clear guidance on how to build defensible arguments to justify test score interpretations and use and echoed Messick's (1989) concerns about the need to justify the inferences and actions based on test scores. Therefore, Kane's work can be seen as an extension of Messick's contribution to validity theory. Another remarkable difference is that in Messick's view, validity is seen as a never-ending process, whereas in an argument-based approach to test score validation, validity has a final stage: when enough evidence has been gathered to support the interpretation and uses of tests. Nevertheless, when potential threats arise and weaken the validity argument more evidence is needed.

The dominant view of validity still considers the notion of construct important, but has downplayed the need to define the construct and certain aspects of the construct model that have transcended to the dominant view of validity (Chapelle, 2012). For example, by focusing on the role of theories (nomological network), Cronbach and Meehl (1955) stressed the need to specify the proposed interpretation before evaluating its validity suggesting that validation should involve an extended research program. In addition, it was also believed that a rigorous research program should "challenge proposed interpretations and should consider competing interpretations (Cronbach and Meehl, 1955). These specifics on validity remain admissible in the current dominant view. However, in the *Standards* there have been shift in terms of terminology: the terms content evidence, criterion evidence and construct evidence are used instead of content validity, criterion validity, and construct validity respectively.

Kane's (2006) approach is pragmatic in the sense that the function of validity is no longer to mirror reality (e.g., to proof the existence of a construct), but to provide a line of reasoning that supports the acceptability of the test interpretation in question. In this view, if the interpretations and uses – supported by the validity argument – made about test takers overcome the critics from rival hypothesis (potential rebuttals) and from the outside public, we are in a safe ground and the claims based on test scores prevail – at least until future evidence challenge such claims. To support the interpretations and use of test scores and the decisions made about individuals based on test performance, Kane (2001, 2006) put forth an argument-based approach to validation. He distinguished five components to build a case for the interpretation and uses

of test scores: scoring of test responses, generalization to a domain of test responses, explanation of performances, extrapolation that go beyond test responses, and decision based on test scores. In the argument-based approach to validation, “test scores are of interest because they are used to support the claims that go beyond (often far beyond) the observed performances” (Kane, 2013, p. 1). This view provides the framework for the evaluation of the claims based on test scores and the core idea is to state the proposed interpretation and use overtly and in some detail, and then evaluate the plausibility of these proposed interpretations (Cronbach, 1988; Kane, 2006, 2009, 2013). In other words, to validate the interpretations and uses of test scores is to evaluate the plausibility of the claims based on the test scores, but to achieve this goal a clear statement of the claims that are proposed is in order, knowing that more ambitious claims require more support than less ambitious claims (Kane, 2006, 2013).

Kane’s (2006) argument-based approach makes use of two kinds of arguments consisting of different components that need to be elaborated to build a case for the justification of the proposed interpretation and use of test scores. These two arguments are the *interpretation/use argument* and the *validity argument*. It is worth noting that Cronbach (1988) coined the term validity argument, but it is still an essential and more elaborated component in Kane’s (2006) work. The interpretation/use argument lays out the proposed interpretations and uses of test scores that originate in the performance of the test takers and end with the conclusions and decisions based on the test taker performance. The interpretation/use argument is presumptive in nature and this quality of presumptiveness always remains inherent in these arguments because they can never be confirmed. The interpretation and uses of test scores may vary from one context to another and to account for this variability it is necessary to be comprehensive in the development of the corresponding interpretation/use arguments. Kane (2006) explained that interpretation/use arguments are informal and their plausibility is subject to empirical challenge, which is evaluated in the light of the soundness of the validity argument.

The validity argument provides an evaluation of the interpretation/use argument and “to claim that a proposed interpretation or use is valid is to claim that the interpretation/use argument is coherent, that its inferences are reasonable, and that its assumptions are plausible” (Kane 2006, p. 23). Thus, the interpretation/use argument is examined rigorously to withstand the challenges and need to be prepared to confront the critics that may rise. A sound interpretation/use argument is one that is supported by a thorough and sound validity argument.

This is similar to a trial in the court of law: the defendant and the plaintiff must provide evidence to support their defense/accusations, and the strongest argument prevails.

Kane’s (2006, 2013) work on argument-based validity used Toulmin’s ([1958], 2003) model of inference to build a validity argument. Under this model, inferences provide a means of making a *claim*, or statement about the examinee’s standing on the targeted construct of assessment on the basis of *grounds* for the claim (e.g., test scores or observations). The inference is supported by *warrants* which can be general principles for making claims. The warrants require *backings* in the form of empirical studies. Inferences can include: scoring, generalization, explanation, extrapolation, and decision inferences. Inferences can be challenged at any time by rebuttals or exceptions, which weaken the strength of the link between claims and its grounds (Chapelle *et al.*, 2008; Kane, 2013). Figure 2 below provides an example of Toulmin’s model of inference.

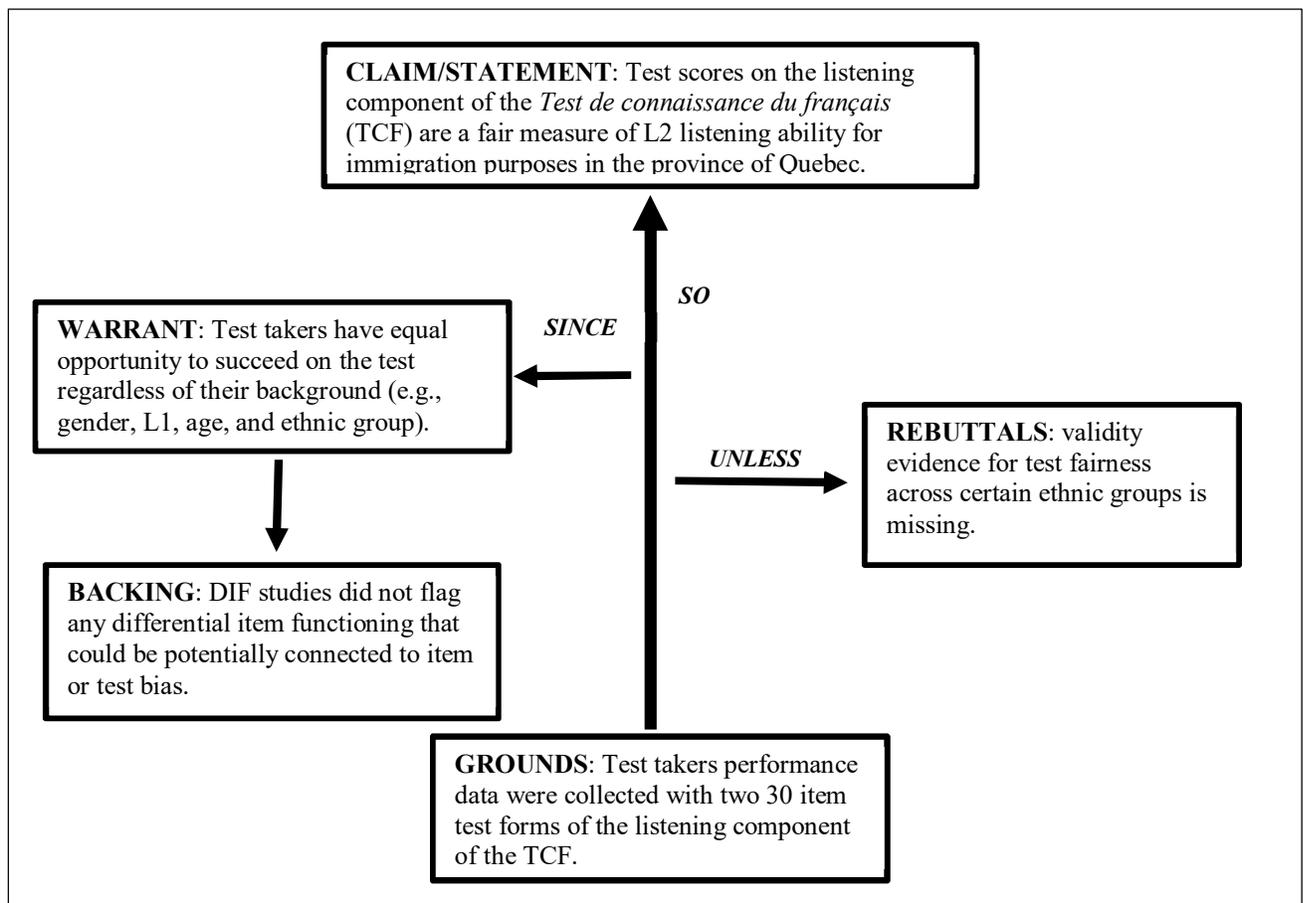


Figure 2 Example structure of an interpretation/use argument for test fairness claim of the TCF listening component in the context of Quebec immigration.

According to Kane (2006), a typical test requires test takers to perform some task and the results are used to draw conclusions and make decisions. This process typically begins by scoring the observed performances and aggregating the tasks' scores, which yields an observed score. This observed score may come from different type of stimuli used in tests. For example, they may be generated by selected-response item types or by constructed-response item types. The scores do not have any meaning standing alone; rather, we want to interpret them as estimates of some more general attribute and as basis for decisions. Although we are not generally interested in a number or score, the scoring inference is important to test score validity because they constitute the foundation for the validity argument. If the instruments utilized to generate the test scores are faulty, we may encounter serious problems in justifying the interpretation and use of tests.

### **2.2.1 Scoring inference**

Among other concerns, the scoring inference requires evidence for the quality of the data collected in a testing program. Data quality can be enhanced by implementing standardized administration procedures and by successful implementation of scoring rubrics and scoring keys. In a validity argument, several kinds of proof may be gathered to justify the quality of the scores. For example, in constructed-response test scoring, empirical data may be used to check on interrater reliability. In selected-response scoring it might be useful to verify that all the items have been marked with the correct answer key. This latter aspect is more accurate today given the fact that most major standardized testing agencies do this type of scoring through automated procedures. Several procedures can be implemented to ensure scoring quality and by conducting distractor analysis of selected response items valuable information can be gathered regarding the functioning of the distractors.

### **2.2.2 Generalization inference**

Whenever we decide to evaluate someone's abilities or competence we do it with a purpose and this normally includes a decision-making step. In education, we normally want to make claims about how well the test taker can perform in a larger domain (Kane, 2013). The generalization component of an interpretation/use argument "relies on the assumptions about

the sampling of the observed performances from the larger universe of possible performances that are of interest” (Kane, 2013, p.10). The generalization inference greatly expands the scope and implications of the interpretation use argument, but this inference is presumptive in the sense that it establishes a presumption in favor of the conclusion regarding the target domain to which the test intent to generalize the test taker performance. However, this is not established definitively and is a serious condition that needs to be evaluated in the validity argument.

In the validity argument, the generalization inference might be evaluated in terms of sampling error and sample representativeness or expert judgment in terms of the representation of test items and tasks of the target domain. However, sampling assumptions are rarely satisfied on both quantitative and qualitative grounds. This notion remains a problematic issue in validity because the generalization inference from the observe score to the universe or domain is uncertain and tests are very limited instruments to account for these criteria. In this regard, Kane (2006, p. 35) posited that “even if we could draw perfectly representative samples estimates of the universe score would still be subject to sampling errors”. With this limitation in mind, the empirical evidence needed to support the backing for a generalization inference can be collected from reliability (Parkers, 2007) or generalizability (Brennan, 2001) studies. However, the estimates of standard errors of measurement derived from generalizability or reliability studies put limits on the precision of estimates of the universe score (Brennan, 2001).

### **2.2.3 Explanation inference**

The explanation inference links the expected score to the targeted construct of the assessment. In the validity argument, backings for the explanation can be gather with factor analytic studies (e.g., confirmatory factor analysis), if the model fit the data and the specified CFA model accounts for most of the variance in the data, the explanation inference is supported. The explanation inference can also be evaluated from a Rasch measurement perspective. That is, misfitting items and erratic patterns of expected responses can be attributed to construct irrelevant variance (Aryadoust, 2013) and be considered as construct contamination. The explanation inference seeks to accumulate evidence on the responses processes hypothesized to underlie the targeted construct. *Explaining* constructs has been in the best interest of validity theorists (e.g., Cronbach, 1971; Cronbach & Meehl, 1955; Messick, 1989) and is an important component of educational assessment because of the link or hypothesized relationship between

tests and curricula. Kane (2006) elaborated on this inference in terms of theory-based interpretation of constructs and Chapelle, Enright and Jamieson (2008) included this inference in the interpretation use argument of the TOEFL iBT to link test scores to the construct of academic English language proficiency.

#### **2.2.4 Extrapolation inference**

Under this component, the interpretive argument seeks to extend presumptively the test score interpretation to predict future performance (e.g., the use of SAT scores to predict college success). According to Kane (2013) in some cases, extrapolations inferences may be very informal and rest mainly on experience. In the validity argument, the extrapolation inference can be evaluated using analytical and empirical evidence. According to Kane (2013), while the former relies on conceptual analyses and on judgments about the relationship between the universe of generalization and target domain, the latter examines the relationships between observed scores and other scores associated with the target domain. This has some similarities with criterion-related validity.

#### **2.2.5 Decision inference**

This final component of Kane's (2006) approach to validation is very straightforward and applies to both the interpretation use and validity arguments. In a nutshell, in almost all cases test scores are used to make decisions about individuals for various reasons and these reasons may include educational or certification purposes. The core idea behind the decision inference is that one needs to provide the necessary evidence to support what we decide about individuals on the grounds of evaluation.

It is important to note here that the above inferences are not to be considered as unique or exhaustive. The definition provided for each of the inferences was to describe in general terms what can be likely included in an interpretation use argument and to briefly elaborate on the components of Kane's (2006) argument-based approach to validity. Moreover, these inferences should not be considered as a checklist, the claims made on behalf of test scores should, to some extent, guide the development of the interpretation use argument. Then, the validity argument should gather the required evidence so that the interpretation use argument holds and stands against potential rebuttals. If this cannot be achieved the interpretation use argument should be

abandoned or modified until it becomes reasonable, plausible and justified by the validity argument (Kane, 2013).

A final remark concerning the current dominant view of validity relates to the attention constructs have received in psychology and in education. Early writings on the topic of validity came from psychologists (e.g., Lee Cronbach, Paul Meehl, and Samuel Messick) and psychology has always shown sparked interest in constructs (e.g., personality, intelligence, happiness, etc.) whereas in education, where nowadays there are more tests, content plays a key role as tests are linked to the intensions and goals of education systems (i.e., curricula and programs of study). Michael Kane attempted to reconcile both fields by putting less emphasis on the construct. And “what forms the basis of the score interpretation is the interpretation use argument rather than a construct” (Chapelle, 2012, p. 19).

### **2.3 Other Views on the Concept of Validity in the 21<sup>st</sup> Century**

Views of test validity have not always converged or agreed with the current dominant view, which have generally followed a linear development through the volumes of *Educational Measurement*, the *Standards* and the contributions of Lee Cronbach, Paul Meehl, Samuel Messick and Michael Kane. Nevertheless, it is interesting and relevant to explore distinct views of test validity as they add value to the ongoing debate and scholarly dialogues on the topic. For instance, in an historical analysis, Markus and Borsboom (2013) identified three interacting processes that have affected test validity theory throughout its history and can help us understand the evolving nature of the concept: *expansion*, *unification*, and *partition*. *Expansion* occurs when researchers or test developers encountered issues not addressed in existing validity theory. For example, content validity be an expansion of criterion related validity since “expansion typically involves new types of validity evidence incorporated into the validation process” (Markus & Borsboom, 2013, p. 5). *Unification* occurs when theorists propose an amalgamation among existing concepts regarding a theory. This process fits well the unified definition of validity suggested by Messick in 1989. Finally, *partition* occurs when researchers advocate for a differentiation of terms treated similarly under existing validity theory. The process of unification has been the basis to critique Messick’s work on construct validity (see Lissitz & Samuelsen, 2007) because some scholars have interpreted his definition as being too general and not offering a roadmap to test validation, undermining the subtleties of a complex concept

such as validity. The addition of content validity to the whole picture of validity mirrors the process of expansion, and the separation between criterion-related and content validity mirrors partition.

If we look at validity from a philosophical standpoint, using the conceptualizations of validity above, “one also finds a successive development of validity in the underlying philosophical assumptions” (Markus & Borsboom, 2013, p. 5). Markus and Borsboom (2013) advocated that the early views of validity theory developed in a manner consistent with a form of descriptive empiricism, reflecting both behaviorism in psychology and positivism in the philosophy of science. In this sense “one finds an emphasis on the idea that claims are not meaningful unless operationalized in terms of observables” (Markus & Borsboom, 2013, p. 7). Following this line of thought, if one is to situate the early views of validity within the philosophical underpinnings of descriptive empiricism, validity, as conceptualized in the early 1900’s favored a theory-neutral approach. This is consistent with the conceptualization of validity put forward by Guilford in 1946 in which he distinguished practical validity with factorial validity. Out of these two categories, the former constitutes a good example of descriptive empiricism due to the lack of a sound theory to bolster the relationship between correlations and what the “test purports to measure” – the criterion. In Messick’s (1989) terms this could be called a value-laden definition for its meaning was based on how well the test scores correlated with the target criterion, and not on a sound theory.

The following section elaborates on views of second language (L2) listening comprehension. The purpose of this section is two integrate key factors in L2 listening comprehension with Kane’s (2006, 2013) validation model. The section starts with a brief historical perspective of L2 listening and expand this discussion to more detailed views of L2 listening including issues, construct definition, and measurement models in L2 listening assessment. The section concludes by recasting the research questions of the study from a validation perspective that integrates both Kane’s (2006, 2013) validation model and key factors relevant to the interpretation use argument of L2 listening assessment in the context of immigration.

## 2.4 Listening Comprehension: Issues and Historical Perspectives

Current conceptions of the listening process maintain that comprehension results from the interaction of numerous sources of information, including the acoustic input and other relevant contextual information. The mind simultaneously processes these incoming stimuli and other information such as linguistic and world knowledge already present in the mind. Listening comprehension is a dynamic process, which continues if new information is made available from many of these sources (Aryadoust, 2013; Buck, 2001; Field, 2013; Ockey, 2013). Listening comprehension is important for in-person communication with estimates that it accounts for more than 45% of the time spent communicating making it the most frequently used language skill in everyday life (Celce-Murcia & Olshtain, 2000; Feyten, 1991). Listening continues to become more important in virtual environments with the increase of communication through technologies such as video broadcasting and Skype (Ockey, 2013). Thus, it stands to say that teaching and assessing the listening skill of immigrants is essential. Unfortunately, the assessment of second language (L2) listening comprehension has attracted little research attention (Buck, 2001), and as a result, understanding of how to best assess it is limited.

Although listening comprehension can be considered as the most neglected language skill least in teaching, research and assessment, there has been a growing interest to learn more about L2 listening ability. In terms of approaches to researching, teaching and assessing L2 listening comprehension, theoretical discussions have proposed different processes of L2 listening comprehension: bottom up, top down, and interactive processes (Buck, 2001; Field, 1999; Geranpayeh & Taylor, 2013; Rost, 2011; Vandergrift, 2004). The bottom up view defines the listening process as local processing levels of information, e.g., acoustic lexical recognition, literal meaning of an utterance and so on. Following a bottom up view, Field (2003) advocated lexical recognition in L2 listening instruction since the attention to speech as a physical phenomenon is an important feature of language to investigate. Thus, it has been considered important to explore how language utterances sound to the non-native listener and to delve into the aspects of spoken language and the obstacles to understanding it at the word-by-word level. The top down view defines the listening process as global processing levels of information where the L2 listener uses metacognitive knowledge, inference skills, background knowledge, and the activation of schema to comprehend aural texts. Moreover, Wenden (1998, p. 574) has

argued that using metacognitive knowledge involving “person, task, and strategy” is of utmost importance for aural comprehension success. Similarly, Field (2003) ascertained that second language learners should be encouraged to bring world knowledge to bear upon their listening experience.

An interactive approach to listening combines both bottom up and top down views as a processing system that occurs simultaneously. According to Buck (2001) this view assumes that listening comprehension is the result of an interaction between several information sources which include, the acoustic input, different types of linguistic knowledge, details of the context, general knowledge and so on. This view argues that listeners use whatever information is available, or what seems relevant to help them interpret what the speaker is saying. Moreover, acoustic lexical recognition and metacognition seem to significantly inform the field of second language listening and provide a theoretical framework that allows researchers to learn more about the interplay among bottom up, top down, and interactive approaches to listening comprehension in a second language (Buck, 2001).

When people listen –whether they are listening to a lecture, a-news broadcast, an announcement, or are engaging in a conversation – they are listening to a stretch of discourse. Although listening might possibly be the most used language skill, it is often overlooked. Weir (2005) argued that L2 listening has been a neglected language skill for much of the thinking on the testing of listening comprehension was based on research in reading comprehension because of the assumption amongst researchers that the comprehension process required in listening shared many of the same routes of the processes in reading comprehension. In this regard, some researchers have referred to L2 listening comprehension as the Cinderella of language teaching and the forgotten skill (Fox, 2004; Vandergrift, 1997a). Lynch (1998) posited that the underlying paradox in listening research is the routine unconscious ease of listening and the extreme difficulty of investigating it, particularly as the process itself is unseen and inaccessible. Lund (1991) noted an interesting paradox about the receptive second language skills of listening and reading. That is, listening has enjoyed a theoretically eminent place in virtually all approaches since audiolingualism (e.g., listening to repeat), but research efforts have been devoted largely to reading. Little attention has been paid to L2 listening assessment until recent research efforts (Aryadoust, 2013; Geranpayeh & Taylor, 2013; Vandergrift & Baker, 2015). In keeping with Lund’s (1991) claim regarding the position of L2 listening in terms of research,

Vandergrift (2004) determined that listening is probably the least explicit of the four language skills, thus making it the most difficult skill to learn [and assess]. The nature of listening, therefore, sets the language skill in a position that attracts the attention of few researchers in applied linguistics.

Although historically listening processes were confounded with reading processes, it is now clear that important differences do occur between the two modalities, for example at the levels of parsing and meaning building, since listeners, unlike readers, must carry forward information in their minds (including information about intonation and stress patterns) without the opportunity to looking back at a text to check comprehension (Weir & Vidaković, 2013). Listening purposes vary according to the situational context in which they occur. This affects the actions and strategies listeners take to understand domain specific spoken discourse. Listening purposes may range from general listening, L2 listening, to academic listening. General listening occurs in our daily routines, it usually starts as unplanned listening experiences that might happen on a bus, at a grocery store, at a bank, etc. L2 listening has a pedagogical focus and can be pre-planned as listening tasks or activities to enhance language learners' listening skills. It may include exercises that require listeners to discriminate minimal pairs for example in French – poisson vs. poison – or summarize the gist of a short story or conversation. This type of listening seeks to draw on language samples of general listening so that the listener is exposed to real world and authentic listening texts.

Moreover, listening purposes vary according to whether learners are involved in listening as a component of social interaction. Brown and Yule (1983) classified listening functions as interactional and transactional. The purpose of interactional listening is to engage in social interaction and its main objective is a two-way communication between two or more interlocutors. This type of listening is regularly encountered in our daily interactions such as going to a bank or calling the doctor. On the other hand, the purpose of transactional listening is mainly one-way communication of information where accurate and coherent comprehension of the message is required for example announcements in subway stations or airports. When involved in transactional listening, L2 listeners need to draw on an array of skills to understand relevant information.

There have been theoretical discussions where second language listening comprehension is conceived in three different ways: (1) the bottom up view, (2) the top down view, and (3) the

interactive view (Buck, 2001; Celce-Murcia & Olshtain, 2000; Field, 2004). The terms bottom up and top down have been borrowed from cognitive psychology (Field, 1999, p. 338), but derive originally from philosophy of science, where they distinguish processes that are inductive from those that are deductive. The bottom up view refers to the use of prior knowledge of language i.e., syntax, lexis, and phonology in decoding the stream of sound. It maintains that knowledge of the phonological system allows the listener to decode acoustic segments as sounds that form words, and words and phrases that form clauses or utterances that are unified by intonation contours having some key prominent element. The top down view refers to the listeners' prior knowledge (also termed content schemata), discourse and socio-cultural knowledge (formal schemata), and assessment of context or speakers' intention. The interactive view refers to the interrelation of the listeners' both prior knowledge of language and knowledge of the world. This assumption is shared among scholars in the field of second language listening who have argued that bottom up and top down approaches to listening are parallel processing systems that interact in decoding the aural input (Buck, 2001; Field, 2008a; Lynch, 2009; Vandergrift, 2004).

Acoustic lexical recognition and metacognition have a close relationship with bottom up and top down cognitive processes respectively. And these two constructs also interrelate in the interactive view of listening. When listeners decode acoustic segments in a linear order, acoustic lexical recognition be a bottom up information-decoding process where listeners attempt to make sense of the in-coming sounds by relating the level of recognition of speech sounds to their knowledge of vocabulary. Conversely, when listeners use their prior knowledge that relates to the listening situation, or resort to using metacognitive knowledge to activate the appropriate background knowledge or strategies to help them make sense of the incoming message, such a process can be identified as a top down approach to listening comprehension. The interactive view reflects an interaction between the acoustic lexical recognition and metacognition. For instance, we can use recognizable acoustic lexical words from an aural text to make inferences that can help us understand the overall message within spoken discourse. This may also happen in reverse where the background knowledge the listener possesses of a text can help the listener understand unknown or unrecognizable lexical words. Due to the importance of acoustic lexical recognition, background knowledge (i.e., content schemata) and metacognition in listening comprehension the sections below elaborates on the different, but harmonizing, views of

listening comprehension and the relationship of metacognition with L2 listening comprehension.

#### **2.4.1 The bottom up view in listening comprehension**

In the 1950's and 1960's experimental psychologists considered listening comprehension to be a bottom up phenomenon (Schramm, 1971). This paradigm of learning and comprehension was mainly based on behaviorist theory that assumed listening comprehension to be a bottom up/linear process of information. The bottom up process was considered as the only cognitive processing model of decoding acoustic input either in listening or reading. In the field of listening, the bottom-up process maintains that human comprehension is driven by the listener's need to process input data accurately. This information-processing model is used to explain how information initially in the form of phonological signals is transformed in the listeners' memory as it undergoes storage and retrieval conversions (Rost, 1994). In listening, the lowest level (i.e., the smallest unit) is the phonetic feature. A simple analysis might present the listener as combining groups of features into phonemes, phonemes into syllables, syllables into words, words into clauses, and clauses into propositions (Field, 1999). This view of listening is rooted in the rationalist tradition in philosophy, the philosophy that gave rise to information processing theory where words were seen as having meaning, and external processes were ignored (Rost, 1990).

However, it is not certain that bottom up processing involves all the levels described above. Some psychologists believe that we process speech into syllables without passing through a phonemic level; others argue that we construct words directly from phonetic features. In this regard, Wilson (2003) points out that the initial sound input is used to match against potential candidate words in the mental lexicon. So, for example, when we hear the initial French sound /rep/, words like "repas", "reprise", "repos" will be activated. These are narrowed down as more sound is processed, so that when we hear /p/ i.e., /rep/ words like "recrue" and "recours" are discarded. This occurs until one match is found, often before the end of the word has been heard. To notice this process, one needs to reflect introspectively pausing the text, or retrospectively with short listening texts, as there is evidence that suggests that listening takes place at a delay of only a quarter of a second behind the speaker (Field, 1999). A quarter of a second is roughly the length of a French syllable, so the listener often begins the processing of

a word before the speaker has finished saying it. Thus, it is worth mentioning that either through introspections or retrospections, reflecting on the listening process is a hard task to do.

Bottom up processing consists of interpreting the stream of sound word by word to build a representation of the discourse (Hansen & Jensen, 1994) and a central tenet of this model is that listening is a sequential process initiated by incoming data (Rost, 1994). Somewhat similar to Wilson's (2003) assumptions described above, Rost (1994) described this process in a four-step sequential order. The listener takes in raw speech and retains a phonological representation of it in working memory, then, tries to organize phonological representation, identifying its content and function. As constituents are identified, the listener uses them to construct underlying propositions, building continually into a hierarchical representation of propositions. Once the propositions for a constituent are identified, the listener retains them in working memory and at some point, purges memory of the phonological representation.

In L2 listening, the bottom up view was strongly manifested with the advent of Audiolingualism in the 1940's. In the Audiolingual method L2 speaking and listening skills were given priority; however, language exercises were very mechanistic and lacked authenticity. The listener had to listen and repeat the utterances verbatim until they master the production of oral discourse close to that of a native speaker. In a sense listening was very discrete and there was no situational context attached to it. Thus, listeners had to process the listening input at the sentence level. Moreover, research was concerned in investigating specific, bottom up aspects of language such as minimal pairs, semantic comprehension of sentences, perception, and decoding of sounds. Due to the complexities of the listening process and the issues it presents to investigate it, the field of applied linguistics did not gain much insight into it and research designs seeking to investigate discrete points of listening embedded in written passages being read aloud. Today we look at the listening process as less rigid than the discrete view of the bottom up approach. This allows the researcher to combine bottom up with top down approaches to gain more insight into L2 listening.

A determining unfavorable factor that affects the bottom up view into L2 listening is the capacity of humans' working memory – the part of short term memory that is concerned with immediate perceptual and linguistic processing (Baddeley, 1986), which provides insights to further understanding the relation between the bottom-up process in listening comprehension and input decoding. As Lynch (1998) noted, there continues to be intense interest in the

interrelationship of memory, listening, and linguistic ability, suggesting that the human being is a limited processor, and one of the constraints we must contend with is working memory. In this respect, Miller (1956) and Baddeley (1986) found that working memory plays a major role during the performance of many basic mental tasks, ranging from serial recall to sentence comprehension, syllogistic reasoning, and arithmetic problem solving. There is empirical evidence that an average person can only store seven, plus or minus two, independent characters in working memory (Baddeley, 1986). Thus, a heavy reliance on bottom up processing skills may overload the capacity of the listener's working memory and may produce communication breakdowns or failures in decoding aural input. Notwithstanding, from an SLA perspective when the aural decoding process occurs automatically i.e., through procedural knowledge, working memory frees itself, constantly allowing comprehension to flow naturally.

#### **2.4.1.1 Acoustic lexical recognition in listening comprehension**

There exists with no doubts an adjacent relationship between acoustic lexical recognition and the bottom up approach to listening comprehension. According to Levelt (1993) for a listener to understand what a speaker says, their acoustic-phonetic processor must first analyze the speech signal and produce a phonetic representation. Although this process is not always straight forward, the steps by which a lexical word is represented in the person's mind is a bottom up process. That is, once the listener has a phonetic representation of a word, the process goes through a stage of phonological decoding and lexical selection of the actual lexicon in the person's cognition. This process ends when a conceptualization of the acoustic input is reached. Because this process normally occurs upwards triggering the recognition of an acoustic signal, such a process is associated with the bottom up view of listening comprehension.

The role of the mental lexicon in human speech comprehension is to mediate between two fundamentally distinct representational and computational domains: the acoustic-phonetic analysis of the incoming speech signal and the syntactic and semantic interpretation of the message being communicated (Marslen-Wilson, 1989). The study of how lexis is processed has a long history in L1 listening research, but has only quite recently attracted wide attention in L2 contexts. Two areas of concern are the nature of the lexicon and how it is accessed, and the fact that there are no consistent gaps between words in connected speech (Field, 2008b).

The acoustic recognition of lexical words is a challenging and demanding task to achieve because aural stretches of discourse can be heavily loaded with lexical words and an attempt to understand each of the words might be cognitively demanding. This holds true not only for the L2 listener but also for the L1 listener. In this regard, Field (2008c) ascertained that the outcomes of linear decoding are tentative for first language listeners, how much more so they must be for those learning a foreign language. Furthermore, Field (2008c) also posited that listeners' incomplete vocabulary repertoire gives rise to a lack of confidence in what they extract from the speech signal. The issue being not so much how many words a listener knows, but how readily the listener can identify known words when they occur in connected speech. Another key issue pointed out by Field (2008c) is the fact that listeners may not be able to fall back upon co-text to support uncertain word recognition the way a native listener can. Thus, appropriate lexical knowledge and recognition of spoken discourse in French play important roles in effective second language listening comprehension.

Lexical representation in the human mind has different advantages and disadvantages. The advantages range from accessible information through visuals, gestures, redundancy, to negotiation of meaning in conversational interactions (Swain, Brooks & Tocalli-Beller, 2002). The disadvantages include phonological modification, i.e., assimilation, elision, and intrusion see Brown (1977). Buck (2001) provided a clear definition of assimilation that refers to sounds that influence the pronunciation of adjacent sounds. For example, in English [won't you] is generally pronounced something like [wonchoo]. In addition, elision occurs when sounds are dropped in rapid speech. For instance, [next day], is usually pronounced like [nexday]. Finally, intrusion occurs when a new sound is introduced between other sounds. For example, in some dialects in England the sound [r] at the end of the word [far] is not normally pronounced, but if the word is immediately followed by a vowel, as in [far away], then it is inserted between the two words. The sounds of a language must be learned to understand speech. These phonological features also exist in French and the L2 French listener needs a good command of them. In this regard, it is not the sounds themselves that cause the most comprehension problems, but the way they vary in normal speech adjacent to each other (Buck, 2001).

In summary, the bottom up view of second language listening assumes that acoustic processing occurs in a linear way and comprehension carries on from phonemes to more complete sounds until full comprehension is achieved. However, the issue here is that unless

this process occurs automatically, an overload of working memory may cause comprehension breakdown. Moreover, vocabulary knowledge does not predict that a listener will be able to identify the matching acoustic signal for each word he or she recognizes in the written register. Issues of ellipsis and liaison and other features of aural discourse add difficulty to this task. The speed in which speech flows also makes the listening task more challenging when the listener attempts to decode the information in a linear way. Thus, the bottom up view comes short when assuming that for successful listening comprehension, a linear decoding process is the only process needed. It is worth noting however that bottom up processing is indeed needed for successful listening comprehension although it needs to be accompanied by top down processing skills. The fact that the listener may start decoding aural messages through acoustic recognition of lexical words, the whole process involved to achieve comprehension is not just bottom up. When listeners identify acoustic signals, they must draw on their world knowledge to have a meaningful acoustic representation of the lexical words. This, in turn, will help them make sense and interpret the spoken discourse.

#### **2.4.2 The top down view in listening comprehension**

In the 1970's and 1980's, the top down view revolutionized the field of second language listening. The theoretical paradigm drew on sociolinguistics and pragmatics to illustrate how the listening process worked. The theoretical assumption of the top down view considers interpretation of an aural text as context driven. And listeners should bear on background knowledge, inference skills, and whatever is available to make sense of the listening situation. Instead of focusing at the sentence level, the top down approach looks at the listening situation at the discourse level.

Top down listening processes involve the activation of schematic knowledge and contextual knowledge both before and during listening in an event that involves spoken discourse. Schematic knowledge is generally thought of as two types of prior knowledge, content schemata that consists of the knowledge of the relevant subject matter, and formal schemata, which consists of knowledge about how discourse is organized with respect to different genres, different topics, or different purposes. For example, in regular daily conversations, listening is viewed as a two-way interactional listening between two or more interlocutors. Contextual knowledge involves an understanding of the specific listening situation

at hand, i.e., listeners assess who the participants are, what the setting is, and what the topic and purpose are. All of this gets filtered through pragmatic knowledge to assist in the processing of aural discourse (Celce-Murcia & Olshtain, 2000).

Top-down listening processes involve the activation of schematic knowledge and contextual knowledge and Buck (2001) referred to this knowledge as non-linguistic knowledge. For example, if we listen to a storyteller, we would expect the stretch of discourse to follow a sequence of events before the peak of the story, and the intriguing and exciting events of the story would normally be towards the middle and the end. Conversely, when we listen to an academic lecture we expect key concepts and relevant information to emerge in an expository manner and a plan of what is going to be covered is generally presented at the beginning of the lecture. Moreover, Buck (2001) argues that background knowledge is very important in listening comprehension. If the listener shares the same knowledge as the speaker, much of what is being said can be understood by means of inferences based on that shared background knowledge.

Top down listening processes are activated when listeners use their understanding of the ongoing discourse or co-text to make meaning of listening input. Pragmatic models tend to be top down in that they posit that comprehension is goal-driven. In this regard, “participants in any interaction pay attention only to the information that seems relevant to their purposes or needs” (Rost, 1994, p. 95). The interaction between linguistic input and top down listening processes “relies heavily on the applications of linguistic processing to background knowledge and context” (Flowerdew, 1994, p. 9). The listeners use what they know of the context of communication to predict how the message will unfold, and use parts of the message to confirm, correct or add to this. Thus, “one of the key processes in top down listening models is inferencing” (Nation & Newton, 2009, p. 40).

Anderson and Lynch (1988) defined top down listening processes as a mental structure consisting of relevant individual knowledge, memory, and experience, which allows listeners to incorporate what they hear into what they know. And Hansen and Jensen (1994) ascertained that top down processing allows the language user to set up expectations about structures, meanings of sentences and the whole text. When involved in this process, listeners create a global macro structure device to interpret subsequent stretches of oral discourse. Furthermore, Field (1999) defined top down processing as more complex than is sometimes suggested and ascertains that contextual information can come from many different sources: from knowledge

of the speaker or from knowledge of the world, from an analogy of a previous situation or from the meaning that has been built up so far in a listening interaction.

Buck's (2001) insights of the listening situation draw our attention to the top down processes of listening. He posited that the situation in which the listening takes place could have a considerable effect on various aspects of the listening process. Firstly, it can determine the topic. For example, a talk in a bread shop (*boulangerie*) is likely to be about bread (*du pain*), and in a chemistry lecture, about chemistry. Similarly, a chat with a friend will require listening to informal language, and a public speech will usually require listening to more formal language. In L2 listening, the top down view may draw on metacognition to explain that while listening to a stretch of discourse, the listeners resort to background knowledge of the topic at hand. The listener examines the situational context and responds according to social and contextual appropriateness if the background knowledge available in the listeners' cognition matches the situation. If there is a gap between the listener and context, the listener might try to compensate by using inferencing skills. The information available to the listener to decode the text is not necessarily available from the text, instead, it is drawn from the listener's schema and co-text clues. A comprehensive (although scanty compared to reading, writing, and speaking) body of research on top-down models of listening can be found in the work of scholars in the field (Field, 2008a; Lund, 1991; Lynch, 1998; Rubin, 1994; Vandergrift, 2004).

The issues underlying top down approaches in listening comprehension are very complex. They involve the listener's background knowledge, which varies from listener to listener. That knowledge can be, in some occasions, fallible and irrelevant to the listening situation. Another problem is the type of genre specific knowledge. For example, a person may be good at decoding messages embedded in conversational French in restaurants settings. However, if the same listener is exposed to a lecture in applied linguistics the background knowledge he possesses in "restaurant talks" would be irrelevant. So, in a sense, background knowledge might help the listener to decode messages that are content and context specific to a domain. In addition, Lynch (2009) discussed that as soon as we examine specific instances of the way people interpret what they have heard, it becomes that we can not necessarily assume that any group of listeners will share common knowledge, so the terms general knowledge and background knowledge raise the questions general to whom? And whose background?

### **2.4.3 Metacognition in L2 listening comprehension**

The concept of metacognition is not identical to background knowledge. Rather, it involves a more conscious strategic way of using – and knowing how to use – one’s background knowledge in the decoding of spoken discourse. Although metacognition can be considered part of a bottom up approach to listening it is more typically associated with top down processes. For instance, one can strategically use the knowledge of phonemes to infer the meaning of a word but the fact that we are thinking at a meta-level of the listening process makes us consider this strategy part of a top down approach to listening. Metacognition is of paramount importance in successful listening comprehension because the listener can use conscious or subconscious strategies to activate the appropriate background knowledge to understand a listening text. Also, it serves to monitor and self-evaluate if one has the apt background knowledge to understand the listening text as it occurs in a given context. For example, after watching a hockey game one might have the ability to understand a conversational English talk about hockey and discuss who played well and so on without being an expert in the game. But if one is not familiar with the positions and the rules of the game, and a more formal discussion takes place using the genre-specific and technicality of the game, our comprehension will not be as good as in the informal conversation. In the latter case our metacognitive knowledge would advise us that we have problems understanding the latter talk because we do not share the same background knowledge as the hockey experts.

Metacognition is a specific focus of cognitive psychology that can further increase our understanding of the complexities of the listening process, refers to an individual’s awareness of personal cognitive performance and the use of awareness to alter that performance (Lundsteen, 1993). The term was coined by Flavell (1979) and involves thinking about one’s own cognitive processes, thinking about one’s thinking, learning, reasoning, and problem solving. This awareness and these beliefs are collectively called metacognitive knowledge, consisting primarily of an understanding or perception of the ways different factors act and interact to affect the course and outcome of cognitive enterprises. Flavell (1979) identified three major categories that comprise the metacognitive knowledge in individuals: person, task, and strategy. Since Flavell’s work, metacognition has informed the field of second language

listening in investigating the relationship and effects of metacognitive awareness between listeners and aural input.

Drawing on Flavell's typology of metacognitive knowledge some scholars in the field of L2 listening have researched the influence of metacognition into listening comprehension to gain more insights into the connection of metacognition and the listening process. According to Goh and Taib (2006) person knowledge consists of general knowledge language learners have about how listening takes place and how different factors like listening contexts and purposes affect comprehension.

**Person knowledge** also includes what learners know about themselves as listeners, and the beliefs they have about what leads to their success or failure in listening comprehension. Goh (1997, 2000) identified four sub-categories of listeners' person knowledge, (1) cognitive processes during listening, where the L2 listener may think of words and spell them out mentally, and reconstruct meaning from keywords that have been successfully understood; (2) Problems during listening in which the listener can mistake one word for another similar-sound one, or cannot remember words/phrases they have just heard or do not recognize sounds of words which are known in writing; (3) Obstacles to listening comprehension, which refers to the limitation of vocabulary, speech delivery rate, types of accent, phonological modification, and (4) obstacles to listening development which refers to one's own personality and social environment.

**Task knowledge** refers to what learners know about the purpose, demands, and nature of listening tasks, this reflects the idea of the top down view that concentrates on formal schemata. It also includes learners' knowledge of the procedures that constitute these tasks. Flavell's illustrations about task knowledge suggest three aspects that can be illustrated in any listening situation. First, learners must have knowledge of the task at hand. Second, they must have knowledge of task demands, and third, they must have knowledge about the information involved in a cognitive enterprise i.e., is it abundant or scarce, familiar or unfamiliar, well or poorly organized. Goh (1997) found that task knowledge in listening comprehension can refer to what the listeners know about the factors that affect listening comprehension such as their existing knowledge, past experiences, and emotional state (e.g., pressure, nervousness, anxiety, fatigue). The usefulness of the input in developing listening abilities is also important to the listeners' task knowledge. This is reflected in the nature of the listening text. For example, if it

is a news broadcast, songs, videos, etc., the aural text will spark different interests in listeners' desire to decode. The notion of tasks posits that learners must be sensitive to when a listening situation will require special effort on their part. In airport announcement, for instance, the listener must discriminate what information should be attended to for future or immediate reference. However, this makes the listening and metacognition very complex because, as Wenden (1998) noted, metacognitive knowledge can be fallible (i.e., what is known is not always empirically supportable and so, may not always be perfectly accurate).

**Strategic knowledge** is the knowing about which strategies are likely to be effective in achieving listening goals. Flavell's description of this variable regarding metacognitive knowledge is brief. He acknowledged that "there is a great deal to be known about the nature and utility of strategies, specifically which strategies can be used effectively in the accomplishment of certain cognitive tasks" (as cited in Wenden, 1987, p. 579). Development of these three aspects of metacognitive knowledge enables L2 listeners to select strategies to improve their listening performance. Based on the factors described above, it can be argued that metacognitive instruction includes both training learners explicitly to employ relevant strategies as well as helping them increase their metacognitive knowledge (Goh & Taib, 2006). Following this line of reasoning Goh (1997) identified two subcategories of strategic knowledge in second language listening: (1) strategies that assist comprehension and recall such as using visual clues, activating knowledge of the contexts from titles, or using existing knowledge to interpret aural discourse, and (2) strategies for developing listening skills such as listening to all kinds of materials, talk to competent speakers frequently, and listening to different variety of accents.

Second language listening comprehension has benefited enormously from research that has drawn on metacognition. It is essential to mention the work of Vandergrift, Goh, Mareschal, and Tafaghodtari (2006) who developed a listening questionnaire designed to assess second language (L2) listeners' metacognitive awareness and perceived used of strategies while listening to aural texts. This questionnaire resulted in a 21-item instrument from which the authors were able to identify five distinct dimensions through factor analysis: (1) problem solving, (2) planning and evaluation, (3) mental translation, (4) person knowledge, and (5) directed attention. These factors are key aspects to consider when researching metacognition in second language listening. However, more research needs to be conducted to replicate these findings in order to draw more convincing conclusions on the effects of metacognitive

knowledge in L2 listening comprehension. The problems with these factors are related to effective pedagogical practices, how can metacognitive knowledge be assessed?

Metacognitive strategies in listening comprehension have played an important role in L2 listening research and by the most part was carried out by Vandergrift and his colleagues in the 1990's. Vandergrift (1997b) suggested that instruction in strategic competence can empower low-level listeners by providing them with useful tools for solving communication problems and enhancing communication. He suggests that an instructional sequence on metacognition could include the following steps: providing students with expressions to clarify meaning and confirming comprehension, developing and presenting training videos where listeners engage in interactive listening, and modeling and practicing the different expressions and strategies in class. Vandergrift (1997a) concluded that because listening in real time is per force a selective process, what listeners decide to select for processing is crucial for comprehension and that the first two years of language learning may be pivotal in acquiring metacognitive strategies to foster successful language learning. Thus, metacognition provides beginning-level listeners with the knowledge and tools for meaningful transfer of learning, so that they know how to listen and understand authentic texts (Vandergrift, 2002).

#### **2.4.4 The interactive view in listening comprehension**

Since the late 1980's to present researchers have realized that listening comprehension is a more interactive process than previously thought (e.g., Buck, 2001; Vandergrift, 2002; Field, 2008a; Lund, 1991). Bottom up and top down listening processes complement one another and although little is known on how this interaction occurs, there is empirical evidence that listeners switch approaches in a non-predictive order (Lynch, 2009). Thus, listening can only work as smoothly as it generally does by being massively interactive and parallel (Buck, 1992).

Research into L2 listening instruction must consider the complex cognitive processes that underlie the listening construct (Vandergrift, 2004). Researchers in L2 listening have reached a common ground in terms of what listening comprehension constitutes for the L2 listener. Our current thinking has moved towards a conception of bottom up and top down views as parallel processes where there is no complete reliance on either of the processing systems. Rather, listening comprehension is viewed as an interactive process in the sense that the various types of knowledge involved in understanding language are not applied in any fixed order. They can

be used in any order or even virtually simultaneously, and they are all capable of interacting and influencing each other (Buck, 2001). The degree to which listeners may use one process more than another will depend on the purpose of the task at hand, and so listening is a contextualized and situated process.

It is also believed that the speed and effectiveness at which listeners carry out both bottom up and top down processes depends on the degree to which the listener can efficiently process what is heard (Vandergrift, 2004). Native language listeners do this automatically, with little conscious attention to individual words. On the other hand, beginning level L2 listeners have limited language knowledge and therefore, little of what they hear can be automatically processed. When L2 listeners encounter comprehension difficulties either communication breaks down or listeners may use compensatory strategies, contextual factors, or any relevant information available to guess at what was not understood (Vandergrift, 2004).

Acoustic lexical recognition and metacognition are factors that may also interact simultaneously. The recognition of aural lexical words, phrases, or whole utterances can activate listeners' background knowledge or metacognition to activate the appropriate background knowledge to understand a text. Also, the listener may draw on previous listening situations to infer the lexical meaning of unknown words. For example, if listener is given preview questions on banking, for instance, before performing a listening task or taking a test, the listener might expect lexical words like mortgage, student loans, savings account, money, credit cards, etc., to have high probability of occurring in that context. It seems then that acoustic lexical recognition and metacognition (in the sense of making inferences) greatly inform listening research and it can be argued that they are cornerstone when conducting research.

In L2 listening authentic recordings of language samples utilized to assess listening comprehension can be conceptualized as an interactive model. The aural text remains the same but the items or tasks are structured in a way that elicit the use of bottom up and top down approaches to listening. For instance, a test item may require students to identify lexical words through gap-filling exercises and at a later stage, students may be asked to pinpoint the gist of the text. Thus, the assessment of listening comprehension calls for an interactive model that includes both bottom up and top down processes.

Rost (1990) offered a general view of the interactive approach to listening comprehension. He claimed that understanding spoken language is essentially an inferential interactive process

based on a perception of cues rather than straightforward matching of sound to meaning. Rost (1990) suggests that the listener must perform five inferential processes while listening: (1) estimating the sense of lexical reference, (2) constructing propositional meaning through supplying case-relational links, (3) assigning a base conceptual meaning in the discourse, (4) assigning underlying links in the discourse and (5) assuming a plausible intention for the speaker's utterances. Thus, the underlying assumption about understanding spoken discourse is much more complex than it seems at a first glance. The interactive view of listening has provided the L2 listening research with an idea of how this process might work, it seems that these processes are not black or white. Rather, the listening process appears to be an interaction of bottom up and top down processes in a continuum, and it is extremely difficult to dissect when one or the other is occurring.

## **2.5 Assessing L2 Listening Comprehension**

### **2.5.1 Construct definition**

Assessment of listening ability has received limited coverage in the language testing literature and is considered the least researched in the literature as well as the least understood in language testing, remaining uniquely undervalued (Brindley, 1998; Buck, 2001; Mendelsohn, 1994). The relatively low profile of listening assessment reflects the inherent difficulties involved in describing and assessing an invisible cognitive operation (Brindley, 1998). In this regard, several L2 listening models have been proposed but very few have been empirically researched or validated to guide testing (Brown & Yule, 1983; Buck, 1991; Dunkel, 1991; Rost, 1990).

Despite the difficulty of assessing second language listening abilities, common points of consensus on the nature of listening processes emerge from the language testing literature, e.g., bottom up, top down, and interactive processes in assessing listeners' abilities. An assumption is frequently made by test developers and that is that there are identifiable listening skills that listeners deploy to comprehend aural texts and that these can be arranged in a hierarchy from lower order, i.e., bottom up assessment (understanding explicit information) to higher order, i.e., top down assessment (understanding implicit information), (see Buck 1990, 1991, 2001; Rost, 1990). The interactive process of listening assessment moves away from the notion of listening as auditory discrimination and decoding decontextualized utterances towards a more complex

and interactive model which reflects the ability to understand authentic discourse in context (Brindley, 1998).

According to Buck (2001) one can define the listening construct in two basic ways. A competence-based listening construct is where it is possible to use a description of listening ability as the basis for defining the listening construct. This is mostly appropriate when we think that consistencies in listening performance are due to the characteristics of the test taker, and that test performance is a sign of an underlying competence that manifests itself across a variety of setting and tasks. According to this model, the listening construct has two parts, language competence and strategic competence. Language competence refers to the procedural and declarative knowledge about the language that the listener brings to the listening situation. Strategic knowledge refers to the cognitive and metacognitive strategies that fulfill the cognitive management function in listening. As noted in Vandergrift *et al.* (2006) the metacognitive strategies include assessing the situation, monitoring, self-evaluating and self-testing. Although strategic knowledge and metacognitive knowledge are important competencies to consider in top down approaches to listening, I would argue from an assessment perspective that a competence based construct definition reflects a bottom up approach because context, background knowledge, etc., are not relevant to assess. Rather, this model considers the listener abilities rather than all the characteristics of real-life listening. That is the relationship between test taker, background knowledge, and situational context.

The alternative way to define the listening construct is as task-based listening construct (Bachman & Palmer, 1996, 2010). This is most appropriate when we think that consistencies in listening performance are due to the characteristics of the context in which the listening takes place, i.e., when we are interested in what test-takers can do. This framework includes characteristics of the setting, characteristics of the rubric, characteristics of the text and the relationship between the text and response. We could claim that this type of construct definition portrays an interactive assessment approach because context, background knowledge, etc., are relevant to assess. The situational context where the listening takes place plays an important role here. For instance, if we were to assess bankers' listening abilities we would define the construct in relation to tasks that need to be performed efficiently in banking situations. This approach to define the listening construct responds much closer to what listeners are to do in a given context.

When we assess listening comprehension, our tests consist of items or tasks where the listeners demonstrate if they have the underlying ability required to answer the item correctly or perform successfully in the task. Tasks and test items also reflect the cognitive processes of listening. If we analyze tasks and test items closely, we could classify them in terms of bottom up items, top down items, and interactive items. The notion of listening tasks in listening tests tends to assess in a more interactive way and with greater scope the situational contexts that may occur in real-life listening.

### **2.5.2 Issues in listening assessment**

One of the issues in assessing listening is that of background knowledge. If we are defining the construct considering background knowledge, this is not only about the aural text but all what the listener brings to bear in the listening test. The more we depend on background knowledge, the more the context makes a difference therefore affecting test performance. A listener might be good in an “X” context or better in another even though his or her language proficiency has not change.

Background knowledge has been defined in many ways by researchers investigating the role it plays in performance-based language testing (Fox, Pychyl & Zumbo, 1997). In the literature, some scholars refer to background knowledge as “scripts” (Schank & Abelson, 1977) and according to Buck (2001) scripts are numerous (e.g., restaurants scripts, birthday party scripts, football scripts, classroom scripts, etc.). In fact, we probably have scripts for all the regularly occurring events in our lives. In listening to a story that calls up a script, the contents of the script automatically become part of the story, even if they are not spelled out. We know what happens when we are waiting in line at a restaurant waiting to be served. When it is our turn, the waiter may only nod at us and we may decode this as “follow me please”. Or, we also know that the waiter will bring the food and drinks to the table and we will pay the bill. Thus, scripts are structures that represent knowledge in memory and are assumed to exist for those things we would want to represent in memory including general concepts, situations, events and our experiences in academic literacy. In L2 listening scripts or schema are activated by both low and high proficiency listeners; Jensen and Hansen (1995) ascertained that low proficiency listeners use schemata. However, they may not select the appropriate schema. Selecting the appropriate schemata depends on having a successful and somewhat automatic interaction

between the input, linguistic knowledge and world knowledge to construct larger units of meaning and to comprehend the discourse.

Domain-specific discourse varies in level of difficulty in accordance with the listeners' knowledge. Background knowledge differs from listener to listener and the successful listener may be the one who shares approximately the same knowledge as the speaker. In this regard, Jensen and Hansen (1995) posited that the relationship between background knowledge and test performance is complex and varies from one test to another, and probably from one topic to another. Nevertheless, in L2 listening tests, Fox *et al.* (1997) found that background knowledge as operationalized as pre-teaching for a test interacts with language proficiency to affect test performance both in terms of overall test scores, as well as how the test taker uses this knowledge while taking the test. In this line of thought Fox *et al.* (1997) also found that the benefit of topic-specific pre-teaching only benefited the high proficiency test-takers whereas low proficiency students did not benefit from topic-specific pre-teaching. This may be the case because through the pre-teaching stage high proficient listeners could activate the appropriate schemata thus making it easier for them to interpret the discourse contained in the test, consequently, awarding them with higher test scores than the control group. This also suggests a language threshold the L2 listener needs to benefit from topic-specific pre-teaching.

### **2.5.3 Measurement models in L2 listening assessments**

The field of second language testing and assessment typically draws on measurement models that help researchers design language tests and validate the scores on these tests (Buck, 1990, 2001; McNamara, 1991). Measurement models are borrowed from the field of educational and psychological measurement to analyze language data. These models can be broadly classified as formative or reflective. A formative measurement model assumes that its indicators (or measures) cause changes in the latent variable or construct. That is, the latent variable is considered a linear composite of its measures. On the other hand, reflective measures are treated as outcomes of constructs. That is, the existence of a construct influences or cause the scores on the indicators (Edwards, 2011). In other words, it is assumed that the construct causes the responses. For example, if we assume that the construct of L2 listening comprehension exists, responses to L2 listening comprehension items in a test are caused by the construct. Reflective

measurement models are generally preferred in language testing. A reflective measurement model of L2 listening comprehension corresponds to the following equation:

$$x_i = \lambda_i \xi + \delta_i,$$

where  $x_i$  is the reflective measure (i.e., the listening test),  $\xi$  is its associated L2 listening construct,  $\lambda_i$  is the effect of  $\xi$  on  $x_i$ , and  $\delta_i$  is the uniqueness of the measure. In other words, a latent variable is introduced (e.g., L2 listening) to account for the covariance between indicators (e.g., test items). According to Schmittmann, Cramer, Waldorp, Epskamp, Kievit, and Borsboom (2013, p. 44) “in most [reflective] models, it is assumed that conditioning on the latent variable makes the covariance vanish (this is an implication of local independence)”. Reflective measures are assumed to represent a single dimension, such that the measures describe the same underlying construct, and each measure is designed to capture the construct in its entirety (Bollen 1984; Bollen & Lennox, 1991; Diamantopoulos & Siguaw, 2006). Because they describe the same dimension, reflective measures are conceptually interchangeable, and removing any one of the measures would not alter the meaning or interpretation of the construct (Bollen & Lennox, 1991; MacKenzie, Podsakoff, & Jarvis, 2005). When designed properly, reflective measures exhibit what DeVellis (1991) calls *useful redundancy*, such that the items have the same meaning without relying on the same terminology or grammatical structure.

As noted earlier in this section, Buck (2001) described two reflective measurement models that are frequently used in the field of language testing and assessment to define language constructs: the competence-based approach and the task-based approach. The former assumes that it is the competence or ability that determines performance, and this exists independent of the context. The latter assumes that it is the context that determines the performance. Buck (2001) ends the discussion suggesting as the best practical approach the use of both competence-based and task-based approaches to define language constructs.

All measurement models are based on some theory or conceptualizations of some kind, and are metaphors that attempt to represent reality. Analyzing test taker performance on listening tests through the lens of a reflective measurement model can be used to learn more about the construct, items, and the characteristics of L2 listening ability. This means that during test construction several theoretical and conceptualization decisions must be made. First the

nature of the construct and the tasks that best reflect the real-life situation and secondly regarding the nature of the data produced and how it can be used for analysis (Buck, 1994).

Psychometric research has produced a powerful array of reflective measurement models and among these we could name a few: Rasch models (1960), where the data must conform to the model, IRT models of Birnbaum (1968) and Samejima (1969), where the models report on the data, common factor models (Jöreskog, 1971; Lawley & Maxwell, 1963), and latent class models (Lazarsfeld, 1959). Reflective measurement models apply complex mathematical procedures to the matrix of item responses to estimate item and person parameters or factors relating to test constructs. Some of the models can be exploratory, confirmatory, and explanatory.

Figure 3 below provides an example of a factor analytic reflective measurement model of general L2 listening comprehension based on subskills identified in relevant research (e.g., Buck, 2001; Ockey, 2013; Rost, 1990; Field, 2013) of L2 listening. This hypothetical L2 general listening model assumes that the construct of L2 listening comprehension causes the responses to the items assessing each of the subskills described in the squares. From left to right, the construct of L2 listening comprehension is divided into two major factors assumed to be correlated: understanding explicit information and understanding implicit information. Each of these factors is subdivided into L2 listening subskills that represent the construct of L2 listening comprehension. It is desirable to assess these subskills in L2 listening assessments to tap into the cognitive processes that have been identified as causing test takers' performance in L2 listening assessment. Failure to assess these subskills and claim that test scores are an indication of L2 listening comprehension may result in construct underrepresentation. Similarly, tapping into skills that are not relevant also contaminate the construct.

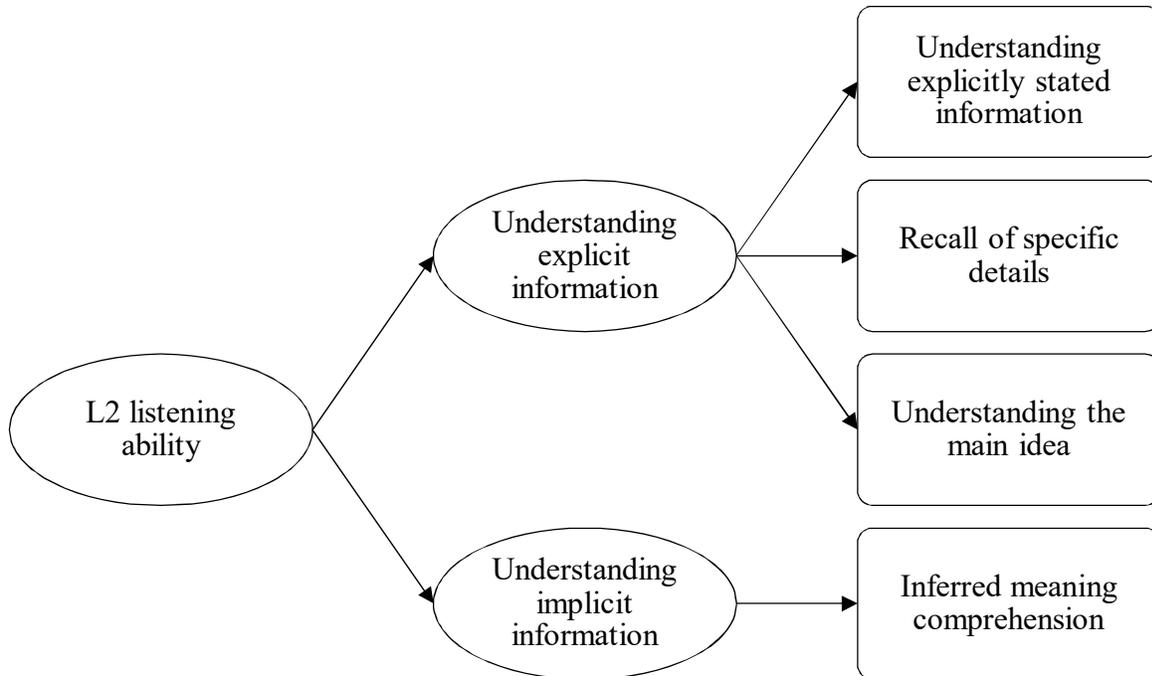


Figure 3 Example of a Reflective Measurement Model in L2 Listening Comprehension

## 2.6 Validity in Language Assessment

### 2.6.1 Theoretical impact on language assessment

This section outlines the impact that validity theory and validation research in educational and psychological measurement have had on language testing and assessment. This review discusses research articles in terms of validation frameworks, and statistical and psychometric analysis used to validate the interpretation and use of test scores of language assessments in different contexts.

The contributions to validity theory stemming from the work of Cureton (1951), Cronbach (1971), Messick (1989), and Kane (2006), among others, have certainly carried an enormous conceptual weight and have set the tone on how validation is understood in language testing and assessment. In this regard, Chappelle (1999) provided a brief historical account of validity and have implemented the argument-based framework advanced by Kane in 2006 (see Chapelle, Jamieson & Enright, 2008). Moreover, the influential work of Messick (1989) on construct validity, but more so the ideas of consequences of testing have been taken very seriously and is salient features in validation dialogues among language testers.

As noted in Chapelle (2012), although the consensus definition of validity has impacted enormously the community of language testing, there are challenges and conundrums to adapt understanding of validity from the measurement literature into the practices in second language research because modern approaches to test score validation are out of reach for a lot of researchers, requiring specialized knowledge (Newton & Shaw, 2014).

Despite the central role of validity and validation research in language assessment, only a handful of English-language assessments have been part of an extensive validation research program (e.g., TOEFL, IELTS). In this vein, Cumming (2013) accentuates the corollary role of validity and validation research in language assessment, outlining the challenges defying key stakeholders when adopting a research agenda in language assessment validation. He further elaborates on the impact of validity theory in language assessment, drawing on the work of Messick (1989) and Kane (2006) to outline validation frameworks in language assessment proposed by Bachman and Palmer (1996) and Bachman and Palmer (2010) respectively. In this regard, the major contribution by Bachman (2005), and Bachman and Palmer (2010) resulted in a comprehensive set of principles and procedures to link test scores and score-based interpretations to test use and the consequences of test use that they termed *Assessment Use Argument*. Since this work originated in the community of language testing and assessment, it is considered by many as an original breakthrough in validation practice. However, these principles were already articulated in the field of educational and psychological measurement. Bachman (2005) and Bachman and Palmer's (2010) credit goes to the integration of these principles rather than their creation. The *assessment use argument* framework has gained popularity in the community of language testing and assessment and several recent studies, including doctoral theses and article published in peer-reviewed journals have endorsed it (e.g., Doe, 2015; Farnsworth, 2013; Jia, 2013; Mann & Marshall, 2010; Pardo-Ballester, 2010; Wang, 2010; Wang, Choi, Schmidgall & Bachman, 2012).

Judiciously, Cumming (2013) emphasizes the relationship between Messick's (1989) work and the targeted "construct" of the assessment and Kane's (2006) work on the rhetoric of arguments for the validity of language assessments. For Cumming (2013), test purpose plays a central role in research validation (e.g., academic, certification, placement, diagnostic, etc.) and influences validation practice. In addition, the article raises concerns about construct definitions in language assessment that are intrinsically related to the adopted theoretical stance on language

constructs. That is, is the language output elicited by a language test due to existence of a trait or construct, a contextual behavior, or an interaction between a person's ability and a context? These are insightful considerations that influence and shape validation research in language assessment and expand on Chapelle's (1999) concerns between the interface of validity theory and language assessment.

### **2.6.2 Validation research in language assessment**

Validation research in language assessment carried between the 1960's and early 2000's adhered to a tripartite model (e.g., content, criterion, construct) of validation, but this can be explained by the prevailing scientific, philosophical, and sociocultural winds of the time. In the past, types of validation evidence such as content, criterion, and construct validities appeared in several titles of research articles published in peer-reviewed journals. For instance, Davies (1984) discussed the validation process of three language tests used for university admission and although the study was about validation, there was no explicit reference to a validation framework. Covertly, both content and criterion evidence appeared to be important to this study as is shown in the abundant use of correlations among the test batteries, which is not surprising given the attention to correlation studies in language assessment in the 1980's. This is somewhat expected since the conception of test validation of the time aligned with correlational studies. However, the study did not report on the assumptions of the statistical analysis involved.

In a similar vein, Dörnyei and Katona (1992) sought to gather validation evidence for the use of c-test to assess general language proficiency. Several criterion measures were collected (e.g., oral proficiency, TOEIC scores) to show the association of c-tests to other measures of language proficiency to showed that c-tests were useful to collect data on general language proficiency. Again, the study did not report on the assumptions of the statistical analysis involved. In terms of validation framework, the term validity occurs several times throughout the article, but no explicit reference is made to a validation framework, implicitly adhering to traditional views of validity, putting concurrent / predictive validity at the heart of the study.

Also, Fulcher (1997) described the implementation of a placement test to reduce the number of students who might have had problems completing their academic degrees because of poor language ability or study skills. Even though the term validity was used in the title and major sections of the article, there is no explicit reference to validity theory. This seems

perplexing since the field of language assessment had already fully embraced a position on validity stemming from Messick's contributions. In terms of methodology, some choices seemed arbitrary. For instance, conceptually, principal component analysis (PCA) assumes that the measured variables (e.g., test items) cause changes in the latent variable or construct implying a formative measurement model. However, changes in indicators of language traits are assumed to be caused by the construct, that is, language proficiency is theorized to predict the score on the question or item (i.e., reflective measurement models). Yet the article draws on PCA to explore the factorial structure of the placement test. Another issue regards the small sample sizes, and the psychometric and statistical analysis that were carried out, potentially highly unstable estimates, threatening the validity of score based decisions.

Unlike the previous studies outlined above, Chapelle (2003) documented very precisely the validation process of a low-stakes web-based ESL test and situated the study within a validation framework. The study drew on notions of validation arguments articulated in Cronbach (1988) and developed a validity argument drawing on Bachman and Palmers' (1996) framework of test usefulness (i.e., reliability, construct validity, authenticity, interactiveness, impact and practicality). This study persuasively builds a context-specific validation framework that fit the purpose of the test and showed a lot of potential. However, some problems with the data hamper the results of the ANOVA and chi-square analyses as the assumption of homoscedasticity was violated for the ANOVA analyses and too many cells had expected frequency of less than five for the chi-square analysis. Although the ANOVA analysis can still be robust if the data are heteroscedastic, if additional violations are present (e.g., asymmetrical distributions in opposite directions) the test may lack robustness.

Although the consensus view of validity theory has impacted enormously the field of language assessment, the challenge to apply validity theory is a daunting endeavor. Guerrero (2000) claimed to have applied Messick's (1989) "framework" to validate the score meaning of a Spanish test for bilingual education teachers but despite addressing important key issues in validation research, the article does not adhere to a validation framework because Messick only outlined concerns about validation research that should be considered when validating test scores. In fact, Borsboom (2016) indicated that Messick (1989) did not offer a guide on how to integrate the evidence relevant to validity to assess the plausibility of the *overall judgement*, leaving many researchers behind in a state of bewilderment. Moreover, the article draws on

global scores of the different components of the test to address validity concerns, but excluded such concerns at the item level, which is a very important step prior to create composite scores. Given the complexity in operationalizing a validation framework that carries Messick's (1989) ideas of validity, over the past decades, many have left Messick's synthesis behind, and have sought to create more light weight validity concepts. The work of Borsboom, Mellenbergh and van Heerden (2004) is a case in point, but also is Kane's (2006) argument-based approach to validity as noted earlier in section 2.1, which in many cases, shuns talk of theoretical constructs and nomological networks. Instead, Kane (2006) stresses the need to produce a strong argument for whatever test score interpretation and use one intends to defend, and presents this as a technique which can be used in tandem with many methodological, conceptual or philosophical inclination (Borsboom, 2016). This innovative approach to validation has been successfully operationalized in validation frameworks highly influential and widely embraced in the language assessment community, and are currently guiding the revision, development and validation of language tests (e.g., Bachman, 2005; Bachman & Palmer 2010; Chapelle, Jamieson & Enright, 2008).

Argument-based approaches to test score validation have followed and have drawn on practical argumentation theories (e.g., Toulmin [1958], 2003), which consist of making claims based on data (i.e., information or facts that serve as the foundation to a claim), concentrating on test score interpretation and use, and the consequences of testing (Bachman, 2005; Kane, 2006; Mislevy, Steinberg, & Almond, 2003). Although the inclusion of the consequences of testing in validation research has been a lingering source of controversy and has spawned considerable debate in the community of educational measurement (Cizek, 2012; Hubley & Zumbo, 2011; Moss, 1998; Popham, 1997), consequences play a pivotal role in validation research in language assessment (Bachman & Palmer, 2010) because life-changing events typically result from high-stakes language assessment (e.g., university admissions, permanent entry to a foreign country, and so on), which requires that potential benefits be weighed against potential unintended negative consequences (Heubert & Hauser, 1999).

Argument-based approaches to test score validation have been widely used in various contexts and areas of language assessment including, but not limited to, the assessment of L2 pragmatics (Youn, 2015), L2 academic listening (Aryadoust, 2013), test fairness (Xi, 2010), automated essay scoring (Chapelle, Cotos & Lee, 2015; Enright & Quinlan, 2010), web-based

language tests (Chapelle, Jamieson & Hegelheimer, 2003; Pardo-Ballester, 2010), test development (Fulcher & Davidson, 2009), L2 literacy (Cheng & Sun, 2015) and scoring rubrics for L2 writing assessment (Knoch & Chapelle, 2017; Lallmamode, Daud & Kassim, 2016). In addition, argument-based approach to test score validation have also been used in published (e.g., Aryadoust, 2013) and unpublished doctoral and master theses (e.g., Chi, 2011; Jun, 2014; Kim, 2010; Le, 2011; Li, 2015; Voss, 2012). This is an indication of how the language assessment community have evolved and have embraced a pragmatic approach to build validity arguments. These studies and theses, however, do not implement a standardized application of argument-based validation and develop different argument structures, providing evidence for some, but not all of the inferences thought to be in a validity argument (this is also the case of the present study). This is a concern expressed by Newton and Shaw (2014, p. 142) who pointed out that “argument-based approach underlies the fact that validation is not simply a one-off-study but a program: potentially a very intensive program”. Despite the discrepancies in its implementation, argument-based validation has proven to be useful in many ways and has overcome the caveats of Messick’s (1989) work in which the researcher was left to decide on the relevance of different kinds of evidence that is needed to inform an evaluative judgement, drawing on validity concepts and definitions that are difficult to disentangle into useful guidelines for validation practice. As noted previously, this was partly the catalyst that spurred alternative thinking to find solutions to the problem and so validation research has been reconceptualized as the process of determining claims and constructing an interpretation/use argument for test scores and in turn a validity argument is constructed to evaluate the plausibility of the interpretation and uses of scores.

Argument-based approaches have also been mapped onto, and used in tandem with, established measurement models. For instance, Aryadoust (2009; 2013) mapped Rasch measurement<sup>13</sup> onto Kane’s (2006) argument-based approach to validity by linking elements of the methodology (e.g., person and item statistics, and fit statistics) to inferences in the argument-based approach. In this regard, Aryadoust (2009) used four inferences (i.e., observation, generalization, explanation, and extrapolation) to illustrate how an inference proceeds from one

---

<sup>13</sup> The classical unidimensional Rasch model will be presented in more detail in section 3.3.1.

validation stage to the next to develop a validity argument within a Rasch-based framework in order to support the claims in the interpretation/use argument. Figure 4 below shows how Kane's (2006) approach to validation can be used with Rasch measurement theory. On the Figure, observation, generalization, explanation and extrapolation inferences are bridges that help proceed from one validation stage to the next. Warrants comprise are the statements that support the postulated inferences and the backings give legitimacy and authority to warrants (e.g., theoretical assumptions and empirical evidence behind the posited warrant).

Aryadoust (2009) posited that warrants for the observation inference in a Rasch-based study can include, for example, the standardization of scoring process, converting raw scores into measured scores and ability. Standardization, in this case, helps to compare item difficulty with person ability or trait levels, since Rasch analysis converts scores to interval like measures making the comparison plausible. To generalize the observed scores into expected scores Aryadoust (2009) proposed person and item reliability, and person and item separation indexes as potential warrants and the theories behind them as backings. The explanation inference bears on the theoretical construct under measurement, where item/person infit and outfit analysis are first warrants and backings include theoretical concepts of fit validity. Investigating item and person fit provides information of construct irrelevant variance.

Lastly, Aryadoust (2009) suggested that we can extrapolate the observation to the target scores. In this regard, differential item functioning (DIF) can be useful. That is, if test takers have similar probabilities to answer an item regardless of their characteristics (e.g., gender, L1, age), this provides further evidence to extrapolate observed scores to the target scores in the population. The extrapolation inference has an element of subjectivity. In this regard, Kane, Crooks and Cohen (1999) indicated that content analysis in the generalization inference can support extrapolation if the universe of generalization corresponds to the target domain. Kane (1992, 2006) also proposed the use of criterion related-evidence to warrant this inference.

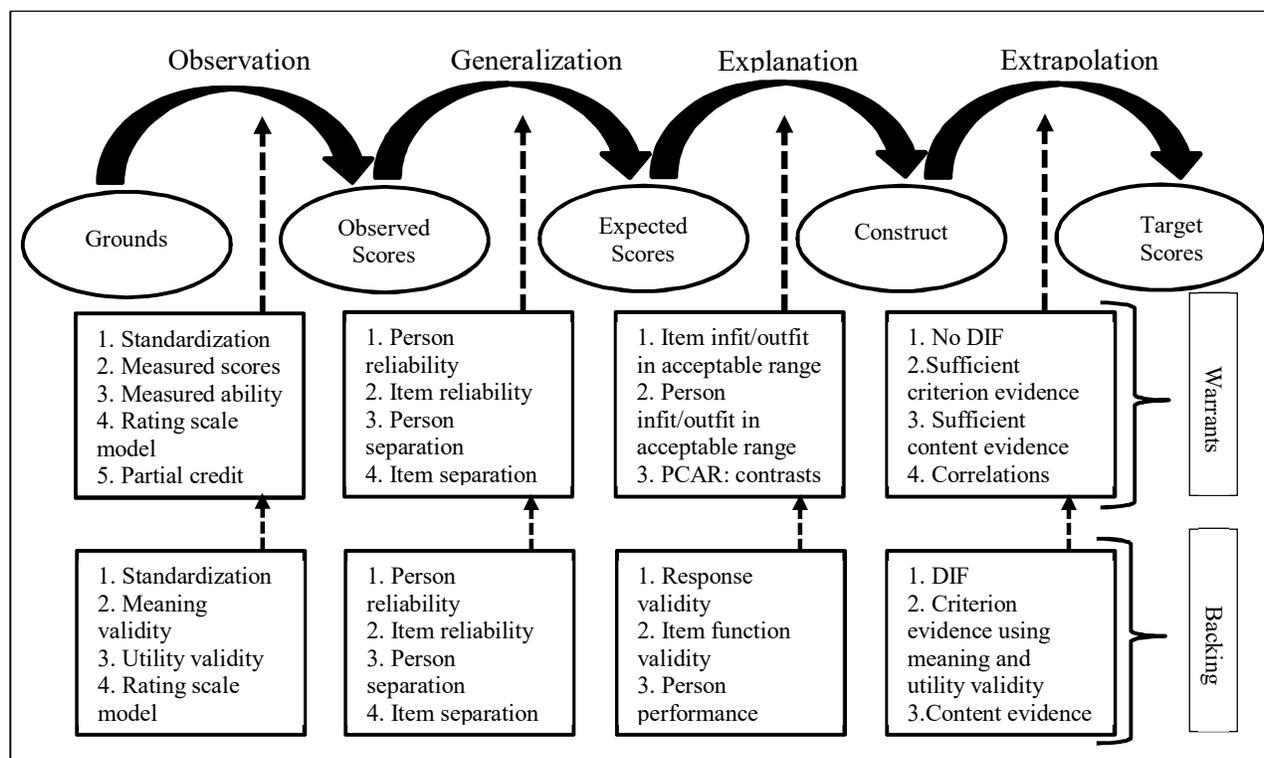


Figure 4 Validity arguments using Rasch measurement (adapted from Aryadoust, 2009)

Although argument-based approaches to test score validation have played a key role in framing validation studies, not all studies have drawn on validity theory or argument-based validation practice to use as a theoretical framework. In fact, a plethora of empirical research, drawing on confirmatory factor analysis and differential item functioning approaches, have been conducted to study the latent structure of tests or to evaluate their fairness. This research has been disseminated and published in peer-reviewed journals and can be directly associated with validation research because the aim has been to gather validity evidence, albeit implicitly so, from instruments that are potentially used to make decisions on behalf of test takers. The subsections that follow review articles that do not adhere to or explicitly adopt a validation framework, but collect the types of evidence that are used as backings to support important warrants in a validity argument.

### 2.6.3 Confirmatory factor analysis in language assessment

The fragmentation of language skills (i.e., listening, speaking, reading and writing) has been researched and discussed in the language testing literature since at least the 1940's, and this has been a central topic of continued interest (Bae & Bachman, 1998). John Ollers' factor

analytic studies conducted in the mid-60's into the *unitary trait hypothesis* of general language proficiency stimulated work in this area in subsequent decades (see the papers in Oller, 1983). Multivariate analyses, such as confirmatory factor analysis<sup>14</sup> (CFA), have played a central role in language related research. CFA is often the analytic tool of choice for developing and refining measurement instruments, assessing construct validity, identifying method effects, and evaluating factor invariance across time and groups (Brown, 2015), has become one of the most commonly used statistical procedures in applied research and is well equipped to address the types of questions language assessment professionals often ask. For instance, CFA helps in answering very popular questions in language assessment such as: what is the factorial structure of the test? Are language skills separable sub-constructs of language ability? In this regard, Bae and Bachman (1998) investigated the factor structure of listening and reading comprehension in a Korean language test. Through CFA, they found that both listening and reading skills could be indeed considered as separable sub-constructs of language proficiency across samples of Korean and non-Korean test takers. And that both sub-constructs correlated highly ( $r = .865$  for the Korean group;  $r = .779$  for the non-Korean group), showing similar factor loading across samples. Although thorough reviews of reporting practices in confirmatory factor analysis research (e.g., Jackson, Gillaspay & Purc-Stephenson, 2009) have shown positive reporting practices to be outweighed by the paucity of reported information in most studies, Bae and Bachman (1998), although dated, provided an exemplary paper in the application of CFA in language assessment. In this regard, they offered a strong rationale for the theoretical and empirical justifications of the models tested: data were assessed for univariate and multivariate normality, the type of matrix analyzed (variance – covariance matrix) was specified, the estimation procedure (maximum likelihood) was discussed, the scale of latent variables was articulated, the software and the version were mentioned, and multiple fit indices (e.g., chi-square,  $df$ ,  $p$ , RMSEA, CFI, TLI) were reported to evaluate the plausibility of the models under investigation.

Various studies in language testing depart from the question of whether language ability is a unitary or divisible construct (e.g., Bae & Bachman, 1998; Sawaki, Stricker & Oranje, 2009)

---

<sup>14</sup> Confirmatory factor analysis is explained in detail in section 3.2.1.

and this interest stems from the work into the unitary trait hypothesis (Oller, 1979) that was called into question by other researchers (e.g., Carroll, 1983; Farhady, 1983). Other confirmatory factor analytic studies have sought to explore the fragmentation of L2 performance assessments in speaking (Kim, 2010; Sawaki, 2007) and writing assessments (Bae & Bachman, 2010), classroom assessment of English proficiency (Llosa, 2007), large scale, high stakes test such as the TOEFL iBT (Sawaki, Stricker & Oranje, 2009) and the TOEIC test (In'nami & Koizumi, 2011), workplace English language proficiency (Yoo & Manna, 2015), partial dictation as a measure of EFL listening proficiency (Cai, 2012) and theoretical models of comprehension skills (Tobia, Ciancaleoni & Bonifacci, 2016). These studies have shed light onto the complexity of language and have accumulated empirical evidence that rejects the most extreme version of the unitary trait hypothesis. That is, a general factor which sufficiently accounts for all the common variances in language tests. Research into the factor structure of language proficiency, as operationalized in language tests, have drawn on theoretical and empirical justification of the models tested. It stands to reason that in general, reporting practices are rigorous (i.e., reporting data screening, parameter estimation, and multiple fit indices). However, these best practices are tainted with the omission of key information that in occasions is missing from the studies. For instance, from the studies mentioned above, most did not specify the type of matrix being analyzed. This is of concern as analyzing a correlation matrix, depending upon the nature of the model and the software used, may lead to incorrect parameter estimates and even fit indices (Cudek, 1989; MacCallum & Austin, 2000). This irregularity in reporting practices is consistent with previous reviews of reporting practices in confirmatory factor analysis in other fields of education (e.g., DiStefano & Hess, 2005; Jackson, Gillaspay & Purc-Stephenson, 2009).

The pervasiveness of confirmatory factor analytic research in language testing can be attributed to the kinds of research questions this analytic tool allows stakeholders to answer. And traditionally, construct validity has been somehow discussed and researched in unison with confirmatory factor analysis. The section below elaborates on another issue akin to validation in language assessment, namely, differential item functioning.

#### 2.6.4 Differential item functioning in language assessment

Similar to studies that have used confirmatory factor analysis as a construct validation tool, not making explicit reference to argument-based approaches to test score validation, differential item functioning<sup>15</sup> (DIF) studies have also played a key role in gathering validity evidence to evaluate the fairness of language test in regard to characteristics of test takers (e.g., gender, first language, nationality, test administration modes, translated tests, etc.). This validity evidence is in turn used to support the interpretation and uses of test scores used to make important decisions on behalf of test takers. An item displays DIF when individuals from different groups do not have the same probability of answering the item correctly after matching on their ability or attribute of interest (Liu, Zumbo, Gustafson, Huang, Kroc & Wu, 2016). Various DIF methods have been introduced to address these kinds of issues (e.g., Angoff, 1993; Dorans & Kulic, 1986; Holland & Thayer, 1988; Liu *et al.*, 2016; Lord, 1980; Swaminathan & Rogers, 1990; Zumbo, 1999) and according to Sireci and Rios (2013) DIF methods can be classified into five categories: (i) descriptive statistical approaches (e.g., conditional proportion correct), (ii) graphical display, (iii) contingency tables, (iv) regression models, and (v) methods based on item response theory (IRT). Camili (2006) differentiated among these methods using only two classifications – those based on observed scores and those based on IRT. The former requires special attention to design (e.g., size of reference and focus groups) whereas the latter requires special attention to both design and data-model fit in the case of the Rasch model and model-data fit in the case of the 2PL and 3PL models. Thus, more assumptions are to be satisfied in Rasch and IRT-based methods for DIF.

In language testing, preoccupations about test fairness can be traced back to the pioneering work of Chen and Henning (1985) who study the systematic bias in a language test used for placement decisions and found that cognates between English and Spanish tended to favor Spanish over Chinese examinees. Subsequent contributions that sought to connect test fairness to validation research in language assessment include Kunnan (1997). DIF investigation has also focused on background languages such as Asian versus European (Kim, 2001), Indo-

---

<sup>15</sup> Rasch-based differential item functioning is explained in detail in section 3.3.1.4

European versus non-Indo-European (Ryan & Bachman, 1992), Arabic and Russian learners of Hebrew (Allalouf & Abranzon, 2008) and Mandarin versus Arabic (Abbott, 2007).

Since the mid-80's language testers have been concerned about the fairness of tests and how test issues flagged in DIF studies can be connected to validation research. Several studies have drawn on different statistical and psychometric methods to address questions about test fairness, using an array of methodologies such as Rasch-based DIF (Aryadoust, 2012, 2013; Aryadoust, Goh & Kim, 2011; Batty, 2014), the likelihood ratio test and logistic regression (Kim, 2001), the likelihood ratio test and Mantel-Haenszel approaches (Pae, 2004, 2012; Ockey, 2007), the standardization method (Harding, 2012; Roever, 2007), the marginal maximum likelihood ratio test (Geranpayeh & Kunnan, 2007), and latent class analysis (Seo, Taherbhai & Frantz, 2016). In regards to the selection of the appropriate method for DIF analysis, the nature of the data to be analyzed will often determine the most appropriate methods where sample size and data type are the two most important factors (Sireci & Rios, 2013). With large sample sizes and dichotomous data, any of the methods used in the research listed above are good candidates. If the data are unidimensional, and effect size criteria are not of interest, then IRT-based methods may be preferred.

One of the major limitations in DIF studies in language assessment is the use of IRT-based or Rasch-based methods (e.g., Aryadoust, 2012; Aryadoust, Goh & Kim, 2011) when sample sizes are not large enough to yield stable item parameter estimation. Sireci and Rios (2013) suggested at least 200 examinees per parameter per group. Similarly, Zumbo (1999) suggested that for binary-scored items in logistic regression-based DIF, 200 examinees per group would be adequate. Another, albeit, different issue in DIF studies is that they can only detect DIF, but cannot disentangle, for example, the effect of gender or first language from other confounders, personal, or contextual factors (Zumbo, 2007). For instance, it is challenging to pinpoint the source of DIF in an item flagged due to gender or first language when there are also existing differences in test takers' learning motivation, social economic status, and knowledge of additional languages. This is more likely the case in educational settings because a lot of confounders covary with outcome variables (Liu *et al.*, 2016). Another issue at a hand arises in disentangling real DIF from artificial DIF (Andrich & Hagquist, 2015), where an item showing large DIF in favor of the reference group (potentially real) induces DIF in other items

(potentially artificial) in favor of the focal group – artificial because it is an artifact of the method for detecting DIF.

### **2.6.5 Distractor analysis of MC items in language assessment**

In validation research in language assessment, classical test theory (CTT) has occupied center stage in distractor analysis. In the case of dichotomous, multiple choice items, CTT calculates the point-biserial correlation between the item in question and total test scores to ensure that those correlations are negative, meaning that high scores on the test are negatively related to choosing any of the distractors. When point-biserial correlation are positive between total score and distractors, items are flagged for inspection, fine-tuning or elimination. The nominal response model (NRM) gives additional information that CTT cannot provide. Persons' thetas (i.e., person ability or standing on the construct) are obtained and the probability of response to all options in a multiple-choice item can be estimated based on persons' thetas. Unfortunately, this model has not been widely used with language data, its use in language assessment is absent and its connection to validation research has not been articulated or linked. Selected response items in the form of multiple-choice are an attractive method of data collection to assess a vast array of traits. This, in part, is due to the objectivity of the method and the relative low cost in item scoring (i.e., they are machine scored in large testing programs). A multiple-choice item is composed of (1) the question or partial sentence called the stem; (2) the choices are the options; the correct answer is often called the "key"; and the incorrect answers are called *distractors*, *distracters*, *foils*, or *misleads* (Haladyna & Rodriguez, 2013).

Traditionally, multiple-choice items have been analyzed in terms of the response patterns of the correct responses or "keys". There exists an array of parametric measurement models (i.e., nominal models) that can be utilized to model the data carried in both the key and distractors of a multiple-choice item, because after all, the distractors are also part of the item (Thissen, Steinberg & Fitzpatrick, 1989). Bock (1972) proposed the nominal response model for the analysis of multiple choice (MC) items. The idea that test takers who answer an MC item incorrectly are unlikely to distribute their responses uniformly over the distractors, motivated the creation of an unordered polytomous measurement model that expressed the probability of response to each alternative of the item as a function of ability.

Research that has fitted the NRM to simulated and real data includes item parameter estimation (De Ayala & Salva-Bolesta, 1999), sample size and parameters recovery (DeMars, 2003), model estimation under nonnormal conditions (Johnson-Preston & Reise, 2014), and scale development (Johnson-Preston *et al.*, 2015). Several other models have been proposed (e.g., Revuelta, 2005; Samejina, 1979; Thissen & Steinberg, 1984; Thissen, Steinberg & Fitzpatrick, 1989) under the label of multiple choice models, and have expanded on the nominal response model to situations where guessing is an issue of concern to the researcher. The nominal response model can be valuable to research in language assessment that seeks to evaluate how the options in an MC item are functioning in relation to level of thetas (i.e., test-taker standing on the construct).

#### **2.6.6 Validation of language tests in the context of immigration**

Although a research program has been recently established for the assessment of English for immigration purposes at Paragon Testing Enterprises in Canada, to my knowledge, the paucity of validation research in language testing in the context of immigration is a prevalent issue. This concern is heightened in the assessment of French for immigration purposes as very little research has sought to develop validity arguments based on current thinking of validation practice and validity theory considering tests in another language. This research not only seeks to contribute to filling this gap, but also to motivate potential readers to pursue validation research in the context of language assessment for immigration purposes.

In Canadian immigration, functional English proficiency (e.g., Paragon Testing Enterprises, British Council, IDP: IELTS Australia and Cambridge English Language Assessment) and functional French proficiency (e.g., Centre International d'Études Pédagogiques, Chambre de Commerce d'Industrie de Paris) are tested by authorized third parties. Test results are used to award points to permanent residence (PR) applications and in Quebec, French is the preferred language of assessment because of its prevalence over English, thus awarding the most points towards PR applications. Validation research on French assessments for immigration purposes is scanty, inaccessible to the public, and absent in peer-reviewed journals. As previously stated, this study seeks to gather validity evidence for the use of the listening comprehension component of the TCF in the context of Quebec immigration. This test is owned by the Centre International d'Études Pédagogiques and authorized by *le*

*Ministère d'Immigration Diversité et Inclusion du Québec* (MIDI) as a French proficiency measure to award points towards PR applications in Quebec.

To conclude, this chapter reviewed the evolution of validity theory in the field of educational and psychological measurement, outlining historical advances that have resulted in the current dominant view of validity (Kane, 2006, 2013) as well as in other views of the 21<sup>st</sup> century (e.g., Markus & Borsboom, 2013). Second language (L2) listening comprehension was also covered, presenting aural comprehension models described as bottom-up and top down, which can be influenced by metacognitive strategies, making the comprehension process interactive in nature. The two last sections explored the construct of second language listening and the subskills that are normally assessed in language tests as well as the impact of validity theory and argument-based approaches to validation on research in the field of language assessment. The table below reorganizes the research questions and statements relevant to the assessment of L2 listening comprehension in the context of Quebec immigration.

Table V Research Questions and Statements for the Interpretation/Use Argument of the TCF L2 Listening Test

Research Questions	Statements
<p>1. What listening sub-skills do the <i>Test de Connaissance du français</i> assess? Do test items contribute exclusively to the assessment of the listening construct? Is there supporting evidence indicating that test items are contaminated by construct irrelevant variance or construct underrepresentation?</p>	<p>1. The <i>Test de Connaissance du français</i> (TCF) assesses listening skills that are relevant to the construct under examination. As such, the listening component of the TCF assesses features of L2 listening such as: understanding of explicit information and implicit information. More specifically, the test is comprised of test items that assess four major sub-skills in L2 listening: (i) comprehending explicit aural discourse, (ii) recalling specific details, (iii) demonstrating understanding of the general topic or main idea, and (iv) inferring implicit ideas.</p>
<p>2. Does the <i>Test de Connaissance du français</i> exhibit differential item functioning (DIF), leading to test bias against gender, first language, age, and geographical location of examinees, threatening the validity of test score-based interpretations and uses of the test in the context of Quebec immigration?</p>	<p>2. The <i>Test de Connaissance du français</i> is a fair measure of L2 listening proficiency. As such, the test does not favor group membership regardless of gender, first language, age, or geographical location (only North and West Africa are included) of candidates who submit test scores to immigrate to Canada through the province of Quebec.</p>
<p>3. What information do the distractors and the keyed option of selected-response (SR) items (multiple-choice) provide to support or attenuate the validity argument for the use of selected-response items to assess listening comprehension in a second language?</p>	<p>3. Selected response items in the form of multiple-choice (MC) are a useful method to gather evidence of L2 listening comprehension. The keyed responses and distractors provide empirical evidence that supports the use of MC items to assess L2 listening comprehension.</p>

## Chapter 3: Methodology

The focus of this study is to gather evidence of the validity of the interpretation of the *Test de connaissance* (TCF) listening test scores to justify the use of the test for immigration purposes in Quebec. In Chapter one, the research questions were presented in a traditional way, in Chapter two the research questions were recast in terms of statements that can be used in an interpretation use argument. This Chapter describes the TCF listening test, participants, methods, procedure and analyses that were used to build the validity argument for the TCF in the context of Quebec immigration.

### 3.1 Test de connaissance du français

The TCF is a French proficiency test that assesses four language skills: listening, speaking, reading, and writing. Speaking and reading are optional examinations unless required by the test user. For this study, only the listening component is investigated and described here. The listening component consists of 30 items that are machine-scored and all use the standard four option multiple-choice format (the key and three distractors). According to the CIEP, the TCF has been adapted for Quebec and this adaptation is acknowledged in their website:

The TCF for Quebec is for anyone regardless of nationality or native language who wishes to begin permanent immigration procedures with the Quebec Ministry of Immigration, Diversity and Inclusion. The TCF for Quebec has been designed to meet the demands of Quebec authorities, which make taking the tests and standardized French examinations obligatory and systematic within the context of procedures for obtaining the Quebec Selection Certificate, leading to the issuance of a permanent residence visa.

The listening component of the TCF is comprised of four major sections:

**(a) Picture recognition** based on the comprehension of four short aural statements (e.g., short interaction between two interlocutors or short monologues) where test takers are asked look at an image, a photo or a drawing and have to select the statement that best describes the image, photo or drawing. These language instances reflect captures of daily life (e.g., walking on the street, park scenes, restaurants, households, etc.).

**(b) Mini-dialogues** where test takers need to comprehend short exchanges (generally between two interlocutors) related to real-life situations (greetings, propositions, phone calls, request for information, store announcements, weather, etc.) and after listening to aural input

(e.g., the stem is uttered by interlocutor A) they need to identify the logical or correct response in the options (e.g., options are uttered by interlocutor B).

**(c) Short conversations** based on short but complete monologues (e.g., announcements) or complete conversations (face-to-face or by telephone) between two interlocutors (up to six exchanges) where test takers are asked a comprehension question based on what they have just listened.

**(d) Longer passages** based on itineraries (e.g., excursions), reports, interviews, long conversations, or radio broadcasts (monologues or between interlocutors). This type of items may exhibit white noise. Test takers are asked to comprehension questions based on what they have heard.

### **3.1.1 Test administration procedures**

The paper-based version of the listening component of the TCF is administered with an audio CD-player and prior to the test date proctors are required to ensure that the audio works properly and that the room acoustics do not interfere with the clarity of the recording. On the test date, ID cards containing the test takers' names are placed on assigned seats and before the test proctors ensure that the audio recording still works and the room lighting is adequate.

Test takers are required to bring a photo ID for identification purposes, are asked to power off electronic devices and leave personal items in a designated place in the room. Then, proctors guide test takers to their assigned seat and when all candidates have been seated, the proctor gives a welcome announcement and details about the listening test (e.g., duration, format, potential sanctions due to fraud and the testing instructions). After the general announcement, booklets and answer sheets are distributed making sure that they match the candidates' name on the ID card previously placed on their seats. The proctor explains how to use the answer sheet and asks if the candidates have questions regarding the procedure. Finally, the proctor writes on the chalkboard the test start time, end time, and the time allotted to complete it (25 minutes). Test takers are asked not to leave the room before the test ends. After all aspects of the test procedures have been covered, the proctors starts the recording as well as the 25 minutes countdown.

### 3.2 Participants

The data for this study stemmed from two active, operational test forms (Form A and Form B<sup>16</sup>) of the listening component of the TCF, consisting of relatively large sample sizes ( $n = 1,501$  and  $n = 2,118$  respectively). Each of the participants took the test to provide evidence of listening proficiency to immigrate to Canada through the province of Quebec. Table 6 below provides the sample size for each of the major nationalities represented in both test forms.

Table VI Nationalities of Test Takers of the TCF for Immigration to Quebec

Nationality	Form A		Form B				
	<i>N</i>	Nationality	<i>n</i>	Nationality	<i>n</i>	Nationality	<i>n</i>
Ivorian	209	Chinese	89	Colombian	287	Egyptian	153
Colombian	159	Israeli	72	Ivorian	286	Chinese	96
Cameroonian	154	Senegalese	68	Iranian	220	Brazilian	94
French	149	Algerian	49	Moroccan	165	Senegalese	79
Moroccan	115	Bulgarian	49	French	163	Tunisian	66
Tunisian	102	Lebanese	46	Cameroonian	156	Algerian	45

*Note.* Other nationalities not included on this table due to small sample size ( $n < 45$ ) include: Russian, Iranian, Indian, Congolese, Egyptian, Togolese, Italian, Syrian, Ukrainian, Spanish, Mexican, Bolivian, and many more. A complete list can be found in Appendix A.

The examinees can be considered as a representative sample of candidates who submitted test scores for immigration to Quebec. The data also included gender information (females = 660 and males = 841 for Form A. And females = 970 and males = 1,148 for Form B). Detailed information about first language and age groups is provided in Table 7 below.

Table VII First Language and Age Groups of Test Takers Taking the TCF for Immigration to Quebec

L1	Form A			Form B			
	<i>n</i>	ages	<i>n</i>	L1	<i>n</i>	ages	<i>n</i>
French	546	18-29	580	French	678	18-29	885
Arabic	281	30-35	571	Arabic	427	30-35	809
Spanish	184	36 +	344	Spanish	322	36 +	423
Russian <sup>17</sup>	104	total	1495	Persian	215	total	2117
Chinese	87	missing	6	Portuguese	97	missing	1
Bulgarian	49	total	1501	Chinese	96	total	2118

*Note.* L1 = first language; other languages were not included on this table due to small sample size ( $n < 49$ ) include: English, Wolof, Italian, Portuguese, Gujarati, Kabyle, and many others. A complete list can be found in Appendix A. Age groups were clustered following Quebec's age grouping used in immigration reports and include all the data from both test forms.

<sup>16</sup> Form A and Form B are used to represent each of the of the test forms in the study.

<sup>17</sup> This language group is composed of descendants or immigrants of the Russian-speaking communities of the Soviet Union who are Israeli ( $n = 68$ ), plus nationals of Russia ( $n = 29$ ), Ukraine ( $n = 3$ ), Belarus ( $n = 2$ ), Argentina ( $n = 1$ ), and Serbia and Montenegro ( $n = 1$ ), whose first language is Russian. This composite explains the absence of Russians on the nationality table.

### 3.2 Panel of experts

Language testing experts and teachers of French as a second language ( $n = 12$ ) were recruited to be part of a panel of experts (hereafter, the panel). The panel consisted of members ( $n = 5$ ) who had already earned a doctoral degree in language pedagogy, language testing, second language acquisition and learning, education, or literacy, of doctoral candidates in educational measurement working with language data in their dissertations ( $n = 2$ ) and language pedagogy concentrating on French pedagogy ( $n = 1$ ), of members ( $n = 3$ ) who had already earned a Master's degree in *Lettres françaises* (French literature), bilingualism studies, and psychology, and an MA student ( $n = 1$ ) in bilingualism studies. The panel had substantive experience (ranging from 1 to 30 years) in teaching, researching and assessing French as a second language. Research experience included: validation of French and English assessments, rater effects in the assessment of speaking in French, and cognitive measurement models [CDM's] applied to reading comprehension data. Assessment experience included: classroom assessment, test development, and examiners of high-stakes French tests (e.g., Diplôme d'études en langue française [DELF], Test d'évaluation de français adapté au Québec [TEFAQ], and tests developed at the School of Public Service of Canada).

The panel was asked to associate each test item (30 items per test form, 60 items in total) to a subskill of listening comprehension in a second language. Four listening subskills (i.e., comprehending explicit aural discourse, recalling specific details, demonstrating understanding of the general topic or main idea, and inferring implicit ideas) were defined drawing on second language listening literature in L2 assessment (e.g., Buck, 2001; Field, 2013; Rost, 1990). From this first iteration, two confirmatory factor models were specified. In a second iteration, the panel was asked how much they agreed (on a Likert type statement ranging from 1 = strongly disagree, 2 = disagree, 3 = agree, to 4 strongly agree) with each of the item-subskill associations. When there was disagreement among the panelists, discussions were held until reaching consensus, or until the panel agreed on a different subskill to be associated with the item in question.

Agreement was calculated using Gwet's  $AC_1$  statistic (Gwet, 2012) for 3 raters or more. The letters AC in "AC<sub>1</sub> statistic" stand for Agreement Coefficient. Gwet's  $AC_1$  statistic represents the probability for 3 raters or more to agree for cause based on all ratings except those

associated with panel members that may produce an agreement by pure chance. This definition is in line with the goal set by Cohen (1960) for Kappa. The percent agreement  $p_a$  used for  $AC_1$  is the same as the one used for Kappa and other kappa-like statistics (Gwet, 2012). However, the problem arises when calculating percent agreement when the number of raters is 3 or more. Gwet (2012) proposed a formula that leads to the multiple-rater version of the  $AC_1$  given by:

$$\hat{\gamma}_1 = \frac{P_a - p_e}{1 - p_e}, \text{ where } \left\{ p_a = \frac{1}{n'} \sum_{i=1}^{n'} \sum_{k=1}^q \frac{r_{ik}(r_{ik} - 1)}{r_i(r_i - 1)} \right\},$$

$$p_e = \frac{1}{q - 1} \sum_{k=1}^q \pi_k (1 - \pi_k),$$

$$\text{and } \pi_k = \frac{1}{n} \sum_{i=1}^n \frac{r_{ik}}{r_i}.$$

In the equation above,  $q$  is the number of categories,  $r_{ik}$  is the count of raters who classified item  $i$  into category  $k$ ,  $r_i$  is the count of raters who rated item  $i$ ,  $n$  is the total count of items, and  $n'$  the count of raters who rated 2 items or more. For this agreement coefficient, the null and the alternative hypotheses are as follows:

$H_0$  = the expert panel categorized the test items randomly.

$H_1$  = the expert panel categorized test items reflecting the intrinsic properties of L2 subskills.

Confirmatory factor analyses models were specified for both Form A and Form B based on the panel's recommendation; the results of these analyses are presented in Chapter four. The section below elaborates on CFA methodology, model specification, identification, estimation as well as fit indices. CFA was used to evaluate claim one in the interpretive argument.

## 3.2 Evaluation of Claim One Using Confirmatory Factor Analysis

### 3.2.1 Confirmatory factor analysis

Confirmatory factor analysis (CFA) is a commonly used technique to study the latent structure of language assessment and gather evidence of construct validity (e.g., Bae & Bachman, 1998; Bae & Bachman, 2010; Cai, 2012; Kim, 2010; In'nami & Koizumi, 2011;

Llosa, 2007; Sawaki, Stricker & Oranje, 2009; Tobia, Ciancaleoni & Bonifacci, 2016; Yoo & Manna, 2015). CFA was used to gather the empirical evidence to support Claim one in the interpretive argument, put forth in Chapter two, that is:

### **Claim one**

The *Test de connaissance du français* assesses listening skills that are relevant to the construct under examination. As such, the listening component of the TCF assesses features of L2 listening such as: the understanding of explicit information and implicit information. More specifically, the test is comprised of test items that assess four major sub-skills in L2 listening: (a) comprehending explicit aural discourse, (b) recalling specific details, (c) demonstrating understanding of the general topic or main idea, and (d) inferring implicit ideas.

CFA is a type of structural equation modeling (SEM), useful for testing latent trait models and “observe the relationships between observed measures or indicators and latent variables of factors” (Brown, 2015, p. 1). CFA is almost always used during the process of scale development to examine the latent structure of a test instrument and widely used at later stages of scale development as well. A fundamental feature of CFA is its hypothesis driven nature, which requires the researcher to specify all aspects of the model and then tests the hypothesis statistically. This specification can be based on relevant theory or empirical evidence of the number of factors that exist in the data (Brown, 2015).

#### **3.2.1.1 CFA model specification**

CFA proceeds in three steps: model specification, in which hypothesized latent variables, observed variables, and the expected relationships between the two are articulated. Model identification refers to the comparison of the number of parameters to be estimated with the number of units necessary for information. Finally, model parameters are estimated and fit statistics are calculated based on the residuals, which are the differences between the theorized CFA model and the actual data.

CFA aims to reproduce the sample variance-covariance matrix by parameter estimates of the measurement solution (e.g., factor loadings, factor covariances, etc.). Figure 5 below presents an example of a reflective CFA model with two correlated latent variables (also call attributes or factors),  $\xi_1$  and  $\xi_2$ , each causing the variation on the indicators or items. The first three measures ( $X_1$ ,  $X_2$ , and  $X_3$ ) are indicators of one latent construct ( $\xi_1$ ), whereas the next three

measures ( $X_4$ ,  $X_5$ , and  $X_6$ ) are indicators of another latent construct ( $\xi_2$ ). Indicators  $X_4$ ,  $X_5$ , and  $X_6$  are congeneric because they share a common factor ( $\xi_2$ ). An indicator is not considered congeneric if it loads on more than one factor. The lambdas ( $\lambda$ ) represent the factor loadings and the deltas ( $\delta$ ) the measurement errors. The squared factor loading represents the proportion of the variance in the indicator (or item) that is explained by the factor and is often referred to as communality.

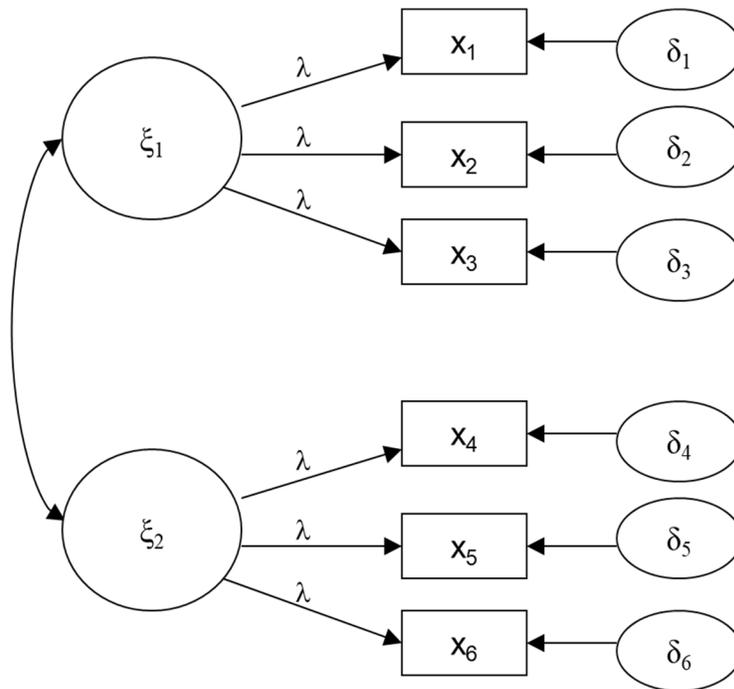


Figure 5 Illustration of a two-factor model with correlated factors

A simple confirmatory factor model is mathematically expressed as  $x = \Lambda\xi + \delta$ , where  $x$  is the reflective measure (i.e., the vectors of test items),  $\Lambda$  is the matrix of factor loadings connecting latent traits to test item performance,  $\xi$  is the vector of latent variables and  $\delta$  is the vector of measurement errors or uniqueness. This equation can be expanded into its components at the level of each individual item. That is,  $x_i = \lambda_{1i}\xi_1 + \lambda_{2i}\xi_2 + \delta_i$ .

### 3.2.1.2 CFA model identification

Brown (2015) suggested that to estimate the parameters in CFA, the measurement model must be identified. A model is identified, if based on known information (i.e., the variances and covariances in the sample input matrix), it is possible to obtain a unique set of parameters

estimates for each parameter in the model whose values are unknown (e.g., factor loadings, factor correlations). If the number of parameters is greater than the known information the model is under-identified and if no difference exists, the model is just identified, but if the known information exceeds the numbers of parameters the model is identified.

### 3.2.1.3 Guidelines for model identification

Brown (2015) and Kenny and Milan (2012) proposed two basic guidelines for model identification in CFA and can be summarized as follows:

1. Regardless of the complexity of the model (e.g., one factor vs. multiple factors, size of indicator set), to identify variables with latent variables the units of measurement of the latent variable must be fixed. This is usually done by fixing the loading of one indicator, called the marker variable, to 1 or fixing the variance of the factor (usually to a value of one).
2. Regardless of the model complexity, the number of pieces of information in the input matrix (e.g., indicator variances and covariances) must be equal or exceed the number of freely estimated model parameters (e.g., factor loadings, factor variances-covariances, indicator error variances-covariances). This point was suggested above.

Note that if the model contains correlated errors, then the identification rules need to be modified. For the model to be identified, then, each latent variable need two indicators that do not have correlated errors and every pair of latent variables needs at least one indicator of each that does not share correlated errors.

### 3.2.1.4 CFA model estimation

The objective of CFA is to obtain estimates for each parameter of the measurement model to produce a predicted variance-covariance matrix ( $\Sigma$ ) that resembles or approximate the sample variance-covariance matrix ( $S$ ) as closely as possible. There are multiple fitting functions to minimize the difference between  $\Sigma$  and  $S$  and Brown (2015) points out that the fitting function most widely used in applied CFA research is maximum likelihood (ML), which can be written as:

$$F_{ML} = \ln |S| - \ln |\Sigma| + \text{trace} [(S) (\Sigma^{-1})] - p,$$

where  $|S|$  is the determinant of the input variance-covariance matrix;  $|\Sigma|$  of the predicted variance-covariance matrix;  $p$  is the order of the input matrix (i.e., the number of the input

indicators); and  $\ln$  is the natural logarithm. The objective of ML is to minimize the difference between these matrices summaries for  $S$  and  $\Sigma$ . The underlying principle of ML estimation in CFA is to find the model parameter estimates that maximize the probability of observing the available data if the data are collected from the same population again. In other words, ML seeks to find the parameter values that make the observed data most likely to be obtained in repeated instances drawing from the similar samples of a population. The discrepancy between  $S$  and  $\Sigma$  yields a residual matrix and this residual matrix is used in the calculation of various goodness-of-fit indices.

Maximum likelihood (ML) is a normal-theory estimator assuming continuous and multivariate normally distributed observed variables. However, many item-level confirmatory factor analysis in language testing are characterized by a dichotomous or binary level of measurement. Although the items of a test are assumed to be measures of a continuous construct, the observed responses are discrete realizations of a small number of categories (e.g., correct or incorrect in binary data) and, thus, lack the scale distributional properties assumed by normal-theory estimators (Moshagen & Musch, 2014). In addition, it is well known that the application of normal-theory estimators based on the product-moment covariance matrix to truly binary data results in biased parameter estimates, standard errors, and test statistics, calling the validity of conclusions drawn from such studies into question (DiStefano, 2002; Muthén & Kaplan, 1985). An alternative approach tailored to the estimation of CFA models based on tetrachoric correlations of binary data is the robust weighted least squares estimator (WLSMV). Moshagen and Musch (2014) explained that the basic idea of the robust WLS is to substitute the full weight matrix in the fitting function,  $F_{wls}(\theta) = (r - \rho)'W^{-1}(r - \rho)$ , by a weight matrix containing only the diagonal elements, that is, the asymptotic variances of the thresholds and tetrachoric correlation estimates. This modification considerably reduces the number of nonzero elements, thereby greatly facilitating inversion. In a simulation study, Moshagen and Musch (2014) found that for binary data, the robust WLS yielded proper convergence rates ( $\geq 99$ ) as a function of the number of indicator per factors ( $\geq 4$ ), primary factor loadings ( $\geq .400$ ), and sample size ( $n \geq 500$ ). Mplus version 8 (Muthén & Muthén, 2017) was used to estimate the CFA models the panel suggested and it was instructed to use the robust WLS for the analyses.

There are various fit indices that are calculated to test how well the model fit the data. Fit indices can be broadly categorized into three categories: absolute fit; fit adjusting for model parsimony and comparative or incremental fit. As a rule of thumb, it is not recommended to adhere to a single fit to determine model fit. In fact, it is suggested to use multiple fit indices to assess the fit of the model to the data. The paragraphs below summarize the most commonly used fit indices:

a) The chi-square ( $\chi^2$ ) test is a classic absolute goodness of fit index. It tests the null hypothesis that  $S = \Sigma$ . Thus, a statistically significant  $\chi^2$  supports the alternative hypothesis that  $S \neq \Sigma$ , that the model estimates do not sufficiently reproduce the sample variances and covariances. The  $\chi^2$  test has been highly criticized and it is rarely used as a sole index of model fit. Important criticisms include its inflation by sample size.

b) Root mean square error of approximation (RMSEA) is a parsimony correct fit index that corrects for the sensitivity of  $\chi^2$  test to sample size. As such, the RMSEA is a population-based index that relies on the noncentral  $\chi^2$  distribution. The RMSEA is an error of approximation index because it assesses the extent to which the model fits reasonably well in the population. Lower RMSEA indices indicate better fit. RMSEA values of 0.05 or below indicates good fit (Browne & Cudek, 1993)

c) Comparative fit index (CFI) and Tucker-Lewis indices are also referred to as incremental fit indices and evaluate the fit of a model relative to a baseline model; indices greater than 0.95 indicate good model fit (Hu & Bentler, 1999).

The following section elaborates on the measurement model that was used to evaluate claim two in the interpretive use argument for the TCF.

### **3.3 Evaluation of Claim Two Using Rasch-based DIF analysis and the Standardization DIF Method**

Various studies have addressed this fairness in language assessment (e.g., Aryadoust, 2012; Aryadoust, Goh & Kim, 2011; Batty, 2014; Pae, 2004, 2012; Ockey, 2007) by drawing on Rasch measurement to estimate item difficulties and examine if test takers from different groups (e.g., gender, age, first language or other person's characteristics irrelevant to the testing context) have different likelihoods of success on an item, after they have been matched on the proficiency of interest (Sireci & Rios, 2013). Although sources of differential item functioning

are difficult to identify concretely, Rasch measurement have been useful in framing empirical studies concerned about test fairness. Nevertheless, Rasch-based DIF analysis requires relatively large sample sizes (e.g., at least 200 examinees per parameter per group for binary scored items) to ensure item parameter estimation stability (Sireci & Rios, 2013; Zumbo, 1999). Since the Rasch model is a simple (one difficulty parameter), albeit not simplistic, measurement model, on grounds of the sample size for this study, it seemed reasonable to adhere to Rasch-based DIF analysis for the relatively large subgroups ( $n < 150$ ). Rasch measurement is a prevalent methodology when assessing the fairness of language tests and has proven being useful when the data conforms to Rasch measurement assumptions (e.g., Aryadoust, 2013). The Rasch model also boasts about the property of measurement invariance or specificity, making it more attractive than IRT-based DIF analysis (e.g., the two-parameter logistic model and the three-parameter logistic model).

After stratifying the test data by different demographic variables (e.g., first language) several groups decreased in sample size and this would have potentially yielded unstable item parameter estimates, thereby threatening the validity of the results. Alternatively, the standardization index, a less restrictive, observed score method, was used to address potential issues of test fairness in groups with relatively smaller sample sizes. This method is described in more detail in a section following Rasch measurement. Rasch-based DIF analysis and the Standardization index were used to evaluate claim two in the interpretation/use argument for the TCF:

### **Claim two**

The *Test de Connaissance du français* is a fair measure of L2 listening proficiency. As such, the test does not favor group membership regardless of gender, first language, age, or geographical location (only North and West Africa are included) of candidates who submit test scores to immigrate to Canada through the province of Quebec.

Before describing the methods for evaluating item bias, the term must be clearly defined. DIF is a statistical observation that involves matching test takers from different groups on the latent trait being assessed and then looking for performance differences on an item or a test. DIF does not necessarily signify item bias. It simply flags the item for statistical significance. Sireci (2011, p. 221) suggested that item bias is present when “an item has been flagged for DIF and the reason for the DIF is traced to a factor irrelevant to the construct the test is intended to

assess”. There has been a plethora of DIF methodologies that have been proposed since the 1970’s (e.g., Angoff, 1972; Dorans & Kulick, 1986; Holland & Thayer, 1988; Holland & Wainer, 1993; Lord, 1980; Raju, 1988, 1990; Shealy & Stout, 1993; Swaminathan & Rogers, 1990). These propositions make use of raw scores or IRT-based measures to calculate DIF analysis and potentially item bias.

Before Rasch-based DIF analyses are conducted the data must conform to the Rasch model expectations. The sections below spell out the assumptions of the Rasch model and new development regarding fit statistics in the context of Rasch measurement.

### 3.3.1 The Rasch model

Danish mathematician Georg Rasch (1960) developed a measurement model for responses to dichotomous item. This model is eponymously referred to as the classical unidimensional Rasch model. The Rasch model is a unidimensional measurement model that calculates the relationship between item difficulty and person ability as the ratio of positive or negative endorsement and expresses the difference in log-odds, or logits (Embretson & Reise, 2000). In addition, “Rasch measurement begins by recognizing that every datum originates as a recoding of the presence or absence of a quality” (Wright, 1996, p. 4). Taking the natural logarithm of the odds ratio yields the log odds form of the dichotomous Rasch model:

$$\ln \left[ \frac{p_{ni}}{1 - p_{ni}} \right] = \beta_n - \delta_i$$

where  $P_{ni}$  is the probability of person  $n$  succeeding or failing item  $i$ ,  $\beta_n$  is the ability of the person  $n$ , and  $\delta_i$  is the difficulty level of the item  $i$  (Wright, 1996). This original form of the Rasch model is only used with dichotomous data. That is, when the raw data take forms as: yes/no, present/absent or right/wrong. In this case, the observations are scored as:  $x = 0, 1$ . Rasch (1977) pointed out that the model specified above has a special property called *specific objectivity*. The principle of specific objectivity is that comparisons between two objects are free from the conditions under which the comparisons are made. For example, the comparison between two persons is not influenced when specific items are used. It is worth noting, however, that specific objectivity is a property of the Rasch model and it is only a property of the data if it can be shown that the data conform to the Rasch model. Applying the Rasch model does not force the data to conform to the Rasch model. Thus, “to claim and deploy the benefits of using the Rasch

model, one must first demonstrate consistency of the data to the model” (Wu & Adams, 2013, p. 340).

The dichotomous unidimensional Rasch model (1960) has inspired other researchers to extend the model to analyze other types of data. For instance, Andrich (1978) extended the dichotomous Rasch model by proposing the rating scale model in order to analyze data that are collected through items using Likert response format. Similarly, Masters (1982) extended the dichotomous Rasch model to a Rasch model for partial credit scoring. One of the advantages of the Rasch model is that it uses raw data and apply a mathematical function to create interval “like” data. Such a property is desirable as “raw counts cannot be the measures sought because in their raw state, they have little inferential value” (Wright & Mok, 2004, p. 2). Thus, to develop metric meaning, the counts must be incorporated into a stochastic process that produces inferential stability (Wright & Mok, 2004). The Rasch model achieves this goal by transforming raw data into interval data through a log-odds transformation of the raw scores. Finally, through Rasch analysis one can explore the “targeting” of the instrument. That is, if the test items are distributed along the ability continuum of the target population,

In summary, Rasch analysis provides a wide range of information that one would not have access to otherwise by using traditional approaches (i.e., classical test theory) to instrument validation. Nevertheless, “all the advantages offered by the Rasch model come with a price” (Chou & Wang, 2010, p. 718). Following this line of reasoning, Rasch models have assumptions that need to be respected so that the Rasch-based inferences are meaningful and valid. Thus, the assessment of dimensionality, local independence and goodness-of-fit of the data to the model are of paramount importance for meaningful Rasch-based interpretations. In this order of assumptions, the following subsection elaborates on each of the requirements expected by the Rasch model assumptions.

### **3.3.1.1 Dimensionality**

Dimensionality refers to the coherence of the data in “measuring” a single or multiple attributes. When using the classical Rasch (1960) model, the data are assumed to be unidimensional and this assumption can be readily tested with principal component analysis of standardized Rasch residuals, according to Linacre (1998a):

The aim of PCAR is to attempt to extract the common factor that explains the most residual variance under the hypothesis that there is such a factor, if this factor is discovered to merely explain random noise then there is no meaningful structure in the residuals (Linacre, 1998a, p. 636).

In other words, once the data fit the Rasch model, a principal component analysis on the remaining random noise (i.e., standardized residuals) should not explain any additional factors; if this is the case one could consider this as a sign of unidimensionality. It is important to note however, that there is an ongoing debate regarding the statistics required (e.g., 1.5 or less eigenvalues of the first contrast) to consider the data unidimensional. In fact, Chou and Wang (2010) suggested that having a fixed cut point for determination of dimensionality through PCAR might not be the best technique to adopt because the values of PCAR statistics are sensitive to sample sizes and test length. Instead, they suggested statistical approaches for multivariate independence to test the correlation matrix obtained from standardized residuals. Following PCAR of a dataset, Raïche (2005) suggested that researchers simulate a second, unidimensional dataset, containing the same number of people and items and with the same person and items reliability indices as the actual dataset, and compare its PCAR results to that in the actual data. If the two PCAR results approximate each other, this is additional evidence of unidimensionality (Linacre, 1998a; Wright, 1994, 1996).

### **3.3.1.2 Local independence**

The unidimensional Rasch model also assumes local independence, which requires weak or zero correlations among the items after the effect of the underlying trait is conditioned out, i.e., the correlation of residuals should be zero (Baghaei, 2008). In the context of the TCF, the response to one item should not lead the test taker to respond another item. If the condition of local independence is not satisfied the items are locally dependent and perhaps a second dimension exist. This can result in spurious and misleading interpretations and can also inflate reliability and “give a fake impression of the precision and quality of the test” (Baghaei, 2008, p. 1105), “affecting parameter estimation in Rasch analysis” (Tennant & Conaghan, 2007, p. 1361).

### 3.3.1.3 Rasch fit statistics

Given the emphasis on the difference between observed response and the expected response by the Rasch model, “it comes as no surprise that many of the fit statistics for the Rasch model are chi-square based” (Tennant & Conaghan, 2007, p. 1360) and “residual-based” (Wu & Adams, 2013, p. 352). Determining whether the data fit the Rasch model is a crucial step in Rasch measurement theory. Since the establishment of Rasch measurement models, there has been an ongoing, rather unsettled debate, regarding the appropriate fit statistics for assessing the goodness-of-fit of the data to the model. Note that in Rasch measurement, it is expected that the *data* fit the model, whereas in item response theory (IRT), *models* are expected to fit the data. Researchers are partially left at the expense of fit indices included in software packages (e.g., R packages: TAM, and eRm; jMetrik; Winsteps; RUMM2030, etc.) to determine data fit. For instance, in RUMM2030 the chi-square statistics compares the difference in observed values with expected values across groups representing different ability levels across the trait to be measured (e.g., L2 listening comprehension). According to Tennant and Conaghan (2007), “the chi-square values are summed up to give the overall chi-square for the item, with degrees of freedom being the number of groups minus one” (p. 1360). If the critical value is  $p < 0.05$ , the item does not fit the model. In other words, “a significant chi-square indicates that the hierarchical ordering of the items varies across the trait, compromising the required property of invariance” (Tennant & Conaghan, 2007, p. 1360).

There exists a plethora of formulas for calculating the fit statistics of Rasch models, and most statistics used to assess the goodness-of-fit of the data to the Rasch models are methods that summarize the squared standardized residuals (R.M. Smith, 2004). R. M. Smith (1991) offered a detailed account on fit statistics of the time and concluded that the ongoing debate regarding the stability of fit statistics led to inconclusive results.

Although the fit statistics used to determine the goodness-of-fit of the data to the Rasch model seemed a bit unstable and unsettled during the 1980’s and 1990’s, there have been recent developments regarding this concern. Wu and Adams (2013) proposed some guidelines that can be helpful when determining the goodness-of-fit of the data to the Rasch model. In this regard, they explained that, traditionally, researchers have tended to suggest rules of thumb (e.g., Bond & Fox, 2007) to suggest critical values for data fit, especially in the case of mean

square statistics. However, Wu and Adams (2013) critique this position by explaining that, for example, the mean square statistics normally used to determine item fit is sensitive to sample size. Thus, a traditional rule of thumb of a mean square of acceptable values between 0.70 and 1.30 might result in spurious interpretation of data fit when the sample size exceeds  $n = 200$  cases. In this regard, they proposed a formula to estimate better mean square statistics based on sample sizes. This formula takes the following form:

$$1 \pm 2\sqrt{\frac{2}{N}}$$

Wu and Adams (2013, p. 352) ascertained that “there should no longer be speculations about the instability of the fit statistics, but rather, there can now be well informed broad guidelines for the use of fit statistics. This is an important contribution because measurement models are idealizations of empirical observations and in the context of Rasch measurement the data are highly unlikely to fit a given Rasch model perfectly. Rather than assessing the fit of the data to the model following fixed rules of thumb, it seems more appropriate to consider sample size as well as the practical utility of the model because “we need to know whether the data fit the model usefully and when misfit is found, how much misfit there is, where it comes from, and what to do about it” (Eckes, 2015, p. 68).

Based on the equation above, Table 8 provides the approximate mean square fit statistics that should be considered for item fit values for different sample sizes. This is indeed a more empirically based fit index to determine if the data conform to the Rasch model expectations. According to this new development in mean square fit statistics, if one bases the data fit indices on traditional practice of mean square values, one could classify problematic items as fitting items. Thus, the dialectic dialogue regarding fit statistics for Rasch measurement models appears to be refining our ideas regarding this concern.

Table VIII Wu and Adams (2013) guidelines for mean square fit statistics

$n$	Lower bound	Upper bound
100	0.72	1.28
200	0.80	1.20
500	0.87	1.13
1000	0.91	1.09
1500	0.93	1.07
2000	0.94	1.06

On Table 8 one can notice that as sample size increases, the range of fit statistics becomes more conservative. This may pose practical difficulties with large sample sizes as there is a risk of encountering a reasonable high number of items that might not conform to the Rasch model. In this case, Eckes (2015) recommends to adhere to Wu and Adams (2013) only when running the first analysis to identify highly misfitting items because their criteria may become too strict and flag misfit that would be too small to distort measurement.

### 3.3.1.4 Rasch-based differential item functioning

Differential item functioning (DIF) compares the difficulty of an item for a person sample of interest, the focal group (F), with the difficulty of the item for another group, the reference group (R). As noted earlier, DIF does not necessarily mean that a given test item is biased towards a given group. However, DIF is a necessary condition to establish potential bias. When the data have conformed to the Rasch model expectations, DIF measures can be instrumental for the investigation of item bias. When comparing the item difficulty parameters between a reference and a focal group, it is important to consider the sample size of both groups, since the difference in logits to flag statistically significant DIF can be affected by sample size.

Linacre (2013) suggests that if having similar standard deviations across sub-groups one can construct a nomogram based on a student's  $t$ -statistic to determine the DIF size required in the analysis to have substantive consequences on the person sample of interest. Using the equation below with two independent groups of different sizes: focal group ( $N_F$ ) and reference group ( $N_R$ ), but equal standard deviations,  $S$ , it is possible to obtain the DIF difference ( $D$ ) that will be required to flag items for differential item functioning.

$$t = \frac{D}{S} \sqrt{\frac{N_F N_R}{N_F + N_R}}$$

Nevertheless, the differential item functioning (if any) resulting from these two groups might be a little unstable due to the discrepancy of group sizes (Linacre, 2013). Thus, if any significant DIF exists the interpretation should be made with caution. In this study, the Rasch measurement software Winsteps 4.0.1 (Linacre, 2017) was used to fit the data to the Rasch model. In the context of Rasch-based DIF under Winsteps, DIF measures (item difficulty in logits) are estimated for the reference and focal groups, then, the difference or “DIF contrast” is calculated between the reference group and the focal groups. If the absolute value for the DIF

contrast exceeds 0.43 logits, the item in question exhibits differential functioning. If the DIF contrast is negative, the item favors the reference group, if positive, the item favors the focal group. Educational Testing Service (ETS) uses Delta measurement units to describe DIF and categorizes DIF in terms of magnitudes: A = negligible, B= slight to moderate, and C = moderate to large (Zwick, Thayer & Lewis, 1999), where one Delta unit equals 0.426 logits and is considered as slight to moderate DIF (B). For DIF to be considered as moderate to large, a logit difference between the compared groups of at least 0.64 is needed.

There are various strategies to address DIF when it is discovered. The item exhibiting DIF can be either removed or resolved. The former simply requires the item with the largest DIF to be dropped, then the analysis is undertaken again to examine if DIF is still present in the remaining items. This is a traditional approach. The latter refers to treating the item(s) in question as different item(s) for different groups (Tennant & Pallant, 2007). That is, splitting the item with the largest DIF into the number of groups under investigation. For example, if a DIF study explores measurement invariance across gender, resolving or splitting the item would consist in creating two items from the item under DIF investigation. Each item would have active data for one group and missing data for the other group. This maintains the same raw scores for test takers, but produces different measures for each group (Linacre, 2016). Resolving or splitting for DIF is an enticing approach to address measurement invariance because real DIF in some items favoring a group can induce the appearance of DIF, or artificial DIF, in other items favoring the other group (Andrich & Hagquist, 2012). This has the advantage of identifying DIF caused by the artifact of measurement while maintaining test length and reliability. Andrich and Hagquist (2012) outlined a sequential, two-step procedure to resolve item(s) exhibiting DIF. First, we hypothesize that the item with the largest DIF is the one with the real DIF, this item needs to be dealt with first. Second, we resolve the hypothesized item into two new items, as in the example above, one item would have the responses just for males and the other responses just for females. These new items can be called the resolved male and resolved female items. This procedure renders structurally missing responses, but such a data matrix with missing data can be analyzed readily according to the Rasch model. Once the item resolved items have been created the data are reanalyzed, if the item that is resolved is the only one with real DIF, then the remaining items will not yield artificial DIF. This procedure continues iteratively until DIF is nonexistent or negligible across all items.

Although resolving items following Andrich and Hagquist (2012) guidelines to address items exhibiting differential item functioning might seem an enticing approach, in this study, the item with the largest DIF for a given class (e.g., gender) was dropped to explore whether DIF in the other items was induced by the item with the largest DIF. Then, the analysis was undertaken again to examine the persistence of DIF. If DIF persisted, a sequential item deletion approach was followed iteratively.

### 3.3.2 The Standardization index

The standardization index was first introduced by Dorans and Kulick (1986). The index essentially computes an average difference across groups in the proportion of examinees selecting a particular response, conditional on the matching variable. There are signed and unsigned indices to address both uniform and non-uniform DIF respectively (Sireci & Rios, 2013). The signed index is,

$$D_{\text{std}} = \frac{\sum_{s=1}^S K_s [P_{fs} - P_{rs}]}{\sum_{s=1}^S K_s},$$

where  $S$  = score level,  $k_s$  = weigh of the score level,  $P_{fs}$  = the proportion correct at score level  $s$  for the focal group,  $P_{rs}$  = the proportion correct at score level  $s$  for the reference group. This calculation is referred to as the signed version because “positive differences at one point of the scale can be offset by negative differences at another point” (Sireci, 2011, p. 226). This will not flag non-uniform DIF (Sireci & Rios, 2013), but to evaluate non-uniform DIF, the unsigned version of the index is used:

$$D_{\text{std}} = \frac{\sum_{s=1}^S |K_s [P_{fs} - P_{rs}]|}{\sum_{s=1}^S K_s},$$

The standardization index matches examinees at particular levels, and compares the difference in  $p$  values (proportion correct) for individual items between the reference and focal groups. A score level can represent either a single total score or a grouping of scores. In deciding how to group examinees into two different score levels, sample size and score distributions may play critical roles. The advantages of the standardization index are (a) it is easy to calculate, (b) it allows for easy interpretation and explanation to stakeholders (Demars, 2011), (c) the results can be easily graphed, and (d) it can be applied to small sample sizes (Muñiz, Hambleton & Xing, 2001).

The signed version of the standardization index ranges from -1.0 to 1.0, while the unsigned version ranges from 0 to 1.0. Although there is no statistical test, an effect size can be computed to assist with the practical interpretations of DIF. Researchers have suggested a critical value of  $\pm 0.10$  to flag items for DIF (Dorans & Kulick, 1986; Muñiz *et al.*, 2001). The idea behind the effect size criterion is that if 10 items are deemed to possess DIF for one group, a one-point total score disparity between groups would be apparent. In this study, the standardization index was calculated with a DOS program (Robin, 2001).

The following section elaborates on the measurement model that was used to evaluate claim three in the interpretive use argument for the listening component of the *Test de connaissance du français* (TCF).

### **3.4 Evaluation of Claim Three Using the Nominal Response Model**

The nominal response model (Bock, 1972) and its extensions have received limited applications to real world measures. In fact, a handful of empirical studies have employed the NRM to model data (e.g., De Ayala & Salva-Bolesta, 1999; Demars, 2003; Johnson-Preston & Reise, 2014; Johnson-Preston *et al.*, 2015; Revuelta, 2005; Samejina, 1979; Thissen & Steinberg, 1984; Thissen, Steinberg & Fitzpatrick, 1989). Among these studies, implementation of the NRM have sought to describe linking under the NRM (Baker, 1993), examine item parameter recovery (de Ayala & Sava-Bolesta, 1999; Preston & Reise, 2014) and develop indices to identify model violations (Glas, 1999). The measurement model presented here sought to provide empirical evidence to support the interpretation/use argument for the listening component of the TCF:

#### **Claim three**

Selected response items in the form of multiple-choice (MC) are a useful method to gather evidence of L2 listening comprehension. The keyed responses and distractors provide empirical evidence that supports the interpretation and use of MC items to assess L2 listening comprehension.

#### **3.4.1 The nominal response model (NRM)**

Most applications of item response theory (IRT) model have been based on dichotomous models, such as the two- and three-parameter logistic models or models that assumed an expected order in the categories of the items (e.g., Samejina, 1969). Bock (1972) introduced the

NRM to model item responses with two or more nominal categories such as multiple choice items in which “item alternatives represent unordered responses” (De Ayala & Sava-Bolesta, 1999, p. 4). The NRM provides a direct expression for obtaining the probability of an examinee with trait level  $\theta$  responding in the  $j$ th category of item  $i$ . The probability of responding in a particular category as a function of  $\theta$  can be represented by the option response function (ORF) (also known as trace line, or category or option characteristic curve). Bock’s (1972) NRM can be written as:

$$P_{ij}(\theta) = \frac{\exp(\zeta_{ik} + \lambda_{ik}\theta)}{\sum_{k=1}^m \exp(\zeta_{ik} + \lambda_{ik}\theta)},$$

where  $\lambda_{ik}$  and  $\zeta_{ik}$  are the slope and the intercept parameters respectively, of the nonlinear response function associated with the  $k$ th category of item  $i$  and  $m$  is the number of categories in item  $i$ . The  $\lambda_{ik}$  parameters are analogous to the traditional discrimination indices. De Ayala and Sava-Bolesta (1999, p. 4) ascertained that “a cross-tabulation of  $\theta$  groups by item alternative shows that a category with a large  $\zeta_{ik}$  reflects a response pattern in which as the group’s theta increases as the number of persons who answer the item in that category increases”. That is, “the larger the intercept parameter  $\zeta_{ik}$ , the greater relative frequency of responses in that category” (Preston & Reise, 2015, p. 388). In the NRM, the slope parameters reflect the interaction between a category’s difficulty and how well it discriminates. These parameters represent the linear relationship between the latent trait (e.g., listening comprehension ability) and the log-odds of responding in a given category (e.g., options in a multiple-choice item).

Penfield (2014) identified a drawback of the NRM applied to multiple-choice items. That is, the item response function for the correct option has a lower asymptote of zero (i.e., approaches zero as the target trait level becomes low), and thus, the NRM lacks the flexibility to account for the chance of guessing the correct option by individuals with lower levels of the trait. Given this limitation of the NRM, alternative models have been proposed that estimate “guessing” or “don’t know” parameters (Revuelta, 2005; Samejina, 1979; Thissen & Steinberg, 1984; Thissen, Steinberg & Fitzpatrick, 1989). These models have been termed multiple-choice models. Note, however, that they are extensions of Bock’s (1972) pioneer work.

To optimize the evaluation of the third claim in the interpretive use argument for the L2 listening component of the TCF, this study included the analysis of the distractors’ data using

Bock's (1972) nominal response model. This allowed to estimate test taker abilities (thetas) and the probability of choosing each of the options (both key and distractors).

In this study, the NRM fit was evaluated based on measurement invariance across gender using the conventional likelihood ratio  $\chi^2$  statistics. The NRM model was fitted independently to both males and females across test forms. The variable gender was chosen to create the groups to maintain relatively large sample sizes and stable model parameters. The log-likelihood difference was used to calculate the statistical significance of parameter invariance across gender. A Bonferroni adjustment ( $\alpha/n = 0.05/30 = p < .001$ ), using the conventional alpha level (0.05) and divided by the number of iterations, that is the number of items in each test form (i.e., 30), was used to control for type I error. The models were fitted using the R package mcIRT (Reif, 2015).

### **3.5 Software used in this study**

Several software were used to analyze the data and to create the graphics in this study. The R language (version 3.4.1) was used to compute the descriptive statistics and create the graphics and the R package mcIRT (version 0.41) was used to fit the nominal response models. R is a free open access software that offers a lot flexibility for analyses and graphics (note that programming in R language is needed). AgreeStat (version 2015.6) was used to analyze the panel data and determine the level of agreement among multiple raters (i.e., panelists). This program is embedded in a stand-alone Excel Workbook and is based on the work of Kilem L. Gwet. The  $AC_1$  index used in the analyses overcomes the limitations of traditional kappa, which can be only calculated for two raters. The CFA models were fitted with Mplus (version 8); Mplus offers a myriad of options that take into account the type of data under analysis. For instance, for binary data, Mplus uses robust weighted least squares estimation to correct for the violation of multivariate normality which inherent in discrete distributions. Finally, Winsteps (version 4.0.1) was used in the Rasch analyses and the output from the DIF analyses were used to create the graphics with an R function using the R package ggplot2.

## Chapter 4: Results

The purpose of this study was to build a validity argument for the listening component of the *Test de connaissance du français*. To achieve this goal, test score results were analyzed with several analytical methods, including item content analysis, confirmatory factor analysis (CFA), Rasch-based differential item functioning (DIF), and the nominal response model (Bock, 1972). This chapter reports on the results of these analyses and attempts to answer each of the research questions. In turn, these results are used to build the validity argument for the TCF. The argument being made on behalf of the TCF in the context of Quebec immigration only includes the scoring, generalization and explanation inferences described in Chapter two. The extrapolation and decision inferences will need to be addressed to complete the validity argument for TCF in the context of Quebec immigration.

Table 9 presents a summary of the interpretation/use argument and the relevant data analysis methods used to address each of the claims. The findings from each of the research questions are used as backings for one or more of the warrants supporting the claims in the interpretation/use argument (Kane, 2006, 2013). The argument-based framework organizes these findings into a coherent validity argument and evaluates whether the collected evidence supports the warrants for each of the inferences supporting the claims in the interpretation/use argument for the TCF in the context of Quebec immigration.

Table IX TCF Listening Validation Studies

Study	Analytical tool	Research Questions	Claims/Statements
1. Factorial structure of the test.	Item coding, confirmatory factor analysis (CFA)	1. What listening sub-skills do the <i>Test de Connaissance du français</i> assess? Do test items contribute exclusively to the assessment of the listening construct? Is there supporting evidence indicating that test items are contaminated by construct irrelevant variance or construct underrepresentation?	1. The <i>Test de Connaissance du français</i> (TCF) assesses listening skills that are relevant to the construct under examination. As such, the listening component of the TCF assesses features of L2 listening such as: understanding of explicit information and implicit information. More specifically, the test is comprised of test items that assess four major sub-skills in L2 listening: (i) comprehending explicit aural discourse, (ii) recalling specific details, (iii) understanding the general topic or main idea, and (iv) inferring implicit ideas.
2. Rasch DIF analysis of the test	Rasch and differential item functioning (DIF)	2. Does the <i>Test de Connaissance du français</i> exhibit differential item functioning (DIF), leading to test bias against gender, first language, age, and geographical location of examinees, threatening the validity of test score-based interpretations and uses of the test in the context of Quebec immigration?	2. The <i>Test de Connaissance du français</i> is a fair measure of L2 listening proficiency and does not favor group membership regardless of gender, first language, age, or geographical location (Only North and West Africa are included) of candidates who submit test scores to immigrate to Canada through the province of Quebec.
3. Distractor functioning	Nominal response model (NRM)	3. What information do the distractors and the keyed option of selected-response (SR) items (multiple-choice) provide to support or attenuate the validity argument for the use of selected-response items to assess listening comprehension in a second language?	3. Selected response items in the form of multiple-choice (MC) are a useful method to gather evidence of L2 listening comprehension. The keyed responses and distractors provide empirical evidence that supports the use of MC items to assess L2 listening comprehension.

## 4.1 Research Question One: TCF Listening Subskills

This section presents the results for the first research question. In turn, this validity evidence will be used to substantiate the interpretation/use argument for the TCF. The results are presented sequentially for both test forms (i.e., Form A and Form B) used in this study. This section addressed the following research question:

- 1) What listening sub-skills do the *Test de Connaissance du français* assess? Do test items contribute exclusively to the assessment of the listening construct? Is there supporting evidence indicating that test items are contaminated by construct irrelevant variance or construct underrepresentation?

### 4.1.1 Descriptive statistics

Table 10 below provides the descriptive statistics for different groups of examinees (e.g., gender, first language, age groups and location in Africa [North Africa: Algeria, Morocco and Tunisia and West Africa: Cameroon, Ivory Coast and Senegal]) who took Form A of the TCF. Except for the French and Arabic L1 subgroups, mean scores tend to be approximately equal across subgroups and the skewness and kurtosis values suggest that the score distribution for these groups approximate a normal distribution. On average, the native speakers of French obtained the highest score ( $M = 21.35$ ,  $SD = 4.98$ ) followed by the Arabic speaking subgroup ( $M = 17.63$ ,  $SD = 5.38$ ).

Table X Descriptive Statistics for Subgroups of Examinees in Form A

Test form	<i>n</i>	<i>M</i>	<i>SD</i>	Skewness	Kurtosis
Form A	1501	16.87	6.58	-0.02	-1.07
<b>Gender</b>					
Male	841	17.19	6.32	-0.091	-1.018
Female	660	16.46	6.87	0.069	-1.122
<b>L1<sup>18</sup></b>					
French	546	21.35	4.98	-0.562	0.024
Arabic	281	17.63	5.38	-0.197	-0.659
Spanish	184	10.52	4.55	1.109	1.288
Russian	104	10.36	3.94	1.187	1.994
Chinese	87	11.75	4.57	0.661	-0.339
Bulgarian	49	11.98	4.53	1.208	1.099

<sup>18</sup> For formatting purposes, this table only presents the first six largest L1 groups in Form A. Appendix A provides the complete list of L1 groups in both test forms.

Table X continued

<b>Age Groups</b>					
18-29	580	17.79	6.28	-0.095	-0.934
30-35	571	16.19	6.55	0.065	-1.136
36 +	344	16.45	6.89	0.018	-1.161
<b>Location<sup>19</sup></b>					
North Africa	266	18.82	5.051	-0.274	-0.632
West Africa	431	19.10	4.643	-0.364	-0.115

This impact on mean score difference is not surprising since native speakers of French would be expected to outperform second language speakers on a language test of general proficiency. The score difference between the Arabic speaking subgroup and the Spanish, Russian, Chinese, and Bulgarian subgroups can be tentatively explained by the prominence of French in Arabic speaking countries such as Algeria, Morocco and Tunisia.

The histogram on Figure 6 depicts the score distribution for Form A, suggesting a bimodal distribution possibly illustrating the pattern observed on Table 10 where native speakers of French and Arabic speaking examinees clearly outperformed the other subgroups.

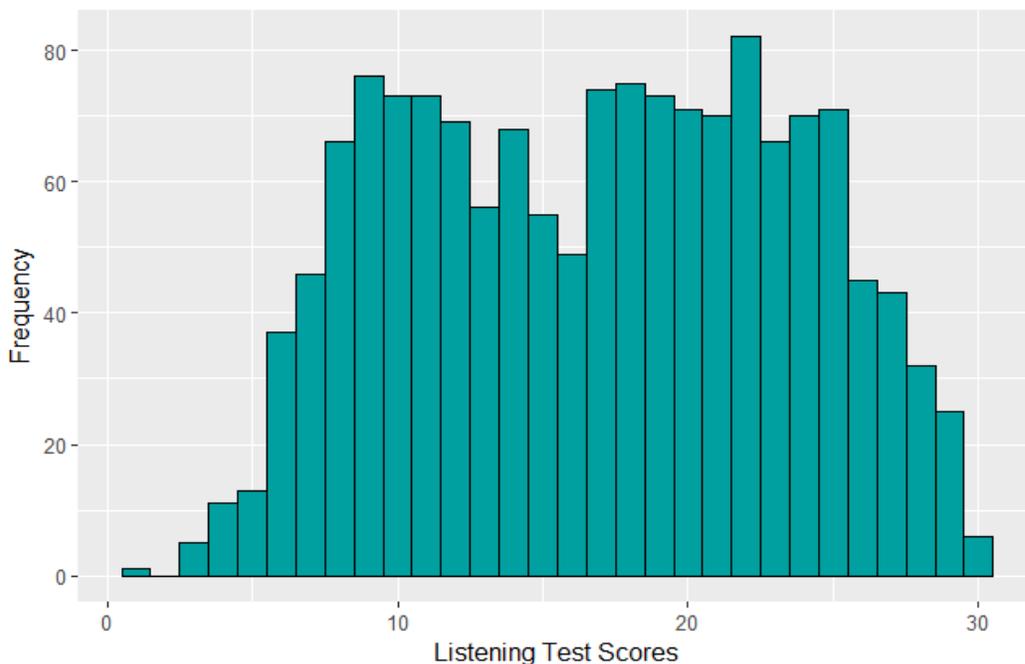


Figure 6 Score Distribution of Test Form A

<sup>19</sup> This distinction was made for DIF purposes and because of the influence of French in different countries across Africa, where Arabic (North Africa) and indigenous languages (West Africa) shape diverse ethnic groups.

The descriptive statistics for test form B on Table 11 also align with the pattern found on Table 10, that is, on average native speakers of French and Arabic speaking examinees seem to outperform the other subgroups. The native speakers of French ( $M = 20.77$ ,  $SD = 5.07$ ) and Arabic speaking ( $M = 16.60$ ,  $SD = 5.76$ ) obtained higher mean scores than the other subgroups. Similar to the examinees in Form A, the Arabic speaking subgroup included Moroccan, Egyptian, Tunisian and Algerian nationals where French plays an important role in various ways. The skewness and kurtosis values across all the subgroups presented on the table suggest that the distribution of scores approximates a normal distribution, except for the Spanish subgroup whose distribution appear a bit leptokurtic.

Table XI Descriptive Statistics for Subgroups of Examinees in Form B

<b>Test form</b>	<i>n</i>	<i>M</i>	<i>SD</i>	Skewness	Kurtosis
Form B	2118	16.17	6.18	0.21	-0.96
<b>Gender</b>					
Male	1054	16.79	6.16	0.067	-0.993
Female	891	15.87	6.35	0.337	-0.987
<b>L1<sup>20</sup></b>					
French	678	20.77	5.07	-0.439	-0.418
Arabic	427	16.60	5.76	0.038	-0.905
Spanish	322	11.80	3.81	0.968	2.144
Persian	215	10.93	3.39	0.408	0.683
Portuguese	97	14.29	4.37	0.253	-0.694
Chinese	96	12.68	4.40	0.894	0.633
<b>Age Groups</b>					
18-29	885	16.65	6.33	0.149	-1.046
30-35	809	15.72	5.99	0.297	-0.845
36 +	423	16.04	6.17	0.156	-0.918
<b>Location<sup>21</sup></b>					
North Africa	276	19.41	4.92	-0.317	-0.643
West Africa	521	18.92	4.47	-0.228	-0.214

The histogram on figure 7 below provides the score distribution for Form B. Unlike the histogram for Form A, the distribution appeared to be a bit skew to the right (the tail of the

<sup>20</sup> For formatting purposes, this table only presents the largest L1 groups in Form B. Appendix A provides the complete list of L1 groups in both test forms.

<sup>21</sup> This distinction is made because of the influence of French in different countries across Africa, where Arabic (North Africa) and indigenous languages (West Africa) shape diverse ethnic groups.

distribution runs from left to right as most of the scores are grouped on the left side of the histogram), indicating that Form B was more difficult than Form A. Further, the native speakers of French obtained a higher average score on Form A compared to the average score that native speakers of French obtained on Form B. Overall, examinees obtained slightly lower scores on Form B ( $M = 16.17$ ,  $SD = 6.18$ ) compare to examinees' overall performance on Form A ( $M = 16.87$ ,  $SD = 6.58$ ). Arguably, the apparent bi-modal distribution noted on Figure 6 could be explained by the difficulty level of Form A. In other words, since Form A was easier than Form B, native speakers of French did really well and their scores were a lot higher compared to non-native speakers, thus creating a bi-modal distribution. In contrast, Form B was more challenging for all examinees and test scores were not distributed as scores on Form A, creating a slightly negative skewed distribution.

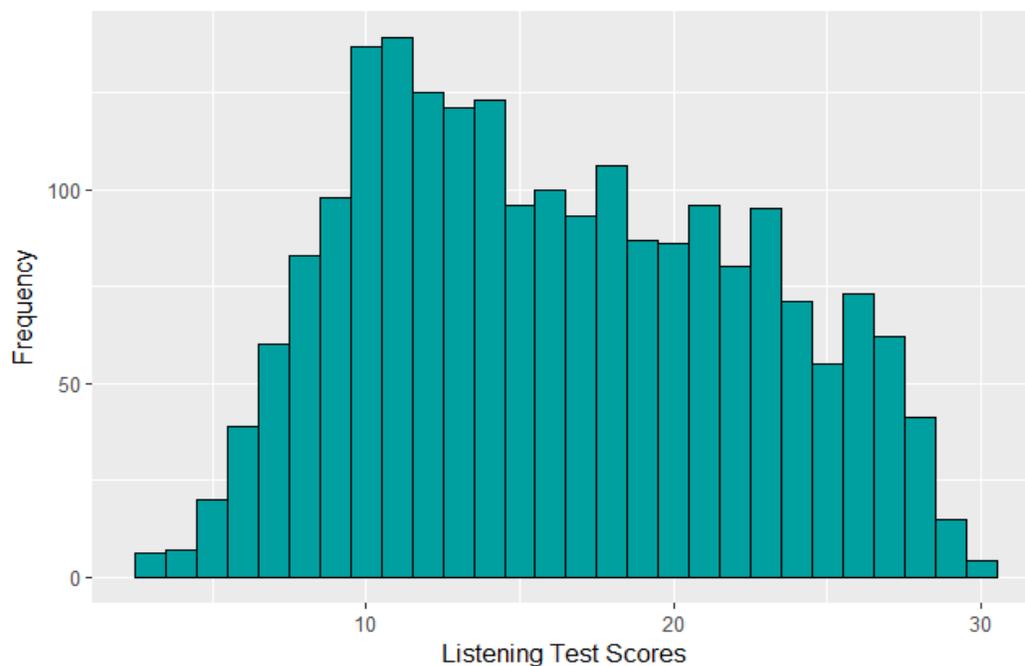


Figure 7 Score Distribution of Test Form B

The Kuder-Richardson 20 (KR-20) internal reliability indices for Form A and Form B were .879 and .865 respectively, suggesting that the items are internally consistent based on the data from the test forms. Tables 8 and 9 provide classical item analysis for both test forms. Item difficulty level seem to increase as the test unfolds and according to Ebel (1965) guidelines, some discrimination indices (Items 1, 5, 27, and 30 in Form A) suggest item revision due to low discrimination index ( $< .30$ )

Table XII Item Level Analysis for Form A

Item	Difficulty	<i>SD</i>	Discrimination	Item	Difficulty	<i>SD</i>	Discrimination
1	0.945	0.229	0.178	16	0.406	0.491	0.408
2	0.847	0.360	0.479	17	0.609	0.488	0.523
3	0.794	0.405	0.479	18	0.544	0.498	0.463
4	0.801	0.400	0.360	19	0.514	0.500	0.484
5	0.533	0.499	0.173	20	0.565	0.496	0.389
6	0.705	0.456	0.555	21	0.599	0.490	0.357
7	0.824	0.381	0.455	22	0.414	0.493	0.452
8	0.585	0.493	0.466	23	0.352	0.478	0.469
9	0.656	0.475	0.627	24	0.557	0.497	0.333
10	0.720	0.449	0.463	25	0.472	0.499	0.366
11	0.638	0.481	0.634	26	0.376	0.485	0.387
12	0.586	0.493	0.353	27	0.335	0.472	0.221
13	0.564	0.496	0.438	28	0.400	0.490	0.542
14	0.410	0.492	0.432	29	0.566	0.496	0.282
15	0.402	0.490	0.437	30	0.155	0.362	0.218

Results from Table 13 for Form B also show similar item difficulty patterns as in Form A where items become increasingly difficult as the test progresses. According to Ebel's (1965) guidelines for item discrimination, Items 2, 3, 22, and 26 through 30 should be revised.

Table XIII Item Level Analysis for Form B

Item	Difficulty	<i>SD</i>	Discrimination	Item	Difficulty	<i>SD</i>	Discrimination
1	0.862	0.345	0.332	16	0.406	0.491	0.334
2	0.583	0.493	0.256	17	0.314	0.464	0.341
3	0.687	0.464	0.197	18	0.477	0.500	0.355
4	0.683	0.466	0.460	19	0.593	0.492	0.523
5	0.722	0.448	0.429	20	0.424	0.494	0.413
6	0.806	0.396	0.459	21	0.501	0.500	0.438
7	0.821	0.384	0.413	22	0.224	0.417	0.264
8	0.720	0.449	0.301	23	0.468	0.499	0.404
9	0.734	0.442	0.387	24	0.484	0.500	0.560
10	0.569	0.495	0.318	25	0.589	0.492	0.513
11	0.655	0.476	0.477	26	0.371	0.483	0.259
12	0.730	0.444	0.473	27	0.336	0.473	0.285
13	0.632	0.482	0.545	28	0.260	0.439	0.234
14	0.551	0.498	0.538	29	0.290	0.454	0.251
15	0.540	0.499	0.575	30	0.145	0.352	0.094

Traditionally, it is common practice to check for multivariate normality before conducting confirmatory factor analysis, however, this procedure was not considered because the estimation

of the confirmatory factor analysis (CFA) models in this study does not proceed using normal-theory estimators like maximum likelihood. Rather, the specified CFA models employed the robust WLS estimator (WLSMV) based on the tetrachoric, sample covariance matrix. This estimator does not assume the distributional properties of data like normal theory estimators. And Robust WLS has shown to yield non-biased parameters when data are dichotomous and large sample sizes (e.g.,  $n > 500$ ) are used to specified CFA models (Moshagen & Musch, 2014).

#### 4.1.2 Item Coding

The review of the literature from second language listening (e.g., Buck, 2001; Field, 2013; Rost, 2011) suggested four major subskills or elements involved in listening comprehension processes. That is, (i) comprehending explicit aural discourse, (ii) recalling specific details, (iii) understanding the general topic or main idea, and (iv) inferring implicit ideas. In an initial iteration, the panel of experts was asked to associate each test item from the two test forms to one of the listening subskill identified in the literature. Then, in a second iteration the panel was asked how much they agreed (on a Likert type statement ranging from 1 = strongly disagree, 2 = disagree, 3 = agree, to 4 strongly agree) with each of the item-subskill associations. When there was disagreement among the panelists, discussions were held until reaching consensus, or until the panel agreed on a different subskill to be associated with the item in question. In the second iteration, it was important to ensure reasonable agreement to substantiate the specification of the CFA models. The panel encountered difficulty in agreeing when items made use of pictures. For example, a few experts considered these types of items to test both understanding of explicit aural discourse and inferring implicit ideas because of the interaction between the picture and the aural input. Importantly, the panel reached consensus when this type of issue arose by giving more weight to one skill over the other. Every expert rated each item from the two test forms used in this study. After consensus was reached, experts either agree or strongly agree with the item-subskills associations. For the test form A, most experts *strongly agreed* with the item-subskills associations. According to the panel of experts, Items 1, 2, 3, 4, 5, 6, 7, 9, 10, 11, 13, 14, 15, 16, 17, 18, 20, 21, 22, 23, 24, 26, 27, and 30 were associated with *comprehending explicit aural discourse* whereas Items 8, 12, 19, 25, 28, and 29 were associated with *inferring implicit ideas*. Table 14 below provides the distribution of experts by agreement category in Form A.

Table XIV Distribution of Experts and Agreement Category for Form A

Raters	Agree	Strongly Agree	Total
Expert A	4	26	30
Expert B	7	23	30
Expert C	2	28	30
Expert D	2	28	30
Expert E	2	28	30
Expert F	3	27	30
Expert G	4	26	30
Expert H	5	25	30
Expert I	5	25	30
Expert J	5	25	30
Expert K	6	24	30
Expert L	3	27	30
Average	4	26	30

Gwet's (2012) agreement coefficient ( $AC_1$ ) for multiple raters was calculated to determine expert agreement. As shown in table 15 below, the  $AC_1$  coefficient was statistically significant ( $AC_1 = 0.703$ ,  $p < 0.001$  95% CIs [0.584, 0.822]), suggesting that item-subskill associations were not due to randomness. This finding should be interpreted with caution given the width of the confidence intervals. In terms of interpretation of agreement coefficients, based on commonly used benchmark scales, expert agreement can be considered as substantial (Landis & Koch, 1977), good (Altman, 1991) and intermediate to good (Fleiss, 1971). Both Gwet's (2012) agreement coefficient and the benchmarking of agreement coefficients supported the CFA model suggested by the panel of experts.

Table XV  $AC_1$  Coefficient and Expert Agreement for Form A

Method	Coefficient	Inference/Subjects & Raters	
		SE	95% C.I.
Gwet's $AC_1$	0.703**	0.061	0.584 to 0.822
Percent Agreement	0.772**	0.036	0.701 to 0.843

Note.  $p < .001$ \*\*

For Form B, most experts mainly *agreed* with the item-subskills associations. After consensus was reached, experts either agree or strongly agree with the item-subskills associations where Items 1, 2, 3, 4, 5, 6, 7, 8, 9, 12, 13, 15, 16, 17, 21, 22, 28, 29, 30 were associated with *comprehending explicit aural discourse*, Items 10, 14, 23, 24, 25, and 26 were

associated with *inferring implicit ideas* and Items 11, 18, 19, 20 and 27 were associated with *understanding the general topic or main idea*. Table 16 below provides the distribution of experts by agreement category in Form B.

Table XVI Distribution of Experts and Agreement Category for Form B

Raters	Agree	Strongly Agree	Total
Expert A	22	8	30
Expert B	23	7	30
Expert C	27	3	30
Expert D	28	2	30
Expert E	25	5	30
Expert F	26	4	30
Expert G	21	9	30
Expert H	27	3	30
Expert I	25	5	30
Expert J	25	5	30
Expert K	26	4	30
Expert L	25	5	30
Average	25	5	30

Gwet’s (2012)  $AC_1$  for multiple raters was calculated to determine expert agreement. As shown in table 11 below, the  $AC_1$  coefficient was statistically significant ( $AC_1 = 0.613$ ,  $p < 0.001$  95% CIs [0.471, 0.755]), suggesting that item – subskill associations were not due to randomness. Again, this finding should be interpreted with caution given the width of the confidence intervals. In terms of interpretation of agreement coefficients, based on commonly used benchmark scales, the indices followed the same classification criteria as in Form A. Both Gwet’s (2012) agreement coefficient and the benchmarking of agreement coefficients supported the CFA model suggested by the panel of experts.

Table XVII  $AC_1$  Coefficient and Expert Agreement for Form B

Method	Coefficient	Inference/Subjects & Raters	
		SE	95% C.I.
Gwet’s $AC_1$	0.613**	0.073	0.471 to 0.755
Percent Agreement	0.720**	0.039	0.644 to 0.797

Note.  $p < .001$ \*\*

#### 4.1.2 Confirmatory Factor Analysis

Based on the item-subskill associations that were recommended by the panel of experts, a correlated two-factor model and a correlated three-factor model were specified for Form A and

Form B respectively to examine their reproducibility in the data. These models are presented in more detail in the next subsections. Additionally, two single factor models and two orthogonal models were specified as competing models. The single factor models were specified given the hypothesized high correlations among the listening comprehension subskills identified by the expert panel. The orthogonal models were specified to explore the separability of L2 listening subskills. The former models are congruent with the view that language skills are based on a rather unitary proficiency factor (Oller, 1983), the latter models are consistent with the view that language skills are separate linguistic entities (Carroll, 1983; Farhady, 1983). All specified models are briefly described below.

1. *A single-factor model for test forms A and B.* This model assumes unidimensionality of the listening items. As such, the existence of a single factor is hypothesized on which all measured variables load. As will be shown, this model was specified given the high intercorrelations between the two-factor models.

2. *A correlated two-factor model for Form A.* This model grouped the items under their respective listening subskills suggested by the panel. After careful examinations of content, the panel considered that only two listening subskills were assessed in this test form. That is, comprehending explicit aural discourse and inferring implicit ideas.

3. *An orthogonal two-factor model for Form A.* This model grouped all items under the respective listening subskill suggested by the panel. However, the correlation between the latent variables (i.e., comprehending explicit aural discourse and inferring implicit ideas) were specified to be zero to compare model fit against the correlated two-factor model. This model gave insights regarding the separability of the listening construct into smaller and self-standing constituents as operationalized in the TCF.

4. *A correlated three-factor model for Form B.* This model grouped the items under the respective listening sub-skills suggested by the panel. After careful examinations of content, the panel considered that three listening subskills were assessed in this test form. That is, comprehending explicit aural discourse, inferring implicit ideas, and demonstrating understanding of the general topic or main idea.

5. *An orthogonal three-factor model for Form B.* This model grouped all items under the respective listening subskill suggested by the panel. However, the correlation between the latent variables (i.e., comprehending explicit aural discourse, demonstrating understanding of the

general topic or main ideas and inferring implicit ideas) were specified to be zero to compare fit criteria against the correlated two-factor model. This model also gave insights regarding the separability of the listening construct into smaller and self-standing constituents as operationalized in the TCF.

All CFA analyses were conducted with Mplus 8 (Muthén & Muthén, 2017) and parameter estimation for item level data was performed using a robust weighted least squares (WLSMV) estimator with the diagonal weight matrix method. This estimator yields more stable parameters compared to the traditionally used maximum likelihood (ML) when data are dichotomous or binary (Moshagen & Musch, 2013). The hypothesized models were evaluated based on goodness-of-fit indices, model parsimony, and reasonableness of individual parameter estimates.

The overall goodness-of-fit statistics from the hypothesized CFA models for both test forms are presented in Table 18 below. Parsimony correction and comparative model fit indices (RMSEA, CFI and TLI) described in Chapter three were used to evaluate the global fit of the models to the data. Except for the orthogonal models, the global fit indices all exceeded the criterion for each index, suggesting a good model fit. In this regard, the results show that the single factor models and the correlated factor models from both test forms yielded the same fit indices.

Table XVIII Summary Results for CFA models in Form A and Form B Listening Sections

Test form	Model	<i>df</i>	$\chi^2$	RMSEA (90% CI)	CFI	TLI
Form A	Single factor	405	878.143**	0.028* (0.025, 0.030)	0.977	0.976
	Correlated two-factor	404	877.891**	0.028* (0.025, 0.030)	0.977	0.976
	Orthogonal two-factor	405	10627.223*	0.130* (0.128, 0.132)	0.510	0.473
Form B	Single factor	405	1226.154**	0.031* (0.029, 0.033)	0.965	0.962
	Correlated three-factor	402	1217.339**	0.031* (0.029, 0.033)	0.965	0.962
	Orthogonal three-factor	405	19431.615*	0.149* (0.147, 0.151)	0.183	0.122

Note.  $p < 0.001$ \*\*;  $p < 0.05$ \*

The single factor model and the two-factor model for Form A yielded good model fit indices (RMSEA = 0.028, CFI = 0.977, TLI = 0.976), whereas the orthogonal model yielded poor model fit (RMSEA = 0.130, CFI = 0.51, TLI = 0.473). Similar results were also obtained for Form B. Both the single factor and the correlated factor models yielded identical goodness-of-fit indices (RMSEA = 0.031, CFI = 0.965, TLI = 0.962) whereas the orthogonal model

showed a poor fit (RMSEA = 0.149, CFI = 0.183, TLI = 0.122). Fit indices for the single and correlated factor models specified in Form B were slightly different than Form A, but still suggested good model fit.

#### 4.1.2.1 The Correlated two-factor model for Form A

Figure 8 below depicts the complete specification of the correlated two-factor model for Form A. On the diagram, the squares represent the measured variables (i.e., the test items), the circles represent the latent or exogenous variables (i.e., the listening subskills), the curved bidirectional arrows provide the covariance between the latent variables, and the numeric values on the unidirectional arrows are the standardized factor loadings. Standard errors or unique variances of the estimates are given in parentheses. The standardized factor loadings are the regression slopes for predicting the indicators from the latent variables (Brown, 2015). For example, the standardized regression coefficient and measurement error for item 2 in form A are 0.788 and 0.018 respectively, and the regression equation describing the effect of the latent variable on performance on this item can be expressed as:  $y_1 = 0.788X + 0.018$ . The inter-factor correlations were high ( $r = .991$ ), suggesting a potentially single factor model. This echoes the similarity of model fit indices between the single factor and the correlated two-factor model. Following Thompson's (2004) guidelines for factor loading strength ( $\lambda > .400$ ), most factor loadings were strongly related to their purported factor except for Items 1, 5, 27, and 30 under the *explicit aural comprehension* factor ( $\lambda_1 = .337$ ,  $\lambda_5 = .240$ ,  $\lambda_{27} = .301$ , and  $\lambda_{30} = .338$ , respectively) and Item 29 ( $\lambda_{29} = .367$ ) under the *inferring implicit ideas* factor.

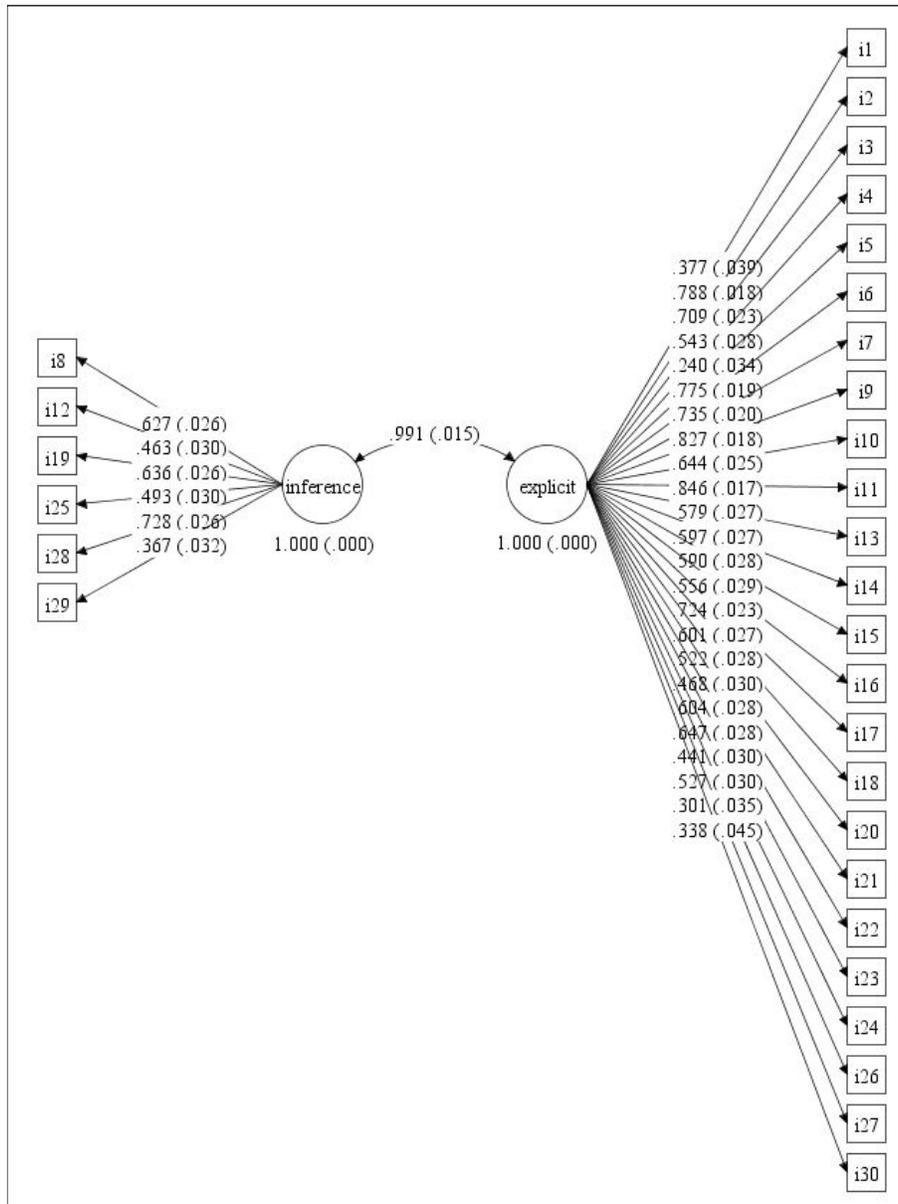


Figure 8 Correlated Two-Factor Model for Form A

Table 19 presents the standardized factor loadings, the standard errors, critical ratios,  $R^2$ , and residual variance for the correlated two-factor model for Form A. The latent variable accounted for a relative high variance in some items (e.g., item 9,  $R^2 = 0.684$ ), but relatively little variance in others (e.g., item 30,  $R^2 = 0.045$ ), potentially indicating that these items engage attributes which are distinct to those assessed by the TCF listening construct.

Table XIX Standardized Factor Loadings and Associated Critical Ratios, Explained and Residual Variances for Form A

Item	Standardized factor loadings	SE	Critical ratio	R <sup>2</sup>	Residual variance
<b>1</b>	<b>0.377</b>	0.039	9.585*	0.142	0.858
2	0.788	0.018	44.430*	0.621	0.379
3	0.709	0.023	31.298*	0.503	0.497
4	0.543	0.028	19.073*	0.295	0.705
<b>5</b>	<b>0.240</b>	0.034	7.060*	0.057	0.943
6	0.775	0.019	39.826*	0.600	0.400
7	0.735	0.020	36.092*	0.540	0.460
8	0.627	0.026	24.056*	0.394	0.606
9	0.827	0.018	47.146*	0.684	0.316
10	0.644	0.025	25.842*	0.414	0.586
11	0.846	0.017	49.114*	0.715	0.285
12	0.463	0.030	15.474*	0.214	0.786
13	0.579	0.027	21.705*	0.335	0.665
14	0.597	0.027	22.178*	0.357	0.643
15	0.590	0.028	20.838*	0.349	0.651
16	0.556	0.029	19.168*	0.310	0.690
17	0.724	0.023	32.164*	0.524	0.476
18	0.601	0.027	22.665*	0.361	0.639
19	0.636	0.026	24.322*	0.404	0.596
20	0.522	0.028	18.386*	0.273	0.727
21	0.468	0.030	15.817*	0.219	0.781
22	0.604	0.028	21.846*	0.365	0.635
23	0.647	0.028	22.759*	0.418	0.582
24	0.441	0.030	14.610*	0.195	0.805
25	0.493	0.030	16.549*	0.243	0.757
26	0.527	0.030	17.438*	0.278	0.722
<b>27</b>	<b>0.301</b>	0.035	8.663*	0.091	0.909
28	0.728	0.026	28.497*	0.529	0.471
<b>29</b>	<b>0.367</b>	0.032	11.579*	0.135	0.865
<b>30</b>	<b>0.338</b>	0.045	7.505*	0.114	0.886

Note.  $n = 1501$ ; the latent variable understanding explicit aural discourse accounted for a low amount of variance (items bolded in red). \* $p < 0.001$

Although there is no single standard for evaluating the magnitude of loading coefficients, Thompson (2004) proposed an absolute structure coefficient of  $\lambda > .400$  to indicate considerable indicator-factor correspondence. Following this rule of thumb, items that yielded weak factor loadings (i.e.,  $\lambda < .400$ ) were dropped and subsequent analyses were carried out. Items bolded in red in Table 19 were dropped and a modified, correlated two-factor model was specified. The chi-square difference between the initial and modified models could not be computed because

the models were not nested. In order to statistically compare non-nested models using Bayes information criterion (BIC), both models were fitted using the maximum likelihood estimator with robust standard errors (MLR<sup>22</sup>), which corrects for non-normality. As suggested by Schreiber, Nora, Stage, Barlow and King (2006) models that provide lower BIC values are preferred. As a result, the modified, correlated two-factor model yielded a better fit to the data (BIC = 42509.49, RMSEA = 0.027, [90% CI 0.024, 0.030], CFI = 0.984, TLI = 0.983) than the initial correlated two-factor model (BIC = 48169.31, RMSEA = 0.028 [90% CI 0.025, 0.030], CFI = 0.977; TLI = 0.976). The diagram on Figure 9 depicts the standardized factor loadings for the modified, correlated two-factor model for Form A.

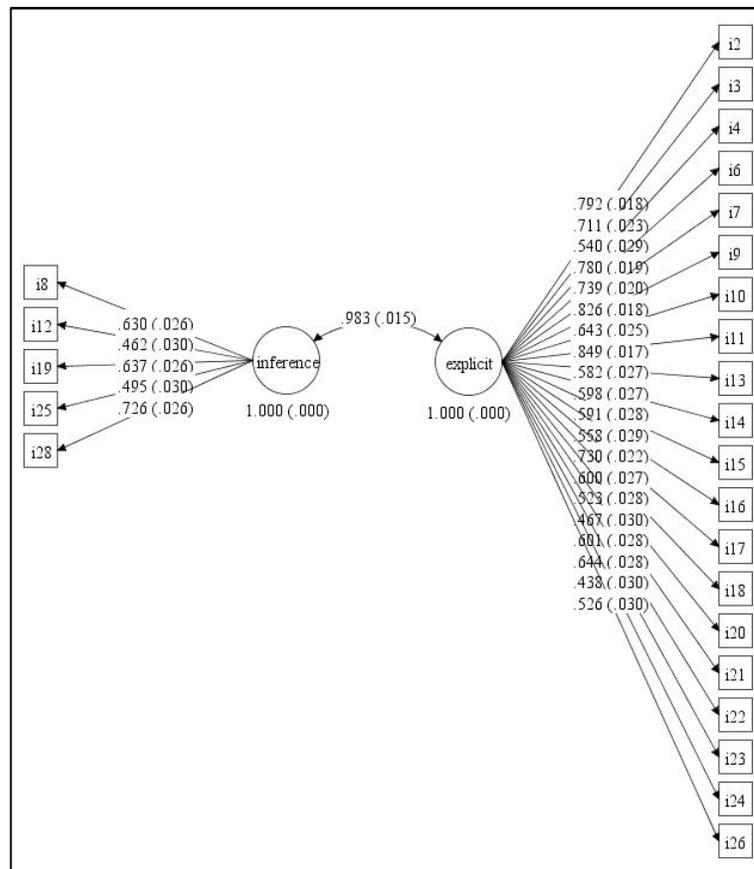


Figure 9 Modified Correlated Two-Factor Model for Form A

The inter-factor correlation was high ( $r = .983$ ) and this correlation along with the standardized factor loadings (ranging from .438 to .849), as well as the statistical significance

<sup>22</sup> The Mplus option of maximum likelihood estimation with robust standard errors was used only to compare the models. For the estimation of fit indices (e.g., RMSEA, CFI, and TLI) and parameters the weight robust least squares estimator was utilized.

of the loadings, suggested that the modified, correlated two-factor model, provided the best representation of the factor structure for Form A. One factor purported to assess the comprehension of explicit aural discourse (explicit) and the other assessed inferencing implicit ideas (inference).

#### 4.1.2.2 The Correlated two-factor model for Form B

The diagram in Figure 10 depicts the standardized factor loadings for each of the items the panel of experts associated with the listening subskills identified in Form B. The panel judged that this test form targeted three listening subskills: *comprehending explicit aural discourse, understanding the general topic or main idea, and inferring implicit ideas.*

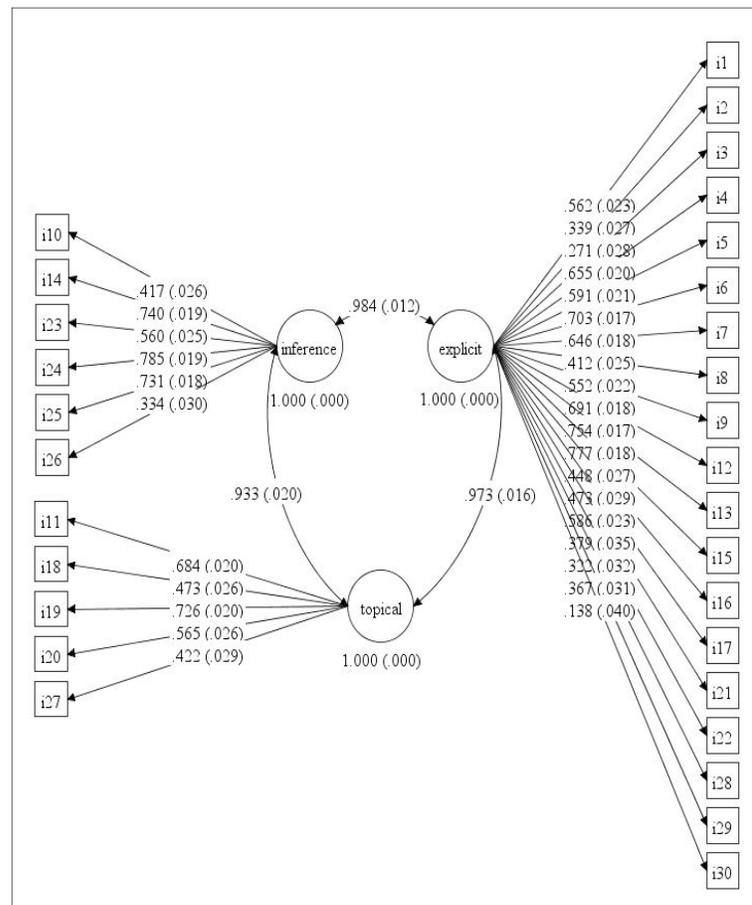


Figure 10 Modified Correlated Three-Factor Model for Form B

Again as in Form A, the standardized regression coefficient and measurement error for item 1 in form B are 0.562 and 0.023 respectively, and the regression equation describing the

effect of the latent variable on performance on this item can be expressed as  $y_1 = 0.562X + 0.023$ . The numeric values on the bidirectional arrows suggested a high covariance among latent variables ( $r = .933$ ,  $r = .973$  and  $r = .984$ ). To address the issue of high inter-factor correlations and to provide a more parsimonious explanation for the correlations among lower-order factors (Brown, 2015), a higher-order factor analysis model was specified. However, the latent variable covariance matrix was not positive definitive and indicated a correlation greater than one between the higher order factor (i.e., listening comprehension) and one of the lower-order factors, in this case, comprehension of explicit aural discourse. This suggested that the model was misspecified, rendering the results inadmissible (Brown, 2015) and the higher order model was discarded.

Table 20 below gives the standardized factor loadings, the standard errors, critical ratios,  $R^2$ , and residual variance for correlated three-factor model in Form B. As noted in Form A, the latent variable accounted for a relative high variance in some items (e.g., item 24,  $R^2 = 0.616$ ), but relative little in others (e.g., item 30,  $R^2 = 0.019$ ). This potentially indicates that low variance items engage attributes that are distinct to those assessed by the TCF listening construct. Consequently, a modified, correlated three-factor model was specified dropping the items that were unlikely to be accounted by the latent variable (e.g., items 2, 3, 22, 26, 28, 29, and 30), depicting a standardized factor loading less than 0.400.

Table XX Standardized Factor Loadings and Associated Critical Ratios, Explained and Residual Variances for Form B

Item	Standardized factor loadings	SE	Critical ratio	$R^2$	Residual variance
1	0.562	0.023	24.565	0.316	0.684
<b>2</b>	<b>0.339</b>	0.027	12.682	0.115	0.885
<b>3</b>	<b>0.271</b>	0.028	9.810	0.073	0.927
4	0.655	0.020	33.518	0.429	0.571
5	0.591	0.021	28.000	0.350	0.650
6	0.703	0.017	41.254	0.494	0.506
7	0.646	0.018	35.817	0.418	0.582
8	0.412	0.025	16.365	0.169	0.831
9	0.552	0.022	25.501	0.305	0.695
10	0.417	0.026	16.013	0.174	0.826
11	0.684	0.020	33.788	0.468	0.532
12	0.691	0.018	39.468	0.478	0.522
13	0.754	0.017	44.714	0.569	0.431
14	0.740	0.019	38.623	0.548	0.452

Table XX continued

15	0.777	0.018	44.377	0.604	0.396
16	0.448	0.027	16.390	0.201	0.799
17	0.473	0.029	16.257	0.224	0.776
18	0.473	0.026	17.885	0.224	0.776
19	0.726	0.020	36.030	0.527	0.473
20	0.565	0.026	22.003	0.319	0.681
21	0.586	0.023	25.278	0.343	0.657
<b>22</b>	<b>0.379</b>	0.035	10.957	0.144	0.856
23	0.560	0.025	22.653	0.313	0.687
24	0.785	0.019	40.880	0.616	0.384
25	0.731	0.018	39.539	0.534	0.466
<b>26</b>	<b>0.334</b>	0.030	11.160	0.112	0.888
27	0.422	0.029	14.322	0.178	0.822
<b>28</b>	<b>0.322</b>	0.032	9.957	0.104	0.896
<b>29</b>	<b>0.367</b>	0.031	11.892	0.135	0.865
<b>30</b>	<b>0.138</b>	0.040	3.430	0.019	0.981

Note.  $n = 2118$ ; the latent variable understanding explicit aural discourse accounted for a low amount of variance (items bolded in red).

In order to assess if the modified three-factor model provided a better fit to the data than the initial three-factor model, the same procedure used in Form A was followed. Based on the Bayes information criterion (BIC), the modified, correlated three-factor model yielded a better fit to the data (BIC = 52591.75, RMSEA = 0.029, [90% CI 0.026, 0.032], CFI = 0.981, TLI = 0.979) than the initial correlated three-factor model (BIC = 69032.64), RMSEA = 0.031 [90% CI = 0.029, 0.033], CFI = 0.965, TLI = 0.962). The diagram on Figure 11 on the next page depicts the standardized factor loadings for the modified, correlated three-factor model for Form B. The inter-factor correlations were high ( $r = .988$ ,  $r = .971$ ,  $r = .932$ ). These correlations along with the standardized factor loadings (estimates ranged from .406 to .786), and the statistical significance of the loadings, suggested that the modified, correlated three-factor model, provided the best representation of the factor structure of Form B. The panel of experts judged Form B tapped onto all but one (i.e., *recalling specific details*) of the subskills identified in the L2 listening literature. After items were dropped, the “Inference” and “Topical” factors remained with five indicators each and 13 indicators were kept in the “Explicit” factor. Items loaded on the respective specified factors recommended by the panel of experts: *comprehension of explicit aural discourse*, *inferencing implicit ideas*, and *understanding the general topic or main idea*.

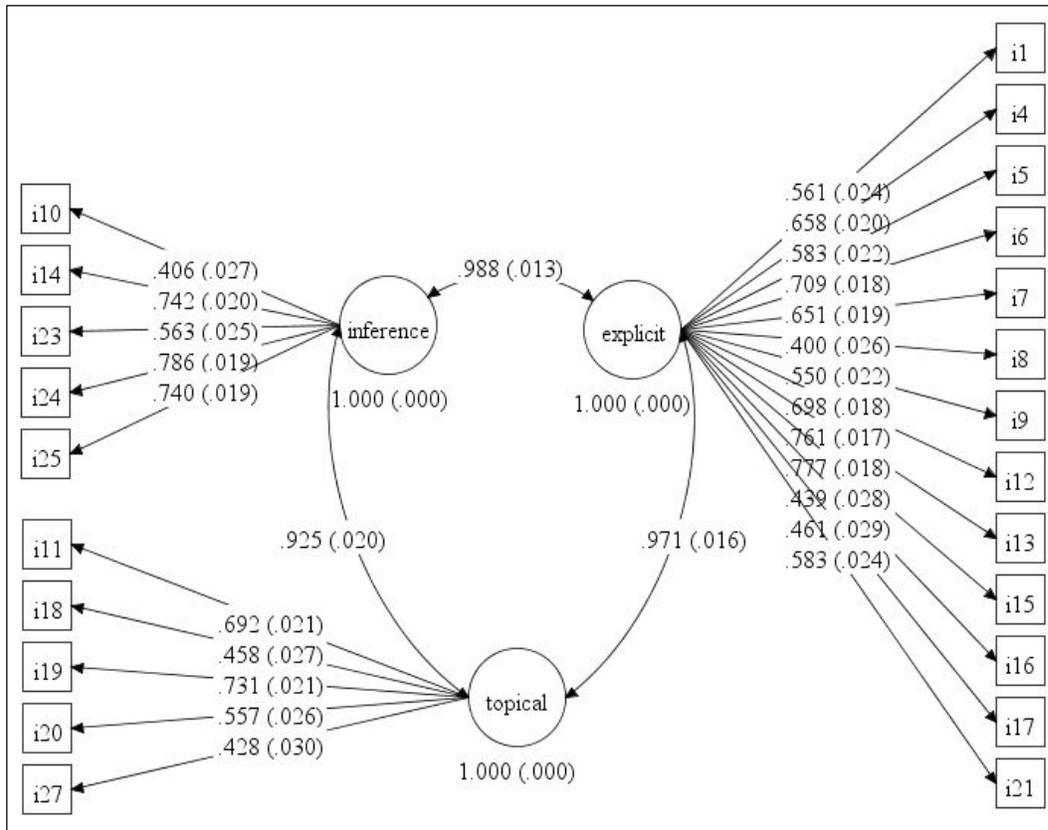


Figure 11 Modified Three-Factor Correlated Model for Form B

#### 4.1.2.3 Similarities and differences between the modified CFA models

Overall, the results from the confirmatory factor analyses for both test forms were consistent with the recommendations of the panel of experts. Of the models examined, the correlated CFA models were successfully reproduced in the sample variance - covariance matrix. The fit indices suggested that the panel recommendation could explain the interrelationships among items and the effect of the latent variables on the items. The correlated models were modified by dropping items with a factor loading inferior to 0.400 and the resulting modified correlated factor models (hereafter, the modified models) fit the data well and were also consistent with the literature in L2 listening comprehension.

Items in the modified models mostly targeted test takers' understanding of explicitly stated information in aural discourse and the panel of experts also considered that both test forms mostly contained items that elicited test takers' understanding of explicit information of aural discourse. To a lesser extent, both test forms exhibited items that targeted inferring implicit

ideas (5 items in each form). According to the panel of experts, Form B also comprised items that assessed the understanding of topical or main ideas. This was the main difference between the two modified models. This difference is further elaborated in the discussion chapter. In terms of the competing models, the fit of the orthogonal factor models to the data was poor and were discarded from further analysis. And although the single factor models fit the data well, conceptually, these models provided little insight about the construct of L2 listening comprehension operationalized in the TCF. Implications of these findings are also discussed further in Chapter five.

The following section addresses the second claim in the interpretation/use argument. That is, the TCF is a fair measure of L2 listening proficiency and does not favor group membership regardless of gender, first language, age or geographical location (only North and West Africa are included) of candidates who submit test scores to immigrate to Canada through the province of Quebec.

#### **4.2 Research Question Two: TCF Fairness**

This section provides the results for the fairness concerns of the TCF across different subgroups of interest in the target population. In turn, the validity evidence gathered in this study will be used to substantiate the interpretation/use argument for the test. The results are presented sequentially for both test forms and since the DIF study used the classical Rasch unidimensional model to estimate the item difficulties for each of the subgroups of interest, the fit of the data to the Rasch model is provided first. The fit of the data to the Rasch model was also helpful in identifying construct irrelevant variance at the item level and corroborated with the results stemming from the CFA analyses. This section addressed the following research question:

2) Does the *Test de Connaissance du français* exhibit differential item functioning (DIF), leading to test bias against gender, first language, age, and/or geographical location (only North and West Africa were included in the analysis) of examinees, threatening the validity of test score-based interpretation and use of the test in the context of Quebec immigration?

In the section that follows, misfitting items were identified using Wright and Linacre's (1994) recommended mean-square fit statistics guidelines for high stakes tests because adhering to Wu and Adams's (2013) sample sensitive criteria would have flagged too many items.

Misfitting items are then compared to the problematic items flagged in the CFA analysis and similarities as well as discrepancies between the two data modeling approaches are outlined.

#### 4.2.1 Rasch fit analyses

Tables 21 and 22 provide the item difficulty, mean-square statistics (infit and outfit) as well as the fit criteria consistent with sample size and the subskill associated with each of the items for both test forms. The column “difficulty” presents item measures in logits. In Form A, Items 1 and 2 were the easiest (-3.05 and -1.78 logits respectively) and item 30 was the most difficult (2.66 logits). In Form B (Table 22), Items 6 and 7 were the easiest (-1.55 and -1.70 logits respectively) and item 30 was the most difficult item (2.53 logits). When adhering to Wu and Adams’ (2013) recommendations to evaluate data-model fit, several large infit and outfit mean-square statistics can be observed in most items. This guideline to assess data-model fit at the item level is very conservative, but optimizes measurement precision as well as model integrity.

Table XXI Difficulty Measures, Fit Statistics, Fit Criteria and Listening Subskills for Form A

Item	<i>n</i>	Difficulty	Model SE	Infit MSQ	Outfit MSQ	Fit cut-offs		Subskill
1	1500	-3.05	0.12	1.03	0.91	0.93	1.07	explicit
2	1500	-1.78	0.08	0.81	0.50	0.93	1.07	explicit
3	1496	-1.35	0.07	0.85	0.69	0.93	1.07	explicit
4	1479	-1.48	0.07	0.98	0.86	0.93	1.07	explicit
5	1469	0.17	0.06	1.35	1.61	0.93	1.07	explicit
6	1486	-0.78	0.06	0.82	0.66	0.93	1.07	explicit
7	1490	-1.62	0.08	0.87	0.59	0.93	1.07	explicit
8	1491	-0.07	0.06	0.96	0.94	0.93	1.07	inference
9	1481	-0.49	0.06	0.75	0.62	0.93	1.07	explicit
10	1496	-0.84	0.07	0.92	0.84	0.93	1.07	explicit
11	1494	-0.36	0.06	0.73	0.65	0.93	1.07	explicit
12	1490	-0.07	0.06	1.12	1.13	0.93	1.07	inference
13	1490	0.05	0.06	1.01	0.96	0.93	1.07	explicit
14	1489	0.89	0.06	1.02	0.98	0.93	1.07	explicit
15	1498	0.94	0.06	1.00	1.02	0.93	1.07	explicit
16	1494	0.91	0.06	1.04	1.08	0.93	1.07	explicit
17	1495	-0.19	0.06	0.88	0.86	0.93	1.07	explicit
18	1489	0.15	0.06	0.98	0.94	0.93	1.07	explicit
19	1498	0.32	0.06	0.95	0.90	0.93	1.07	inference
20	1483	0.02	0.06	1.07	1.10	0.93	1.07	explicit
21	1494	-0.14	0.06	1.10	1.14	0.93	1.07	explicit
22	1489	0.86	0.06	0.98	0.99	0.93	1.07	explicit
23	1490	1.22	0.06	0.93	0.99	0.93	1.07	explicit

Table XXI continued

24	1493	0.08	0.06	1.13	1.16	0.93	1.07	explicit
25	1491	0.55	0.06	1.11	1.14	0.93	1.07	inference
26	1496	1.08	0.06	1.06	1.13	0.93	1.07	explicit
27	1478	1.31	0.06	1.26	1.44	0.93	1.07	explicit
28	1494	0.94	0.06	0.86	0.83	0.93	1.07	inference
29	1497	0.04	0.06	1.21	1.25	0.93	1.07	inference
30	1496	2.66	0.08	1.11	1.85	0.93	1.07	explicit

*Note.* Fit cut-offs are based on Wu and Adams (2013) recommendations for fit criteria; 304 missing data points (.7%) are distributed across the items and explains the variation of  $n$ .

In more practical terms, however, liberal guidelines would have not flagged most of these items as misfitting. For instance, Wright and Linacre (1994) suggested that for productive measurement mean-square fit statistics could range from 0.5 to 1.5 where values between 1.5 and 2 would be considered unproductive for the construction of measurement, but not degrading and values above 2 would degrade the measurement system. Adopting more liberal criteria would have not otherwise flagged these items as misfitting, but how much liberal is too good to be true? In this regard, Wright and Linacre (1994) also suggested that for selected response, high stakes tests, a reasonable range for mean-square statistics should be between 0.80 and 1.20 where fit values greater than 1.20 would distort measurement and exhibit too much variation (*underfit*) and fit values less than .80 would be too predictable (*overfit*), but would not degrade the measurement system.

Adhering to Wright and Linacre (1994) criteria for fit assessment in high stakes testing, large and underfitting outfit mean-square statistics were observed in Items 5, 27, 29 and 30 in Form A, and Items 2, 3, 22, 26, 27, 28, 29, and 30 in Form B, which represents a larger amount of problematic items for the latter test form. Only Items 5, 27, 29 in Form A and item 30 in Form B exhibited underfitting infit mean-square statistics. These items are also termed *noisy* items because they disrupt or distort the measurement system. Overfitting items (mean-square fit statistics < .80) such as Items 2, 3, 6, 7, 9, and 11 in Form A, and Items 1, 5, 6, 7, 11 through 15, and 19 in Form B were abundant across test forms, but are of less concern because the measurement procedure is not degraded by overfitting items. These items are also termed *muted* because they add redundancy to the measurement instrument.

Table XXII Difficulty Measures, Fit Statistics, Fit Criteria and Listening Subskills for Form B

Item	<i>n</i>	Difficulty	Model SE	Infit MSQ	Outfit MSQ	Fit cut-offs		Subskill
1	2107	-2.05	0.07	0.91	0.74	0.94	1.06	explicit
2	2097	-0.24	0.05	1.16	1.21	0.94	1.06	explicit
3	2084	-0.84	0.05	1.18	1.27	0.94	1.06	explicit
4	2094	-0.8	0.05	0.89	0.81	0.94	1.06	explicit
5	2100	-1.02	0.05	0.91	0.79	0.94	1.06	explicit
6	2111	-1.55	0.06	0.83	0.61	0.94	1.06	explicit
7	2102	-1.70	0.06	0.88	0.63	0.94	1.06	explicit
8	2110	-0.99	0.05	1.06	1.04	0.94	1.06	explicit
9	2116	-1.06	0.05	0.96	0.84	0.94	1.06	explicit
10	2100	-0.16	0.05	1.11	1.09	0.94	1.06	inference
11	2109	-0.61	0.05	0.89	0.77	0.94	1.06	topical
12	2112	-1.04	0.05	0.86	0.70	0.94	1.06	explicit
13	2103	-0.49	0.05	0.82	0.73	0.94	1.06	explicit
14	2112	-0.05	0.05	0.85	0.78	0.94	1.06	inference
15	2111	0.00	0.05	0.80	0.73	0.94	1.06	explicit
16	2096	0.7	0.05	1.09	1.16	0.94	1.06	explicit
17	2110	1.24	0.05	1.07	1.17	0.94	1.06	explicit
18	2113	0.34	0.05	1.07	1.07	0.94	1.06	topical
19	2108	-0.27	0.05	0.86	0.77	0.94	1.06	topical
20	2109	0.61	0.05	1.01	1.01	0.94	1.06	topical
21	2103	0.2	0.05	0.97	0.96	0.94	1.06	explicit
22	2106	1.85	0.06	1.12	1.47	0.94	1.06	explicit
23	2081	0.35	0.05	1.02	1.03	0.94	1.06	inference
24	2111	0.3	0.05	0.82	0.81	0.94	1.06	inference
25	2106	-0.25	0.05	0.86	0.81	0.94	1.06	inference
26	2109	0.91	0.05	1.18	1.30	0.94	1.06	inference
27	2110	1.11	0.05	1.18	1.25	0.94	1.06	topical
28	2110	1.59	0.06	1.19	1.41	0.94	1.06	explicit
29	2105	1.39	0.06	1.19	1.35	0.94	1.06	explicit
30	2108	2.53	0.07	1.26	2.18	0.94	1.06	explicit

*Note.* Fit cut-offs are based on Wu and Adams (2013) recommendations for fit criteria; 387 missing data points (.6%) are distributed across the items, this explains the variation of *n*.

To improve data-model fit, a sequential item deletion procedure would require dropping the noisiest items first, one at time, and then re-running the analysis and reexamine the fit statistics. This iterative process should stabilize mean-square statistics and potentially improve

data-model fit. This is common practice in Rasch measurement as the empirical data can never fit perfectly the Rasch model expectations (Bond & Fox, 2015). In order to draw valid inferences about persons' standing on the construct targeted in the TCF, misfitting items were dropped, until reasonable fit was achieved. In Form A, Items 30, 5, 27, and 29 were dropped sequentially. Again, the noisiest item was dropped first, the analysis was undertaken and the next noisiest item was dropped and so on. Three of the dropped items assessed the comprehension of explicit aural discourse and only one item targeted inferring implicit ideas. When the noisiest items in Form A were excluded from the analysis following Wright and Linacre's (1994) fit criteria for high stakes tests, Items 21, 24, 25 and 26 did no longer fit the model, but in order to maintain a reasonable degree of construct representation the item deletion iterations stopped at this stage. This decision is further elaborated in a section of the discussion chapter that elaborates on how CFA and Rasch measurement can complement each other in constructing measures. This deletion procedure yielded a set of calibrated items underlying the construct of listening comprehension in French as operationalized in Form A. A similar procedure was implemented for Form B where Items 30, 22, 28, 29, 26, 3 and 2 were dropped sequentially and excepting Item 26, which targeted inferring implicit ideas, the deleted items assessed the comprehension of explicit aural discourse. Note that Item 27, which targeted *understanding the general topic or main idea* did not fit the model, but was not dropped from the analysis to maintain a reasonable amount of items underlying this listening subskill within the test. Similar to the Rasch analysis in Form A, after the item deletion procedure other items appeared to misfit the model (Items 8, 10, 16, 17, 18 and 27), but were not dropped from the analysis. Again, this decision is further elaborated in the discussion chapter.

Tables 23 and 24 below provide the item difficulty, mean-square statistics (infit and outfit) as well as the fit criteria and the subskill associated with each of the items for both test forms after items were dropped. In general, the data from Form A provided a better fit to the unidimensional Rasch model after highly underfitting items were deleted and as noted in the first iteration of analyses, the overfitting or muted items were kept in both test forms because they do not represent a serious measurement problem.

Table XXIII Difficulty Measures, Fit Statistics, Fit Criteria and Subskills for Form A after Item Deletion

Item	<i>n</i>	Difficulty	Model SE	Infit MSQ	Outfit MSQ	Fit Cut-offs		Subskill
1	1500	-2.99	0.12	1.06	1.00	0.93	1.07	explicit
2	1500	-1.68	0.08	0.80	0.49	0.93	1.07	explicit
3	1496	-1.24	0.07	0.86	0.70	0.93	1.07	explicit
4	1479	-1.38	0.07	1.01	0.92	0.93	1.07	explicit
5				DELETED				explicit
6	1486	-0.65	0.07	0.81	0.65	0.93	1.07	explicit
7	1490	-1.53	0.08	0.87	0.60	0.93	1.07	explicit
8	1491	0.09	0.06	1.00	0.99	0.93	1.07	inference
9	1481	-0.35	0.06	0.76	0.62	0.93	1.07	explicit
10	1496	-0.72	0.07	0.95	0.88	0.93	1.07	explicit
11	1494	-0.21	0.06	0.75	0.66	0.93	1.07	explicit
12	1490	0.09	0.06	1.18	1.19	0.93	1.07	inference
13	1490	0.21	0.06	1.04	1.00	0.93	1.07	explicit
14	1489	1.1	0.06	1.07	1.04	0.93	1.07	explicit
15	1498	1.16	0.06	1.05	1.10	0.93	1.07	explicit
16	1494	1.12	0.06	1.10	1.17	0.93	1.07	explicit
17	1495	-0.04	0.06	0.89	0.90	0.93	1.07	explicit
18	1489	0.32	0.06	1.02	1.01	0.93	1.07	explicit
19	1498	0.5	0.06	1.00	0.95	0.93	1.07	inference
20	1483	0.19	0.06	1.12	1.17	0.93	1.07	explicit
21	1494	0.02	0.06	1.16	1.27	0.93	1.07	explicit
22	1489	1.07	0.06	1.03	1.06	0.93	1.07	explicit
23	1490	1.45	0.06	0.97	1.07	0.93	1.07	explicit
24	1493	0.25	0.06	1.21	1.32	0.93	1.07	explicit
25	1491	0.74	0.06	1.18	1.26	0.93	1.07	inference
26	1496	1.31	0.06	1.11	1.23	0.93	1.07	explicit
27				DELETED				explicit
28	1494	1.16	0.06	0.91	0.90	0.93	1.07	inference
29				DELETED				inference
30				DELETED				explicit

*Note.* Fit cut-offs are based on Wu and Adams (2013) recommendations for fit criteria; 304 missing data points (.7%) are distributed across the items this explains the variation of *n*.

The Rasch fit analyses flagged misfitting items that were also identified as problematic in the CFA analyses, however, some differences were also evident. For example, when adhering to guidelines suggested by Thompson (2004) to retain indicators in factor analysis and by Linacre (1992) to diagnose misfitting items in Rasch measurement, the CFA and the Rasch analyses for Form A flagged Items 5, 27, 29 and 30 as problematic. However, Item 1 in Form A fit the Rasch model while exhibiting a factor loading of 0.377, which was dropped in the purification stage following Thompson’s (2004) guidelines. In a similar vein, when adhering to the adopted guidelines, both CFA and Rasch analyses flagged the same items as problematic in Form B, but in this occasion the Rasch analysis also flagged an additional item (Item 27), which was not flag under the CFA analyses because of its moderate factor loading ( $\lambda = 0.422$ ). This is further elaborated in a section of the discussion chapter that summarizes how CFA and Rasch analyses can complement each other and how guidelines can impact decisions on item deletion.

Table XXIV Difficulty Measures, Fit Statistics, Fit Criteria and Subskills for Form B after Item Deletion

Item	<i>n</i>	Difficulty	Model SE	Infit MSQ	Outfit MSQ	Fit	Cut-offs	Subskill
1	2107	-1.85	0.07	0.93	0.78	0.94	1.06	explicit
2				DELETED				explicit
3				DELETED				explicit
4	2094	-0.53	0.05	0.92	0.87	0.94	1.06	explicit
5	2100	-0.76	0.06	0.97	0.89	0.94	1.06	explicit
6	2111	-1.33	0.06	0.84	0.62	0.94	1.06	explicit
7	2102	-1.48	0.06	0.9	0.64	0.94	1.06	explicit
8	2110	-0.73	0.06	1.14	1.27	0.94	1.06	explicit
9	2116	-0.81	0.06	1.01	0.96	0.94	1.06	explicit
10	2100	0.15	0.05	1.22	1.28	0.94	1.06	inference
11	2109	-0.33	0.05	0.92	0.8	0.94	1.06	topical
12	2112	-0.79	0.06	0.88	0.72	0.94	1.06	explicit
13	2103	-0.2	0.05	0.83	0.74	0.94	1.06	explicit
14	2112	0.28	0.05	0.88	0.84	0.94	1.06	inference
15	2111	0.34	0.05	0.83	0.76	0.94	1.06	explicit
16	2096	1.1	0.05	1.2	1.34	0.94	1.06	explicit
17	2110	1.7	0.06	1.18	1.43	0.94	1.06	explicit
18	2113	0.7	0.05	1.2	1.24	0.94	1.06	topical
19	2108	0.04	0.05	0.89	0.82	0.94	1.06	topical
20	2109	1	0.05	1.11	1.14	0.94	1.06	topical
21	2103	0.55	0.05	1.04	1.08	0.94	1.06	explicit

Table XXIV continued

22				DELETED					explicit
23	2081	0.71	0.05	1.1	1.12	0.94	1.06		inference
24	2111	0.65	0.05	0.85	0.83	0.94	1.06		inference
25	2106	0.05	0.05	0.89	0.84	0.94	1.06		inference
26				DELETED					inference
27	2110	1.55	0.06	1.29	1.48	0.94	1.06		topical
28				DELETED					explicit
29				DELETED					explicit
30				DELETED					explicit

*Note.* Fit cut-offs are based on Wu and Adams (2013) recommendations for fit criteria; 387 missing data points (.6%) are distributed across the items, this explains the variation of  $n$ .

In the sections that follow, the unidimensionality and local independence assumptions for both test forms are verified to ensure that the model assumptions are satisfied before the DIF analysis is undertaken.

#### 4.2.1 Assessing dimensionality and local independence

In the context of the classical Rasch unidimensional measurement model, both unidimensionality and local independence are central tenets for the integrity of the model. Thus, in order to draw valid inferences the data need to conform to the model expectation, otherwise, violations to the model assumptions can seriously affect the validity of the results. Unidimensionality refers to the assessment of a single construct (e.g., listening comprehension in French) and local independence refers to the relationship among the responses. That is, if local independence holds, a response given to one item should not influence subsequent responses to other items. The assumptions of unidimensionality and local independence are commonly verified through principal component analysis and inter-item correlations of standardized Rasch residuals respectively (Linacre, 1998b). Once the Rasch dimension is conditioned out or extracted from the data, there should not be any discernable patterns in the data and the PCA of standardized Rasch residuals should not yield a meaningful factor structure and the inter-item correlation should be nonexistent or negligible.

Although both unidimensionality and local independence are assumptions for the unidimensional Rasch model, the PCA and the inter-item correlations of standardized Rasch

residuals need to be conducted after fitting the data to the model because residuals are only generated after data modeling. The subsections that follow examine these assumptions in two steps: first all items are included in the analyses and in a second iteration, the dropped, misfitting items are excluded to explore the effect of misfitting items on dimensionality and local independence.

#### **4.2.1.1 Examining dimensionality through principal component analysis of standardized Rasch residuals**

Table 25 provides the PCA of the standardized Rasch residuals (PCAR) for Form A where lines three and four indicate that the primary Rasch dimension explained 33.10 % of the observed variance in the data, which is the sum of the variance explained by the persons (7.112 [15.90%]) and by items (7.704 [17.20%]). Initially this appeared to be a low amount of explained variance, but the Rasch model also predicts that there will be a random aspect of the data and this aspect of the randomness depends on the spread of the distributions of persons and items (Linacre, 2008). A wide spread of items and persons should explain most of the variance in a given measurement context. For example, when the person and item standard deviations are around 1 logit, only 25% of the variance is explained by the Rasch measures, but when the person and item standard deviations are around 5 logits, approximately 80% of the variance could be explained by the Rasch measures. Before item deletion in Form A, the standard deviations for both persons and items were 1.33 and 1.11 logits respectively (See Table 31) and this is coherent with the amount of variance explained by the Rasch measures in Form A.

To examine the dimensionality of Form A, the unexplained variance in the first contrast is an important number of reference. The first contrast ( $\lambda = 2.057$ ) suggests that a potential second dimension with the strength of at least two items is causing the noise. In other words, after the Rasch dimension was conditioned out of the data, a second dimension comprised of two items could be extracted from the remaining residuals. Although this is very weak to confirm a secondary dimension, E. V. Smith (2002) posited that an eigenvalue greater than 1.5 for the first contrast signified a violation of unidimensionality under the 500-person / 30-item condition in Rasch modeling. However, Chou and Wang (2010) suggested that having a fixed value for the first contrast could lead to erroneous decisions and in this regard in a simulation

study, Raïche (2005) concluded that the first contrast could be as large as two eigenvalues when the assumption of unidimensionality holds.

Table XXV Standardized Residual Variance in Eigenvalue Units for Form A

Variations	Eigenvalue	Observed Variance
Total raw variance in observations	44.815	100.00%
Raw variance explained by measures	14.816	33.10%
Raw variance explained by persons	7.112	15.90%
Raw Variance explained by items	7.704	17.20%
Raw unexplained variance (total)	30.00	66.90 %
Unexplained variance in 1st contrast	2.057	4.60%
Unexplained variance in 2nd contrast	1.443	3.20%
Unexplained variance in 3rd contrast	1.291	2.90%
Unexplained variance in 4th contrast	1.230	2.70%
Unexplained variance in 5th contrast	1.203	2.70%

Further investigation into the dimensionality of Form A examined the plot of the standardized residuals for the first contrast. Figure 12 depicts the loadings for the first contrast and each of the item clusters (i.e., cluster 1 cluster 2, and cluster 3, column furthest right). Of particular interest, the factor loadings for items “A”, “B”, and “C” and “a”, “b”, and “c” corresponding to Items 2, 7, 11, and 30, 29 27 respectively, exhibit the greatest absolute magnitude of loadings ( $\lambda > .48$  for A, B and C, and  $\lambda > .40$  for a, b and c). A closer examination of the content of these items revealed that except for item 27 (which assessed inferring implicit ideas), the remaining items assessed comprehension of explicit aural discourse. This suggests that this minor violation to dimensionality is not present in the data because of different strands (e.g., addition and subtraction), different dimensions (geography and history) or different item formats (e.g., pictorial monologues or short conversations). Note however that items 2, 7, 11 are relatively located at the beginning of the test, whereas items 27, 29 and 30 are located at the end of the test and if time constraint is an issue to complete the last items on the test, it could be hypothesized that speededness introduces some unwanted noise. This observation is revisited again in the Wright maps which depict the difficulty hierarchy of the items.

To gather more concluding evidence regarding the dimensionality of Form A, the correlation coefficients corrected for attenuation from the three clusters of items was examined. In Winsteps, each cluster constitutes a pseudo-test within the whole test. According to Joreskog

(1971), if the disattenuated correlation between two tests approaches unity, it can be concluded that the subtests are assessing the same trait. In this regard, the correlations corrected for attenuation between cluster one and three, and cluster one and two approached unity ( $r = .972$  and  $r = 1.00$  respectively), but the correlation corrected for attenuation between cluster two and three was less strong ( $r = .892$ ). Under cluster three, items a, b, and c (27, 29, and 30) were the last items on the test and given the nature of the testing instrument, a timed, second language listening test, it could be argued that time or pressure to answer the last items introduced noise to the data.

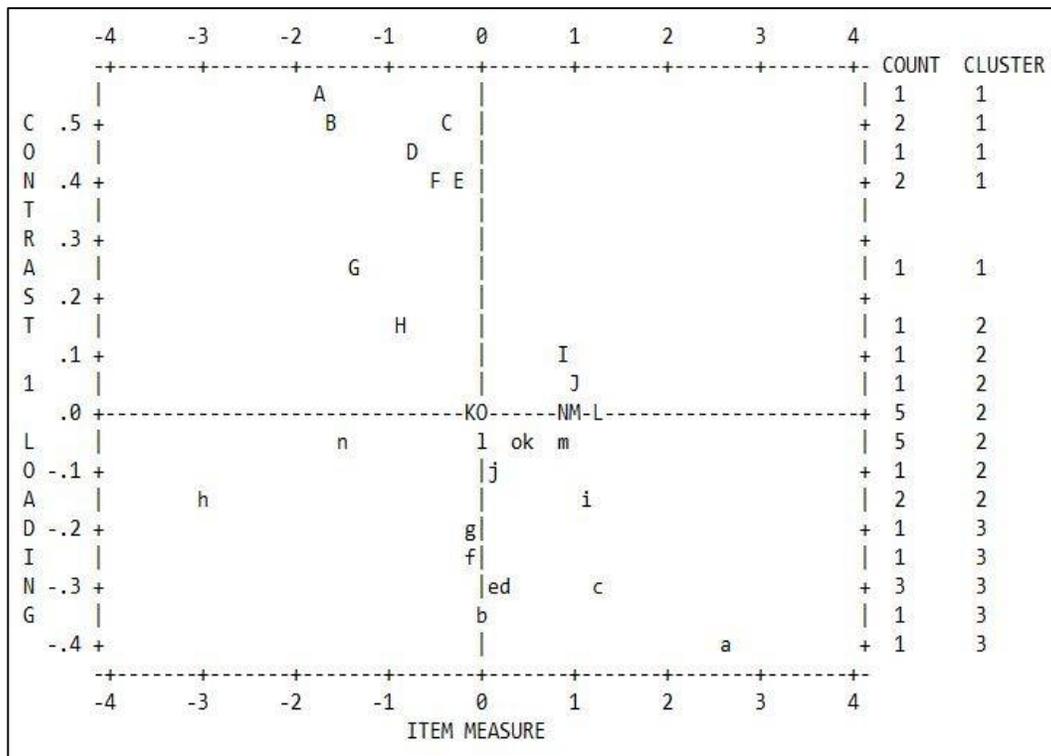


Figure 12 Standardized Residual Plot for the First Contrast in Form A

To gather more evidence regarding the dimensionality of Form A, underfitting items (5, 27, 29, and 30) were dropped and a new dimensionality assessment was carried out (See Table 26). In this iteration, the first contrast of the PCAR was inferior ( $\lambda = 1.78$ ) compared to the first contrast ( $\lambda = 2.057$ ) that included all the items, suggesting that the data were potentially unidimensional after the underfitting items were excluded from the analyses. In addition, the correlations corrected for attenuation between cluster one and three, and cluster one and two ( $r = 1.00$  and  $r = 1.00$  respectively), as well as the correlation corrected for attenuation between cluster two and three ( $r = .988$ ) improved, suggesting that the data were unidimensional.

Table XXVI Standardized Residual Variance in Eigenvalue Units for Form A without Noisy Items

Variates	Eigenvalue	Observed Variance
Total raw variance in observations	39.50	100.00%
Raw variance explained by measures	13.50	34.20%
Raw variance explained by persons	7.14	18.10%
Raw Variance explained by items	6.36	16.10%
Raw unexplained variance (total)	26.00	65.80 %
Unexplained variance in 1st contrast	1.78	4.5%
Unexplained variance in 2nd contrast	1.38	3.5%
Unexplained variance in 3rd contrast	1.24	3.1%

From a psychometric point of view, the results from the dimensionality assessment of Form A is consistent with the results from the CFA study because the modified CFA model also improved after problematic items were excluded from the analyses. However, this raises an important conundrum: how relevant is it to drop problematic items based on CFA and Rasch modelling procedures? This question is further elaborated in the discussion chapter.

The procedures used to examine the dimensionality of Form A were also applied to examine the dimensionality of Form B. That is, the eigenvalue of the first contrast from the standardized Rasch residual variance, the residual plot from the first contrast and the correlations corrected for attenuation among clusters were used to flag potential noise in the data. The dimensionality analysis in Form B yielded similar results as in Form A. For instance, the first contrast ( $\lambda = 2.169$ ) in Form B suggests a potential secondary dimension with the strength of two items. The sum of the variance explained by the persons (6.152 [15.90%]) and by the items (8.212[18.50%]) accounts for the amount explained by the Rasch dimension (32.40 %), which is similar to the variance explained in Form A.

Table XXVII Standardized Residual Variance in Eigenvalue Units for Form B

Variates	Eigenvalue	Observed Variance
Total raw variance in observations	44.369	100.00%
Raw variance explained by measures	14.369	32.40%
Raw variance explained by persons	6.157	13.90%
Raw Variance explained by items	8.212	18.50%
Raw unexplained variance (total)	30.000	67.60%
Unexplained variance in 1st contrast	2.169	4.90%

Table XXVII continued

Unexplained variance in 2nd contrast	1.596	3.60%
Unexplained variance in 3rd contrast	1.218	2.70%
Unexplained variance in 4th contrast	1.169	2.60%
Unexplained variance in 5th contrast	1.136	2.60%

In Figure 13, the residual plot of the first contrast flagged moderate factor loadings in items 12, 13 and 25, and items 26, 28, and 30 (A, B and C, and a, b and c, respectively), but these factor loadings were weaker than in Form A. The absolute magnitude of the factor loadings for these items was  $\geq .31$ , and according to the panel of experts, all items assessed comprehension of explicit aural discourse, except for items 25 and 26, which assessed inferring implicit ideas. Again, this suggests that this minor violation of dimensionality is not present in the data because of different strands of the construct, multidimensionality or different item formats.

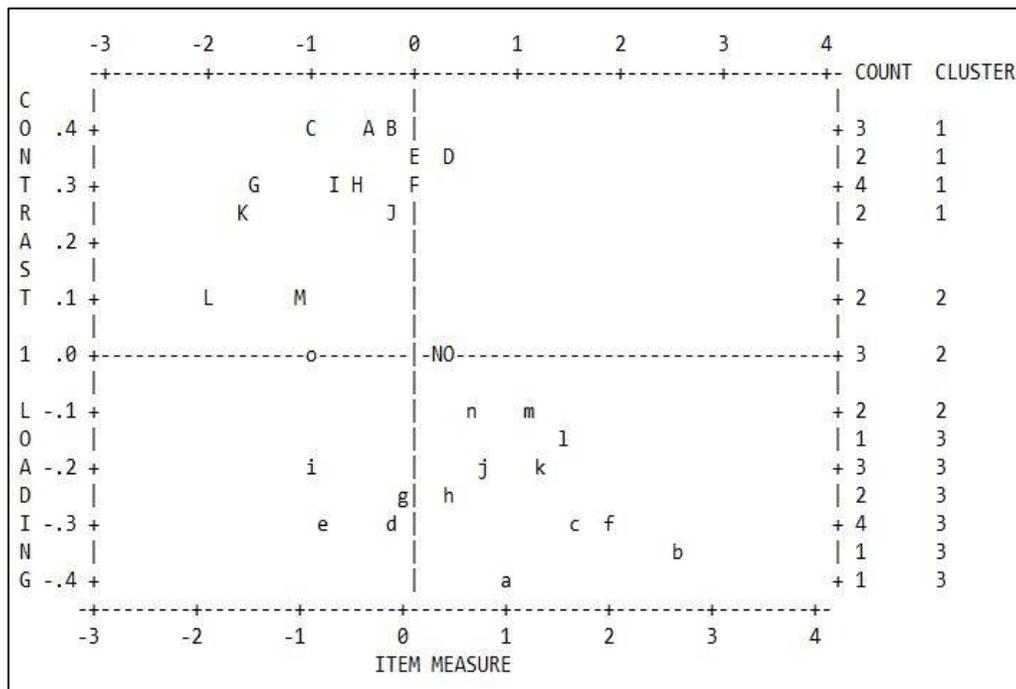


Figure 13 Standardized Residual Plot for the First Contrast in Form B

The correlations corrected for attenuation of the item clusters suggested similar trends as in Form A. The correlation corrected for attenuation for clusters one and two was unity ( $r = 1.00$ ), and for cluster two and three was strong ( $r = .943$ ). However, clusters one and three yielded a weaker correlation ( $r = .864$ ). Cluster three consisted of items 26, 28, and 30, three of

the last five items of the test. As in Form A, to gather evidence to support the unidimensionality assumption underlying the model, in a second iteration, the underfitting items were dropped from the dimensionality assessment. This time, the principal component analysis of standardized Rasch residuals for the first contrast yielded a weaker eigenvalue ( $\lambda = 1.66$ ) compared to the PCAR including all the items ( $\lambda = 2.169$ ), suggesting that the data were unidimensional after excluding noisy items.

Table XXVIII Standardized Residual Variance in Eigenvalue Units for Form B without Noisy Items

Variations	Eigenvalue	Observed Variance
Total raw variance in observations	33.98	100%
Raw variance explained by measures	10.98	32.3%
Raw variance explained by persons	5.56	16.4%
Raw Variance explained by items	5.42	16.0%
Raw unexplained variance (total)	23	67.7%
Unexplained variance in 1st contrast	1.66	4.9%
Unexplained variance in 2nd contrast	1.48	4.4%
Unexplained variance in 3rd contrast	1.18	3.5%
Unexplained variance in 4th contrast	1.12	3.3%

In addition, when the noisy items were dropped, the correlations corrected for attenuation between clusters one and three improved from  $r = .864$  to  $r = .992$  and the correlation between clusters one and two as well as clusters two and three improved to unity, suggesting that the data were unidimensional.

#### 4.2.1.2 Examining local independence

The assumption of local independence was verified through the inter-item correlations of standardized Rasch residuals. Once the Rasch dimension has been conditioned out or extracted, the correlation between any items should be zero, or very weak and non-significant (Baghaei, 2008). The table below provides the inter-item correlations for the five largest standardized Rasch residual coefficients in both test forms before dropping the underfitting items. From the five pairs of standardized Rasch residual correlations for Form A and Form B, the largest coefficients are weak ( $r = -.179$  and  $r = -.150$ , respectively). All residual correlations support the assumption that the test items are locally independent in both test forms.

Table XXIX Largest Residual Correlations for both Test Forms

Form A			Form B		
Correlation coefficient	Item	Item	Correlation coefficient	Item	Item
-.179	2	30	-.150	6	22
-.172	5	6	-.146	7	26
-.136	4	25	-.137	3	25
-.135	21	28	-.133	1	30
-.135	6	29	-.129	3	15

*Note.* Complete correlation matrices of standardized residuals are included in Appendix B

Table 30 provides the inter-item correlations for the five largest standardized Rasch residual coefficients in both test forms after the underfitting items were dropped and although item pairs were not the same as in the previous analyses, the correlation are still weak and the assumption of local independence still holds.

Table XXX Largest Residual Correlations for both Test Forms without Noisy Items

Form A			Form B		
Correlation coefficient	Item	Item	Correlation coefficient	Item	Item
.145	6	7	.106	1	6
.129	2	7	-.137	6	16
.101	2	11	-.134	10	25
-.147	2	12	-.133	18	24
-.133	21	28	-.127	7	16

*Note.* Complete correlation matrices of standardized residuals are included in Appendix B

The assessment of dimensionality and local independence for both test forms suggested that the data are suitable for Rasch analysis. Minor violations to the assumption of unidimensionality were identified in both test forms, but after dropping underfitting items these violations were considered negligible, given the size of the eigenvalue of the first contrast for both test forms. Further examination of the factor loadings of the first contrast and the correlations corrected for attenuation between item clusters suggested potential problems with the last items of the test. Again, after problematic items were dropped the correlation corrected for attenuation between all item clusters suggested that the data were unidimensional. It is hypothesized that the noise in the data potentially suggesting a second dimension with the strength of two test items in both test forms, might be the result of the nature of the test. That is,

a timed, second language listening test where the last items are responded under time pressure, which can lead to several Guttman errors in the data, thus affecting fit statistics in several ways. The underfitting items across test forms were excluded from further Rasch analysis because “the purposes of [classical] Rasch analysis are to maximize the homogeneity of the trait and to allow greater reduction of redundancy [or noise] at no sacrifice of measurement information by decreasing items to yield a more valid simple measure” (Granger, 2008, p. 1122). And this often times requires extracting from messy data measures that are homogenous, as a result, underfitting items are not considered in the Rasch-based differential item functioning. Once the misfitting items were dropped the data fit the model in Form A ( $\chi^2 = 36208.55$ ,  $df = 36293$ ,  $p = .622$ ) as well as in Form B ( $\chi^2 = 46767.55$ ,  $df = 46898$ ,  $p = .664$ ).

#### 4.2.2 Rasch modeling of the TCF listening Test

Figure 14 depicts the Wright maps for both test forms after underfitting items were dropped. Form A on the left and form B on the right display concurrently the persons and items along the listening continuum operationalized in the TCF. The measurement unit of the scales is in logits and are centered at zero. Each “#” on the left side of the Wright map for Form A represents 7 examinees and each dot “.” represents a range of one to six examinees. For Form B, each “#” represents eleven examinees and each dot “.” represents a range between one and ten examinees. Each item on both test forms is preceded by the letter “Q”. Test items range from easiest at the bottom and most difficult at the top of the Wright maps, and the examinees range from lowest scorers at the bottom to highest scorers at the top. On each side of the dividing dash lines, *M* represents the mean, *S* implies one standard deviation from the mean, and *T* represents two standard deviations from the mean. Comparison of the mean location logit scores of the examinees to the mean location logit scores of the items provides an indication of how well targeted or well matched the items are for the examinees in the sample. In both instances, on average, the samples were located at a higher level of L2 listening ability than the average difficulty of the scale. Although Form A appeared to be slightly easier than Form B, as noted in the descriptive statistics at the beginning of this chapter, after the underfitting items were dropped, both test forms exhibited similar level of difficulty. Test items cover a wide range of abilities from -2.99 logits (Item 1,  $SE = 0.12$ ) to +1.45 logits (Item 23,  $SE = 0.06$ ) for Form A, and from -1.85 logits (Item 1  $SE = 0.07$ ) to +1.70 logits (Item 17,  $SE = 0.06$ ) for Form B.

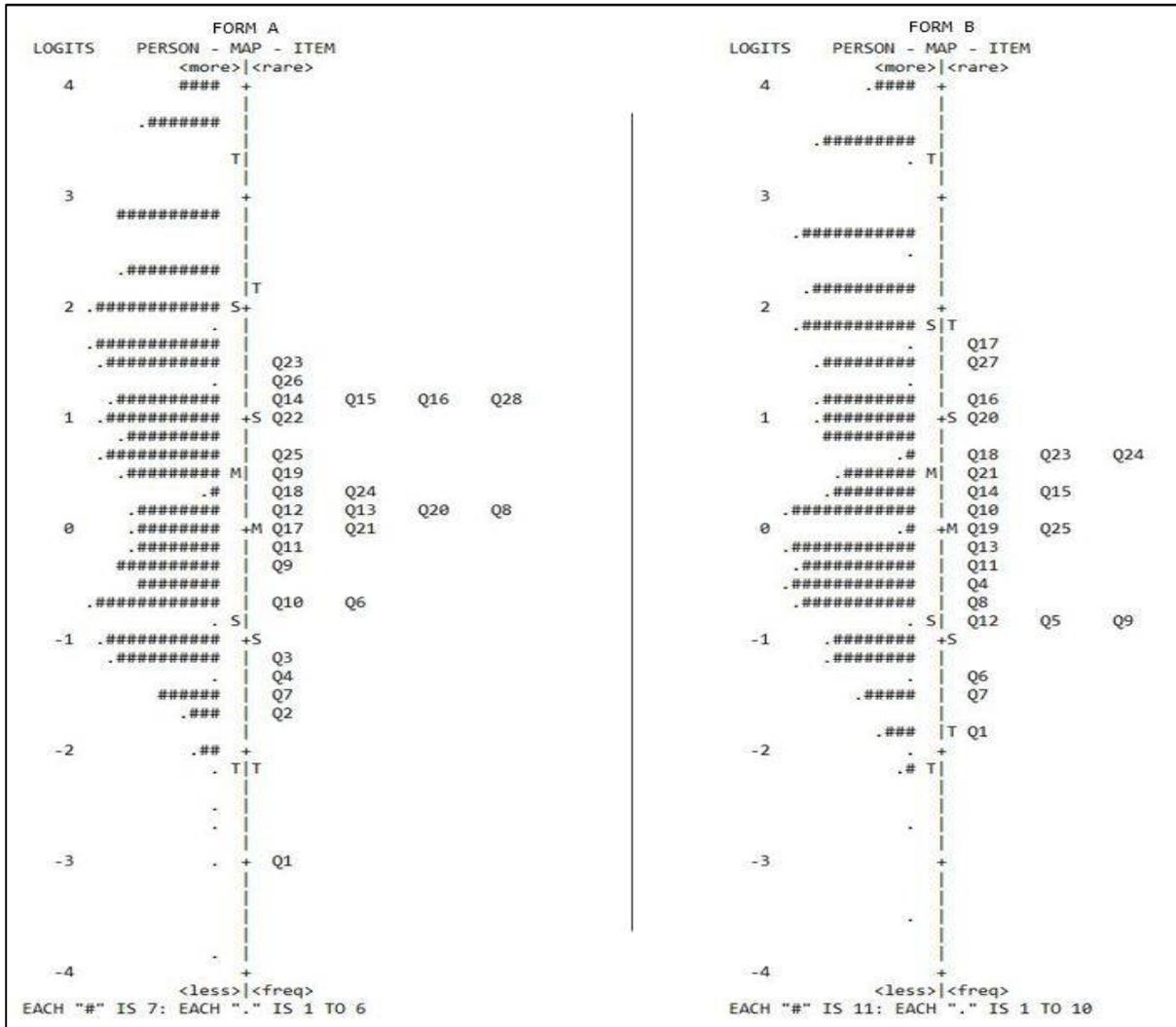


Figure 14 Wright Maps for both Test Forms

Most items clustered around the mean, where most of the test takers are located. The test takers with the highest and lowest ability levels do not have enough items corresponding to their ability levels. This produces larger errors of measurement at the top and at the bottom of the Wright maps due to poor targeting of extreme abilities. Ideally, items that are more difficult are needed to improve the targeting of the tests. In general, the targeting of both test forms is excellent. Furthermore, an interesting pattern occurring in both test forms is the difficulty hierarchy of test items where the most difficult items are the last items of the tests and the easiest items are generally the first items on the tests. This might be purposefully done by the test developer, but it could also be that the current allotted time is not enough to complete the last items on the test.

Table 31 below presents the Rasch summary statistics for both test forms before and after underfitting items were dropped. As noted earlier on the Wright maps, Form A and Form B exhibited a similar difficulty level after underfitting items were dropped (Form A: average person ability = 0.64 logits and Form B: average person ability = 0.60 logits).

Table XXXI Rasch-based Summary Statistics for both Test Forms after Item Deletion

Summary statistics	Before Item Deletion		After Item Deletion	
	Form A	Form B	Form A	Form B
Person Model <i>SE</i>	0.48	0.46	0.54	0.53
Person <i>M</i>	0.44	0.27	0.64	0.60
Person <i>SD</i>	1.33	1.24	1.49	1.39
Person separation Index	2.39	2.31	2.27	2.15
Item Model <i>SE</i>	0.07	0.05	0.07	0.05
Item <i>M</i>	0.00	0.00	0.00	0.00
Item <i>SD</i>	1.11	1.08	1.07	0.88
Item separation Index	16.21	19.35	15.34	15.76

*Note.* *M* = mean; *SD* = standard deviation; *SE* = standard error

In terms of the reliability for the positioning of persons and items, separation indices were used as large sample sizes and test length can inflate traditional reliability measures (Bond & Fox, 2015) and as recommended by Wright and Masters (2002), the number of person and item strata were calculated with the formula  $\frac{4G+1}{3}$  where G is the separation index for persons or items. Person separation indices and strata for both test forms (2.27 [strata = 9.41] and 2.15 [strata = 8.93] for Forms A and B respectively) suggested that the listening test could distinguish well between high and low performers, and as noted in the Wright maps this can be improved by developing more difficult items to better target the high performers. Item separation indices (15.34[strata = 61.69] and 15.76 [strata = 63.37] for Form A and B respectively) suggested that the item hierarchy in terms of difficulty is reliable. That is, if the listening tests were administered to similar samples of test takers, item difficulties should follow approximately the same hierarchy. The results from the Rasch analysis yielded insightful information regarding misfit diagnosis and dimensionality, in turn, this information supported the results from the CFA modeling, but also provided more evidence that should be considered when dropping items in test construction. This will be revisited in the discussion chapter.

Although fit statistics and differential item functioning (DIF) are usually independent, and item deletion has little impact on DIF, in this study, the underfitting items that were dropped in the previous Rasch analyses were not included in the DIF investigation. This decision aligns with the idea that assessment of model fit and refinement of the measurement system are always required when valid inferences underlying the construct are to be drawn (Sireci & Rios, 2013). This is consistent with Rasch measurement philosophy where data model fit is a necessary condition to draw valid inferences based on Rasch measures. In the sections that follow Rasch-based differential item functioning is investigated to identify potential issues that could attenuate or weaken claims regarding the fairness of the test.

### **4.2.3 Differential item functioning across gender groups**

The following sections report on the results of the measurement invariance of both test forms across several subgroups of interest in the targeted samples of examinees. More specifically, Rasch-based differential item functioning (DIF) was used to flag items exhibiting DIF across gender, first language (L1), age, and geographical location (for North and West Africa only). These stratifying variables are relevant for DIF investigation in the context of immigration since permanent resident applicants, with the same level of language ability, should have similar probabilities of success regardless of their gender, first language, age, or geographical location. Rasch-based DIF was used with relatively large sample sizes per group ( $n \geq 149$ ) to warrant the stability of difficulty parameters (Sireci & Rios, 2013; Zumbo, 1999). Due to the large amount of DIF graphics and tables produced across the subgroups under examination (i.e., gender, L1, age, and geographical location), only the graphics that depicted statistical significant DIF (i.e., absolute DIF contrast  $\geq 0.43$  logits) are discussed in the DIF sections. For illustration purposes a few graphics exhibiting no DIF are also included but all DIF graphics and tables are included in Appendix B. Items with large DIF were deleted sequentially to verify if these items were causing artificial DIF in others and when DIF was still present, deleted items were reinstated in the analysis.

Rasch-based DIF across gender was investigated as the groups consisted of relative large sample sizes for both test forms (males:  $n = 841$ ; females:  $n = 660$  for Form A and males:  $n = 1,148$ ; females:  $n = 970$  for Form B). Figure 15 depicts the item characteristic curves for items 4, 8, 16 and 23 in Form A. Items 4 and 8 do not exhibit differential item functioning since the

empirical item characteristic curves for both males and females are almost identical, depicting similar probabilities of correct response given a level of theta ( $\theta$ , L2 listening ability). Males were set as the reference group and an absolute DIF contrast exceeding 0.43 logits (i.e., 1 Delta unit) was considered as an indication of statistical significant DIF (Zwick, Thayer & Lewis, 1999).

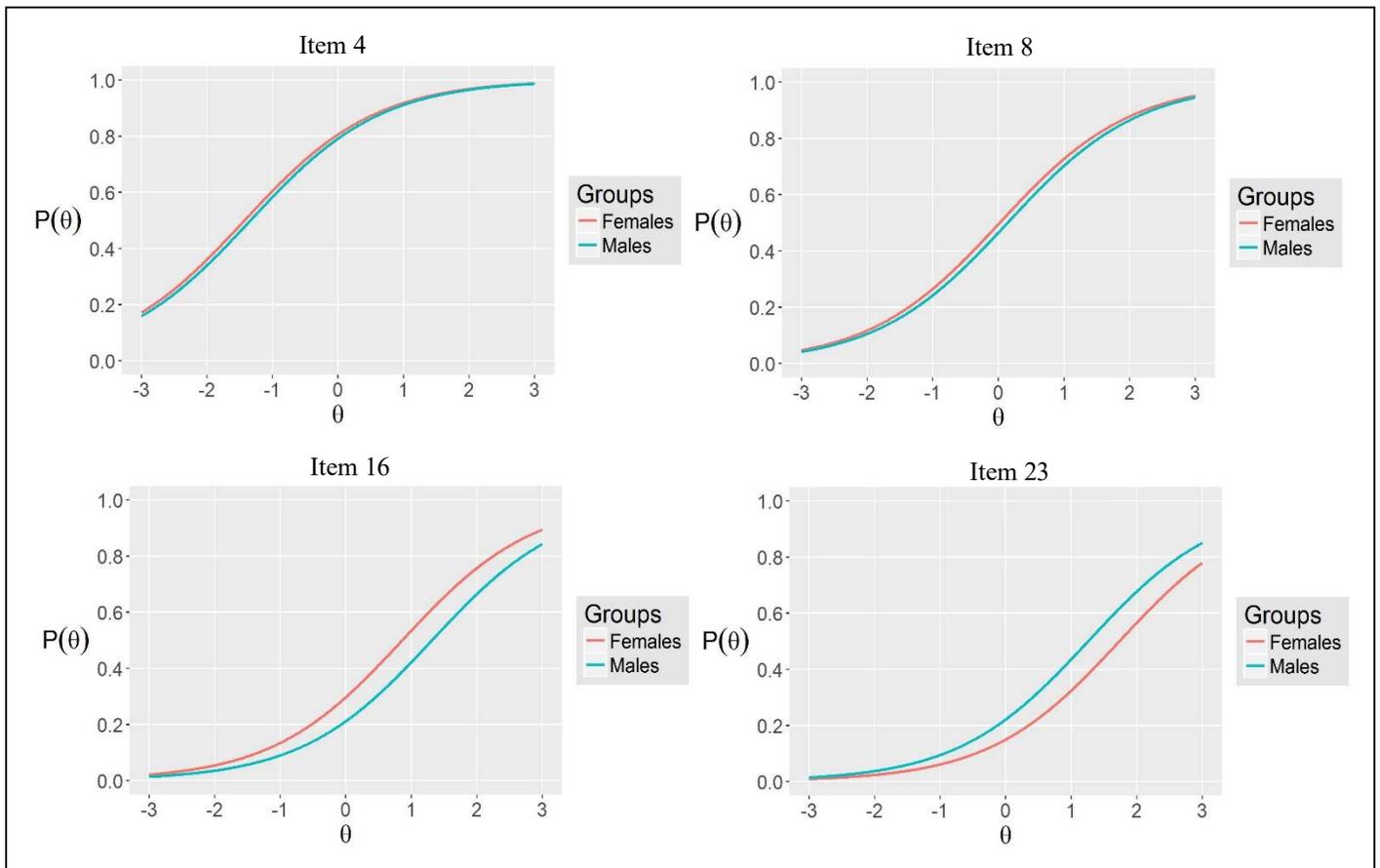


Figure 15 Item Characteristic Curves for Items 4, 8, 16 and 23 in Form A

A negative DIF contrast indicated that the item favored the reference group (i.e., males) and positive DIF contrast indicated that the item favored the focal group (i.e., females). Item 16 ( $\delta_{\text{males}} = 1.32 - \delta_{\text{females}} = .86$ ,  $\text{DIF}_{\text{contrast}} = .46$ ,  $p < .05$ ) and Item 23 ( $\delta_{\text{males}} = 1.26 - \delta_{\text{females}} = .1.74$ ,  $\text{DIF}_{\text{contrast}} = -.48$ ,  $p < .05$ ) exhibited slight to moderate differential item functioning, where item 16 was easier for the focal group (i.e., females) and item 23 was easier for the reference group (i.e., males). The sample-based effect size can be readily calculated with the formula:  $\frac{\text{DIF}_{\text{contrast}}}{\text{Person sample measure SD}}$ , where the numerator is the difference between the DIF

measures (e.g., the reference and focal groups), and the denominator is the standard deviation for the sample measure. Adhering to Cohen’s (1988) guidelines of effect sizes for independent group designs, the effect size for Items 16 and 23 is small ( $d = .308$ ,  $d = .322$ , respectively). Items 16 and 23 were dropped sequentially and subsequent analyses suggested that the remaining items did not exhibit substantive DIF.

Each item flagged as having a substantial level of DIF (i.e., Items 16 and 23) was subjected to a content analysis. The content of Item 16 consisted of a short dialogue between a man and a woman where the female voice was predominant in the dialogue and also uttered the stem of the question. In this regard, Kramer (1977) found gender differences in speech perception where men perceived trait characteristics of female speech as fast and with a wide range in rate and pitch, perhaps this could have introduced some noise in the item, favoring females. For Item 23, the pattern was reversed while a male voice was predominant in an aural exchange between two interlocutors. Note however that these hypotheses require further investigation. Form B did not exhibit substantive gender DIF as the largest DIF was only flagged in Items 9 and 17 and this was negligible ( $DIF_{contrast} = .38$ ,  $p < .05$ ;  $DIF_{contrast} = .39$ ,  $p < .05$ , respectively).

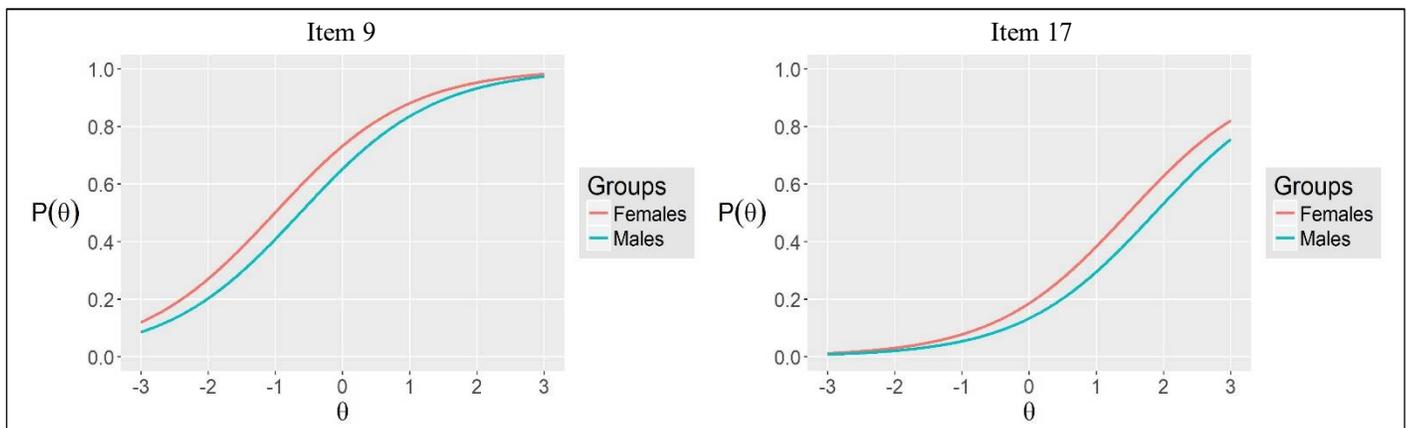


Figure 16 Item Characteristic Curves for Items 9 and 17 in Form B

#### 4.2.4 Differential item functioning across first language

Differential item functioning across examinees’ first language (L1) was also examined and the subgroups of interest for the Rasch-based L1 DIF included: Arabic ( $n = 281$ ), French ( $n = 546$ ), and Spanish ( $n = 184$ ) for Form A, as well as Arabic ( $n = 427$ ), French ( $n = 678$ ), Persian ( $n = 215$ ), and Spanish ( $n = 322$ ) for Form B.

Figure 17 below depicts the empirical item characteristics curves for Items 1, 2, 3, 4, 6, and 7 in Form A with first language as the grouping variable. Based on the results, most of these items exhibited DIF magnitudes that required attention. For example, in Item 1, Spanish, low ability examinees ( $\theta \leq 0$  logits), had a higher probability of answering the item correctly than the Arabic and French groups. This is at odds with the reference group (French) as French is their mother tongue and are expected to outperform nonnative speakers of French on a test of functional proficiency. In addition, some Arabic speaking subgroups (e.g., Morocco and Tunisia) use French as a lingua franca or have had French influence in the past and yet had a lower probability of correct response than the Spanish examinees at lower levels of ability. It is probable that low ability examinees whose first language was Spanish relied on guessing to answer this item. With regards to the French group, inattention could have affected their performance, especially since native speakers of the language of the test are not accustomed to take these type of assessments and items that are too easy might trigger inattention. The item consisted of listening to four aural utterances to match the correct utterance with a picture (picture recognition item). A content analysis of the item did not reveal a clear pattern as of why the item was biased against the low ability French and Arabic groups. This situation is somewhat similar to the groups' performance on Item 4 where the item was easier for the Spanish speaking examinees ( $\delta = -1.78$ ) compared to the Arabic ( $\delta = -1.33$ ) and French ( $\delta = -0.88$ ) candidates.

Items 2, 3, 6 and 7 depicted a more coherent pattern since Spanish speaking examinees had a lesser probability of correct response than the Arabic and French groups. Given the hypothesized higher language ability for both French and Arabic examinees over Spanish speaking candidates, it stands to reason that this difference in performance can be potentially attributed to impact rather than bias. In this regard, it is expected that Arabic examinees whose second language is likely to be French, and French examinees, whose first language is the targeted language of the assessment, have a higher rate of success than Spanish speaking candidates. This hypothesis is particularly heightened in Item 7 where Arabic and French examinees have similar probabilities of correct response at all levels of theta (i.e., listening ability). A content analysis of these items did not flag any potential bias against Spanish speaking examinees.

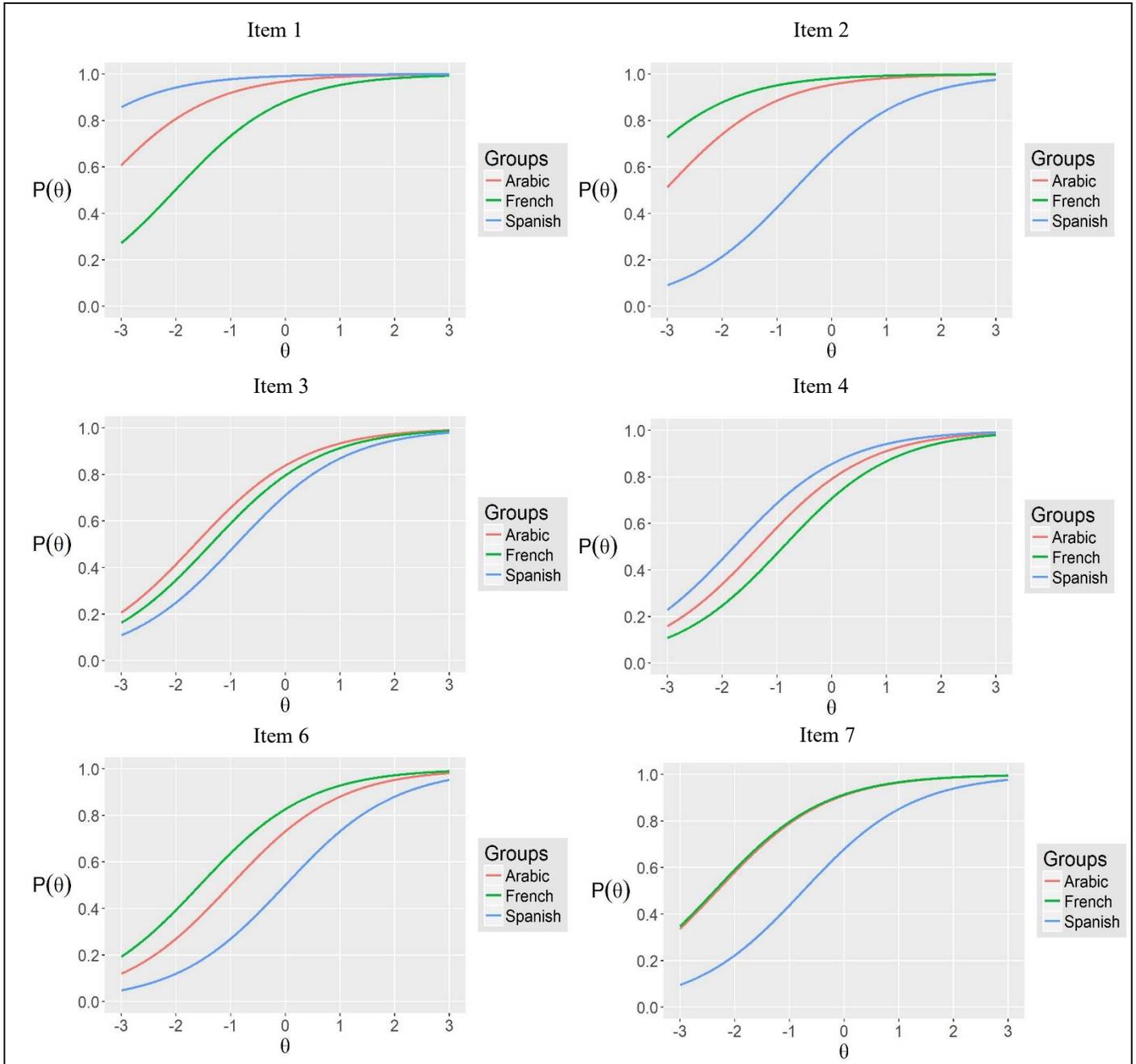


Figure 17 Item Characteristic Curves for Items 1, 2, 3, 4, 6 and 7 in Form A

Figure 18 below depicts the empirical item characteristics curves for Items 8 through 13. From this set items, Item 10 is the only item that did not exhibit differential item functioning. Items 9 and 11 depict the idea of impact attributed to the expected weaker performance of Spanish examinees on a French assessment when compared to French and Arabic examinees where French is the mother tongue of the former group, and the latter group speaks French as either lingua franca or a second language.

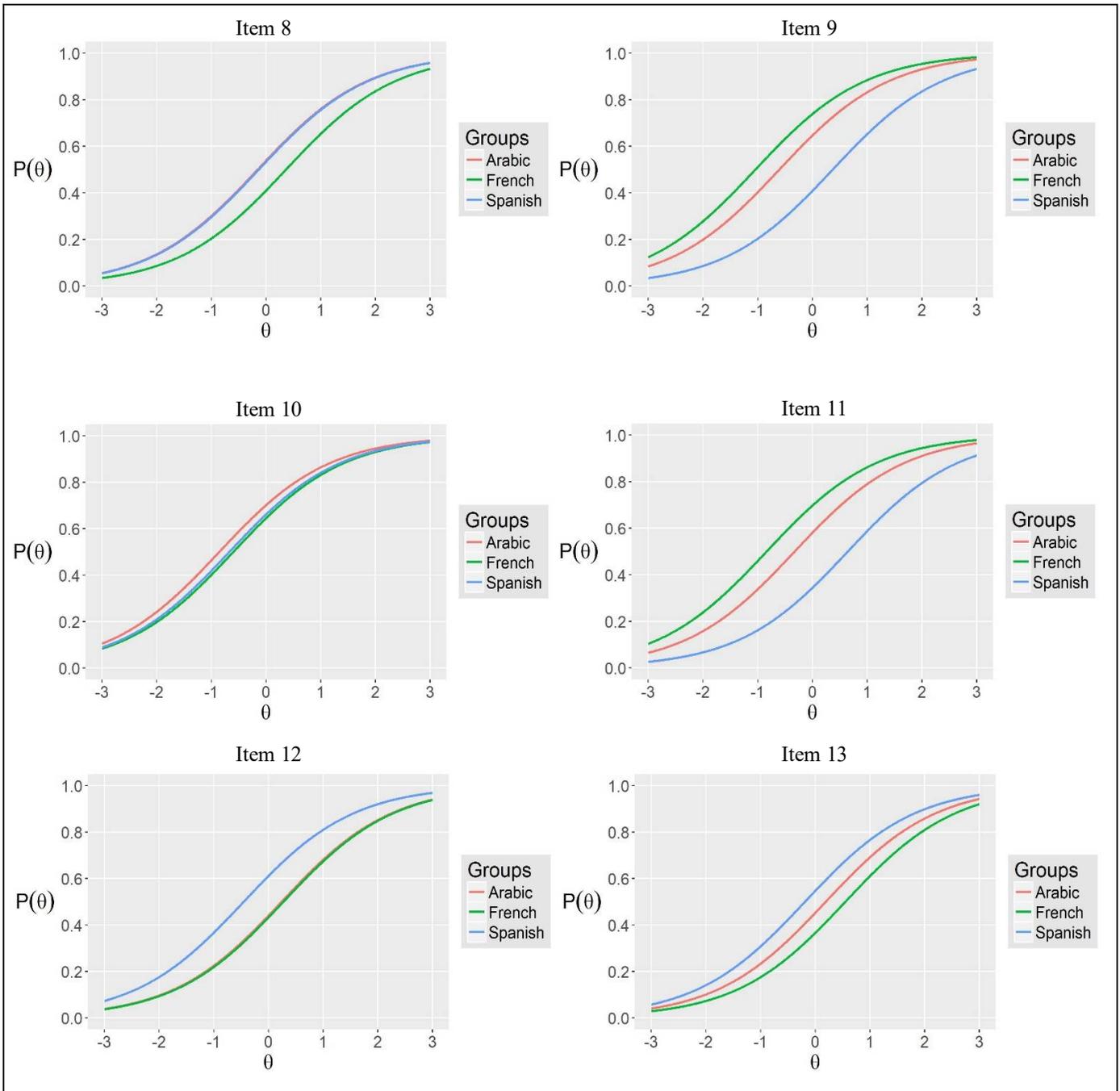


Figure 18 Item Characteristic Curves for Items 8 through 13 in Form A

Items 8, 12 and 13 provided a pattern worth paying attention because again the group whose first language is Spanish obtained a higher probability of correct response compared to the native speakers of French. On Item 8, the difficulty level for the subgroup whose first language was Arabic ( $\delta = -0.15$ ) was similar to the Spanish speaking subgroup ( $\delta = -0.13$ ),

thus obtaining approximately the same probability of correct response across all levels of theta (i.e., listening ability). This trend is similar in Item 13 where the absolute DIF contrast (0.38) suggested negligible DIF between the Spanish and Arabic speaking examinees. On Item 12, the difficulty for French and Arabic speaking examinees ( $\delta = 0.27$  and  $\delta = 0.24$ , respectively) suggested a similar probability of correct response across all levels of theta. Thus, on the graphic, both item characteristic curves for the French and Arabic speaking subgroups overlap. Figure 19 below depicts the empirical item characteristics curves for Items 14 through 19 for Form A. From this set of items, item 14 depicts a slight to moderate but non statistical DIF between the Arabic and French speaking examinees ( $DIF_{contrast} = 0.49, p > .05$ ) in favor of the Arabic speaking subgroup. Negligible DIF was identified between the Arabic and Spanish speaking examinees ( $DIF_{contrast} = 0.42, p > .05$ ) in favor of the Arabic subgroup as well. In this vein, Item 15 also depicts negligible DIF between the French and Arabic speaking examinees ( $DIF_{contrast} = -0.42, p < .05$ ), but in favor of the French speaking subgroup. Comparisons between the Spanish speaking subgroup and the French speaking subgroup did not suggest any DIF in Items 14, 15 and 16. In contrast, for the Arabic speaking subgroup ( $\delta = 1.45$ ), the item was more difficult compared to the French ( $\delta = 1.00$ ) and Spanish ( $\delta = 0.98$ ) speaking subgroups. Item 17 follows the same idea of impact evoked on behalf of Items 9 and 11 in that Spanish speaking examinees were outperformed by native or natively like speakers of French, which is somehow to be expected and considered as impact. Items 18 and 19 followed a similar trend to Item 12. French and Arabic speaking examinees performed similarly on Item 18 ( $\delta = 0.59$  and  $\delta = 0.52$ , respectively) and on Item 19 ( $\delta = 0.62$  and  $\delta = 0.68$ , respectively) suggesting a similar probability of correct response across all levels of theta. Thus, on the graphics, for both items, the item characteristic curves for the French and Arabic speaking subgroups overlap.

A content analysis of the Items 14 through 19 did not reveal potential sources of bias rooted in the test items that could have placed a given L1 subgroups in either an advantage or in a disadvantage. It is puzzling, however, how native speakers of French, taking a French test developed by a French firm, had a lower probability of correct response on certain items of the test. Perhaps, native speakers are not used to this type of assessments and too easy items trigger inattention to the aural input, thus missing the key information to respond the item correctly. Or perhaps, items are too easy for them and they over think too much.

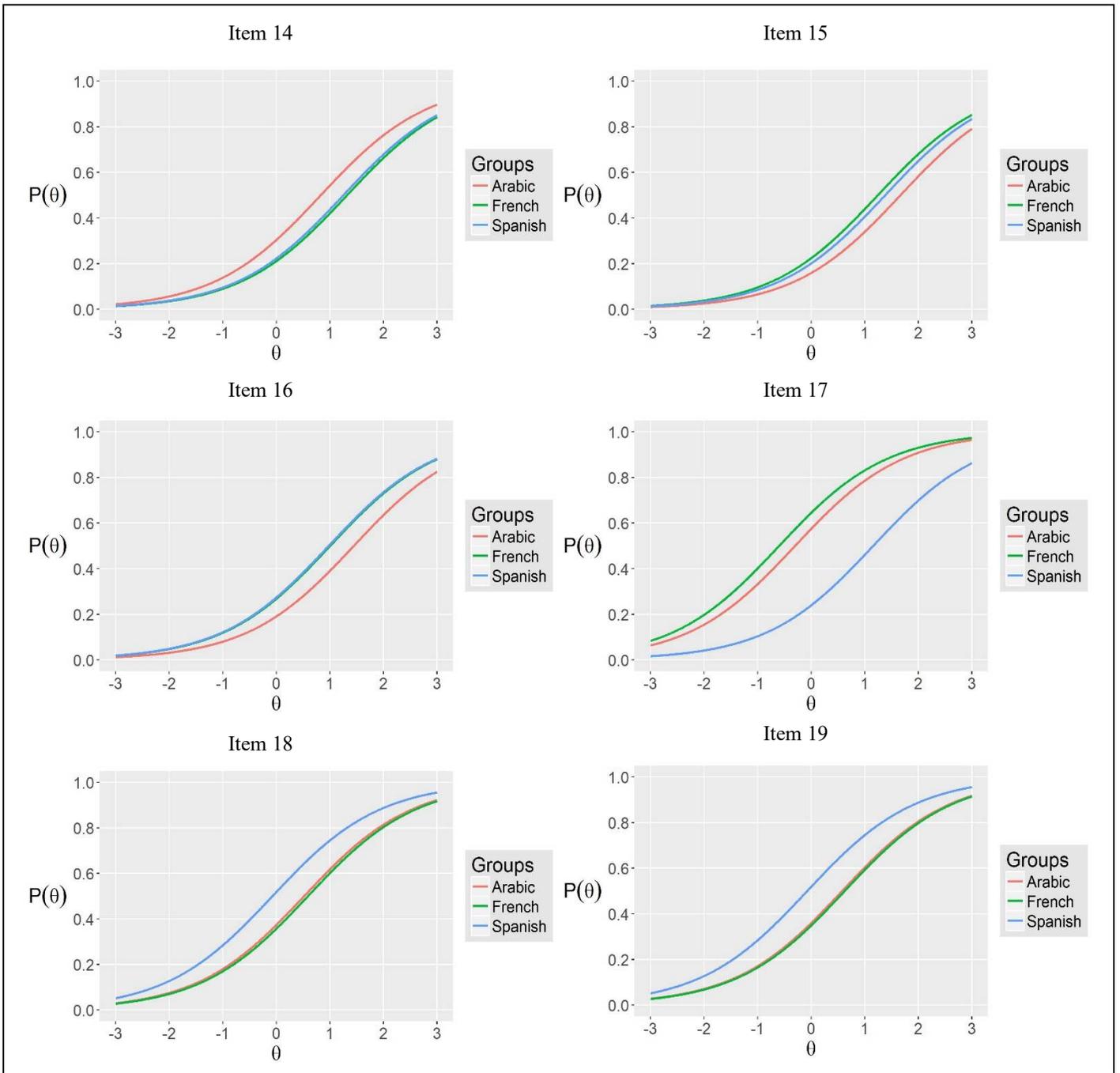


Figure 19 Item Characteristic Curves for Items 14 through 19 in Form A

Figure 20 below provides the item characteristic curves for Items 20 through 25 in Form A. Items 20 through 23 only exhibited negligible DIF across all comparisons made between the subgroups of interest. Graphics for Items 24 and 25 depict the Spanish speaking examinees as

the strongest subgroup where Item 24 was easier for the Spanish speaking subgroup ( $\delta = -1.08$ ) compared to the French ( $\delta = 0.44$ ) and the Arabic ( $\delta = 0.88$ ) speaking subgroups.

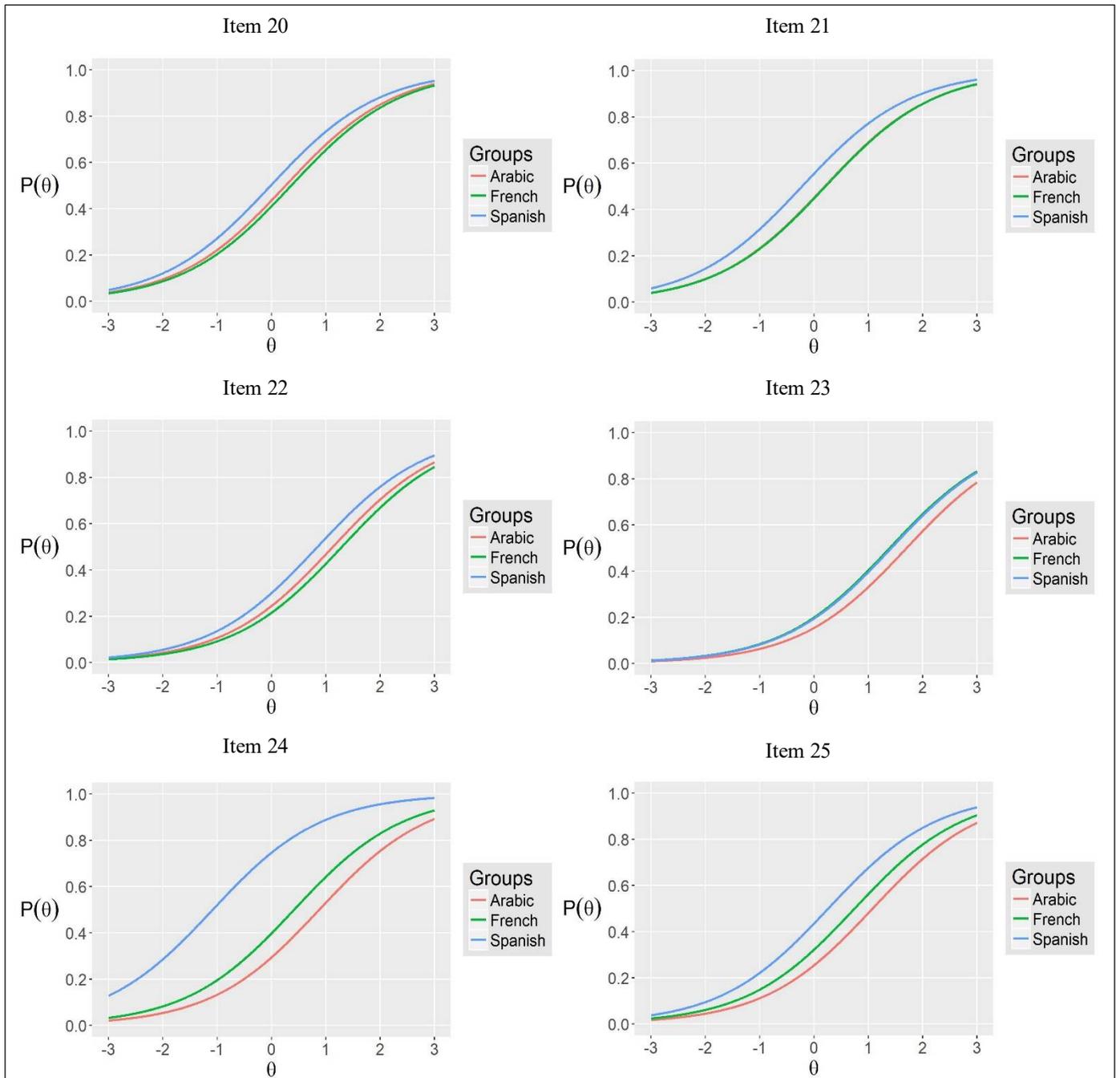


Figure 20 Item Characteristic Curves for Items 20 through 25 in Form A

A content analysis of the items revealed that these items consisted of longer stretches of aural discourse. Perhaps, the Spanish speaking subgroup are decoding the stream of sound using more effectively listening strategies than the French and the Arabic speaking subgroups.

Figure 21 provides the item characteristic curves for Items 26 and 28 for Form A. Item 26 was difficult for the French speaking subgroup ( $\delta = 1.80$ ) and much easier for the Arabic ( $\delta = 1.01$ ) and Spanish ( $\delta = 0.89$ ) speaking subgroups. Item 28 was difficult for the Arabic speaking subgroups ( $\delta = 1.54$ ) and less difficult for the French ( $\delta = 0.94$ ) and Spanish ( $\delta = 0.92$ ) speaking subgroups. A content analysis of these two items did not reveal potential issues of item bias across all the possible comparison among the examinees from the subgroups under investigation.

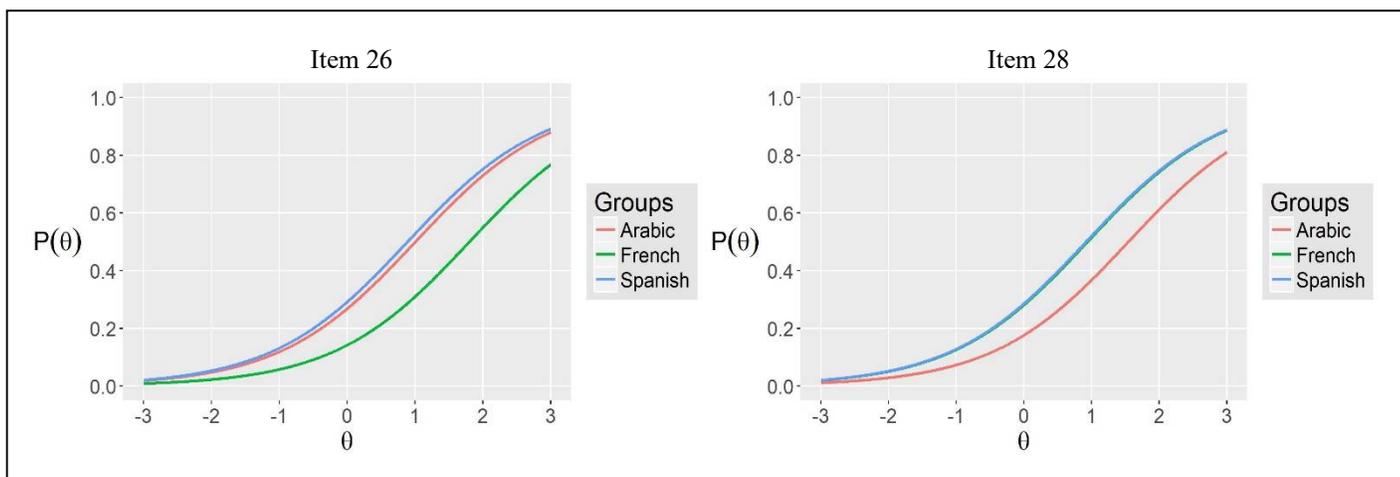


Figure 21 Item Characteristic Curves for Items 26 and 28 in Form A

The graphics on Figure 22 provide the item characteristic curves for the first language DIF in Form B. The subgroups of interest included Arabic ( $n = 427$ ), French ( $n = 678$ ), Persian ( $n = 215$ ), and Spanish ( $n = 322$ ). Figure 22 below depicts the item characteristics curves for Items 1, 4, 5, 6, 7 and 8. On Item 1, the Arabic speaking subgroup obtained a higher probability of correct response at lower levels of theta or listening ability. Item difficulty was approximately the same for French ( $\delta = -1.65$ ), and Spanish ( $\delta = -1.62$ ) speaking subgroups and a bit easier for the Persian speakers ( $\delta = -1.93$ ). Item 4 shows the French speaking group as having a higher probability of correct response than the Arabic, Persian and Spanish subgroups. This is expected since French is the first language of these examinees. An interesting patterned is observed in Item 5 as the Arabic and Spanish speaking subgroups found the item easier ( $\delta = -1.10$  and  $\delta = -$

1.11, respectively) compared to the French ( $\delta = -0.18$ ) and the Persian ( $\delta = -0.61$ ) subgroups. Items 6 and 7 follow a more expected hierarchy and can be considered as impact as both French and Arabic speaking subgroups, who speak French as either a first language or lingua franca, had a higher probability of correct response than subgroups that are likely to speak French as foreign language (i.e., Persian and Spanish subgroups).

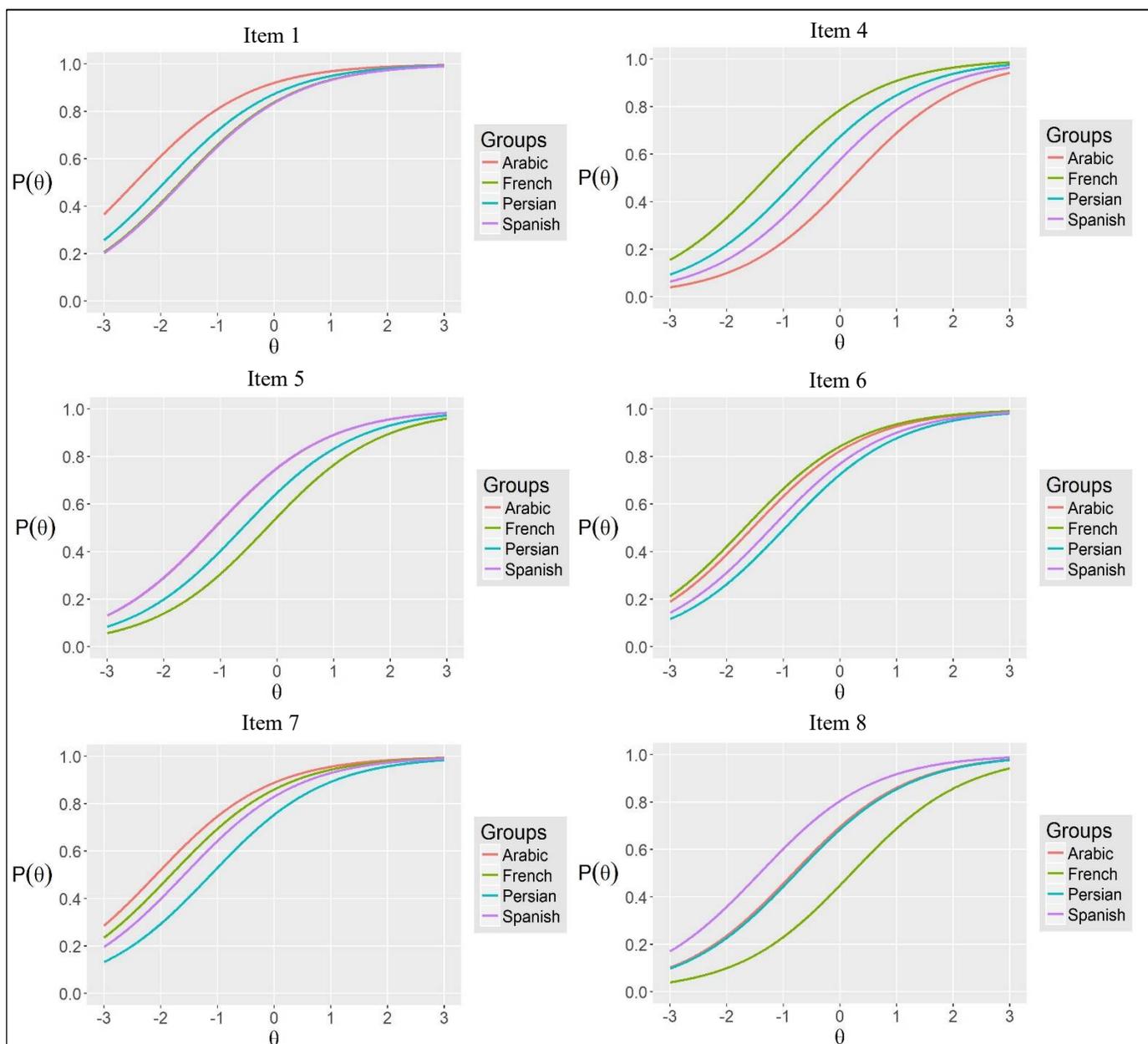


Figure 22 Item Characteristic Curves for Items 1, 4, 5, 6, 7 and 8 in Form B

A content analysis of the Items 1, 4, 5, 6, 7 and 8 did not reveal a potential issue that would flag the difference in performance. Items 5 and 8 were fairly easy, but surprisingly the French

speaking subgroup had a lower probability of correct response than the other groups. This is suspect of inattention or carelessness on the part of the French speakers.

Figure 23 below provides the item characteristics curves for Items 9 through 14 in Form B. The item characteristic curves (ICCs) for Item 9 provide an ideal situation of No DIF, suggesting that all the item characteristic curves overlap and all examinees have a similar probability of correct response regardless of their first language.

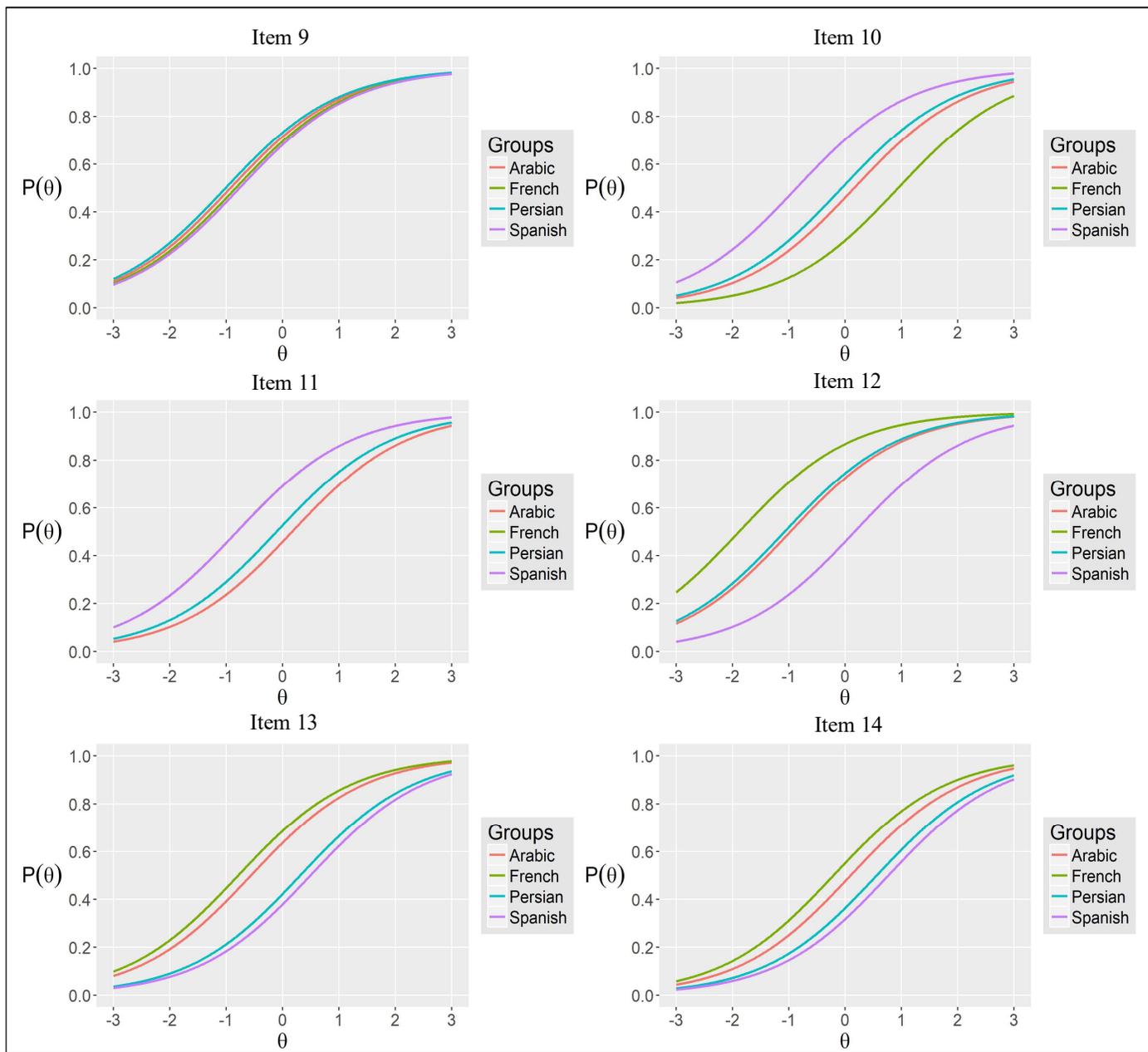


Figure 23 Item Characteristic Curves for Items 9 through 14 in Form B

Item 10 suggests the Spanish speaking examinees obtained the highest probability of correct response. Again and unexpectedly, the French speaking examinees had the lowest probability of correct response on Item 10 ( $\delta = 0.94$ ), and Arabic as well as Persian speaking examinees had a similar probability of success ( $\delta = 0.16$  and  $\delta = -0.6$ , respectively). On Item 11, French and Spanish speaking subgroups obtained a similar probability of correct response as the ICCs follow the same ogive curve across all levels of theta ( $\delta = -0.81$  for both groups), but the Arabic and Persian speaking subgroups performed a bit different from each other ( $\delta = 0.17$  and  $\delta = -0.11$ , respectively). Item 12 was relatively easy for the French ( $\delta = -1.88$ ) speaking subgroup, but more challenging for the Arabic ( $\delta = -0.97$ ), Persian ( $\delta = -1.07$ ), and Spanish ( $\delta = 0.17$ ) speaking subgroups where the Arabic and Persian speaking examinees performed fairly similar on this item. The probability of correct response on Items 13 and 14 followed a similar trend in which a hierarchy of subgroups can be readily identified and explained. In this regard, on Items 13 and 14 the French ( $\delta = -0.78$  and  $\delta = -0.21$ , respectively) and Arabic ( $\delta = -0.56$ ;  $\delta = 0.10$ , respectively) speaking subgroups obtained a higher probability of correct response across all levels of theta when compared to the Persian ( $\delta = 0.31$  and  $\delta = 0.56$ , respectively) and the Spanish ( $\delta = 0.50$  and  $\delta = 0.77$ , respectively) speaking subgroups. This is in line with the following hypothesis: native or natively like (i.e., French and Arabic speakers from Africa) speakers should have a higher probability of correct response on all TCF items and this is heightened when the subgroups' ability distributions of different groups are mismatched (Liu *et al.*, 2016). A content analysis of this set of items did not reveal a potential problem that could explain why the items (excluding Item 9) functioned differently across the subgroups. Items 12, 13, and 14 can be just a question of impact. Items 10 and 11 were relatively easy, but the French speaking subgroup had the lowest probability of correct response across all levels of theta. As in previous descriptions of this type of trend, inattention might be a potential explanation to this tendency.

Figure 24 below depicts the ICCs for Items 15 through 20 in Form B. For Items 16, 17, 18 and 20 the native speakers of French obtained a lower probability of correct response where the Persian speakers outperformed the other subgroups. For instance, Item 16 was easier for Persian speakers ( $\delta = 0.51$ ) when compared to Arabic ( $\delta = 1.23$ ), French ( $\delta = 1.42$ ) or Spanish ( $\delta = 0.98$ ) speakers. Despite the higher probability of correct response for the Persian speaking

subgroup, the ICCs on Item 16, only provided an indication of substantive DIF between the French and the Persian subgroups ( $DIF_{contrast} = 0.91$  [ $\chi^2 = 2.62, p > .05$ ]). The ICC's for Items 15 and 19 suggest that the French speaking group obtained a higher probability of correct response compared to the Arabic, Persian and Spanish speaking subgroups. A content analysis of this set of items did not reveal content that would have triggered potential bias favoring any of these L1 subgroups.

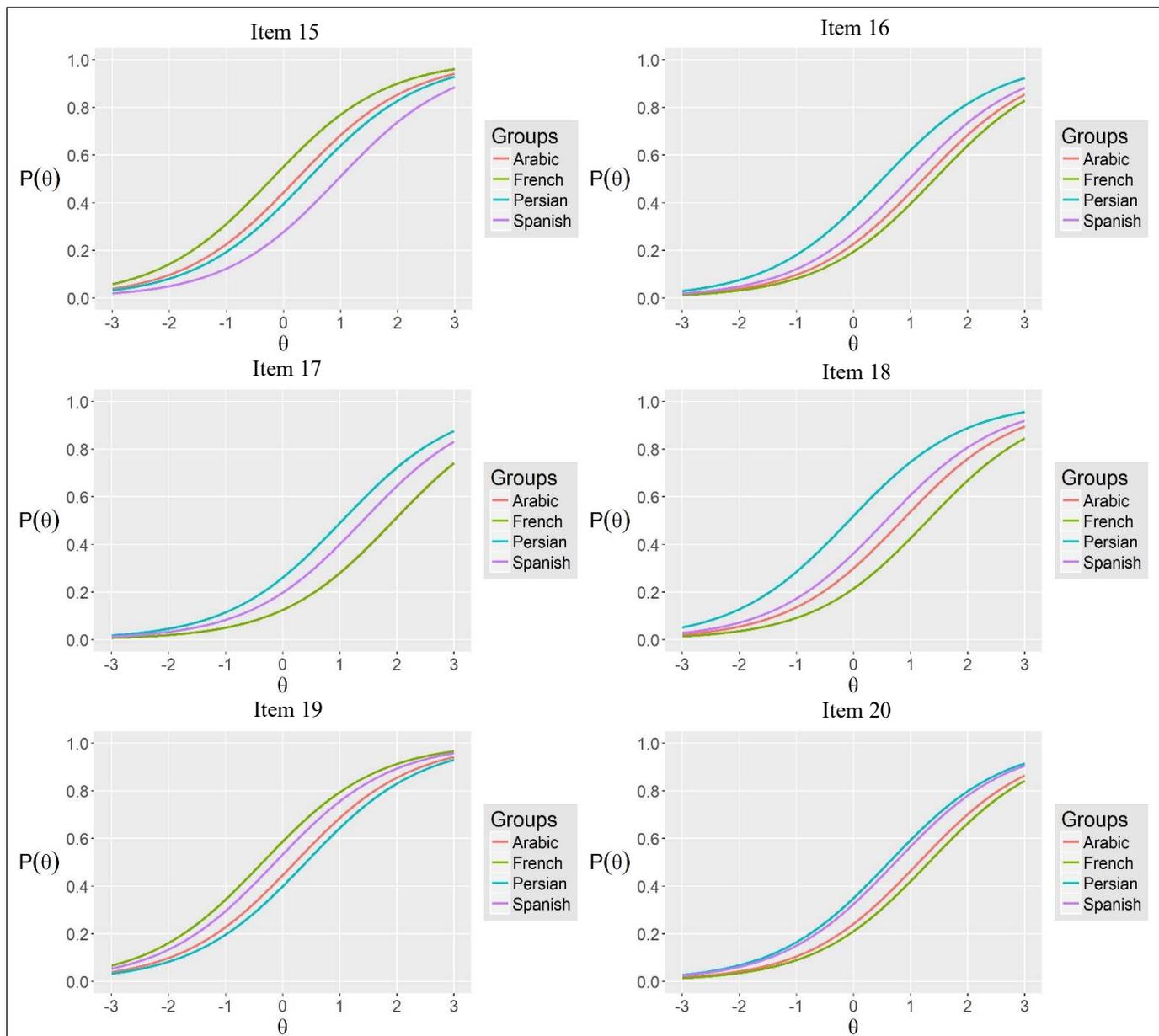


Figure 24 Item Characteristic Curves for Items 15 through 20 in Form B

Figure 25 below displays the ICCs for Items 21, 23, 24, 25 and 27 in Form B across first language subgroups of interest (i.e., Arabic, French, Persian, and Spanish speaking examinees). Items 24 and 25 reflect the idea of impact that was previously evoked on behalf of Items 4 and 12. That is, native speakers or nativelike users of the target language of the assessment should have or are expected to have a higher probability of correct response on test items.

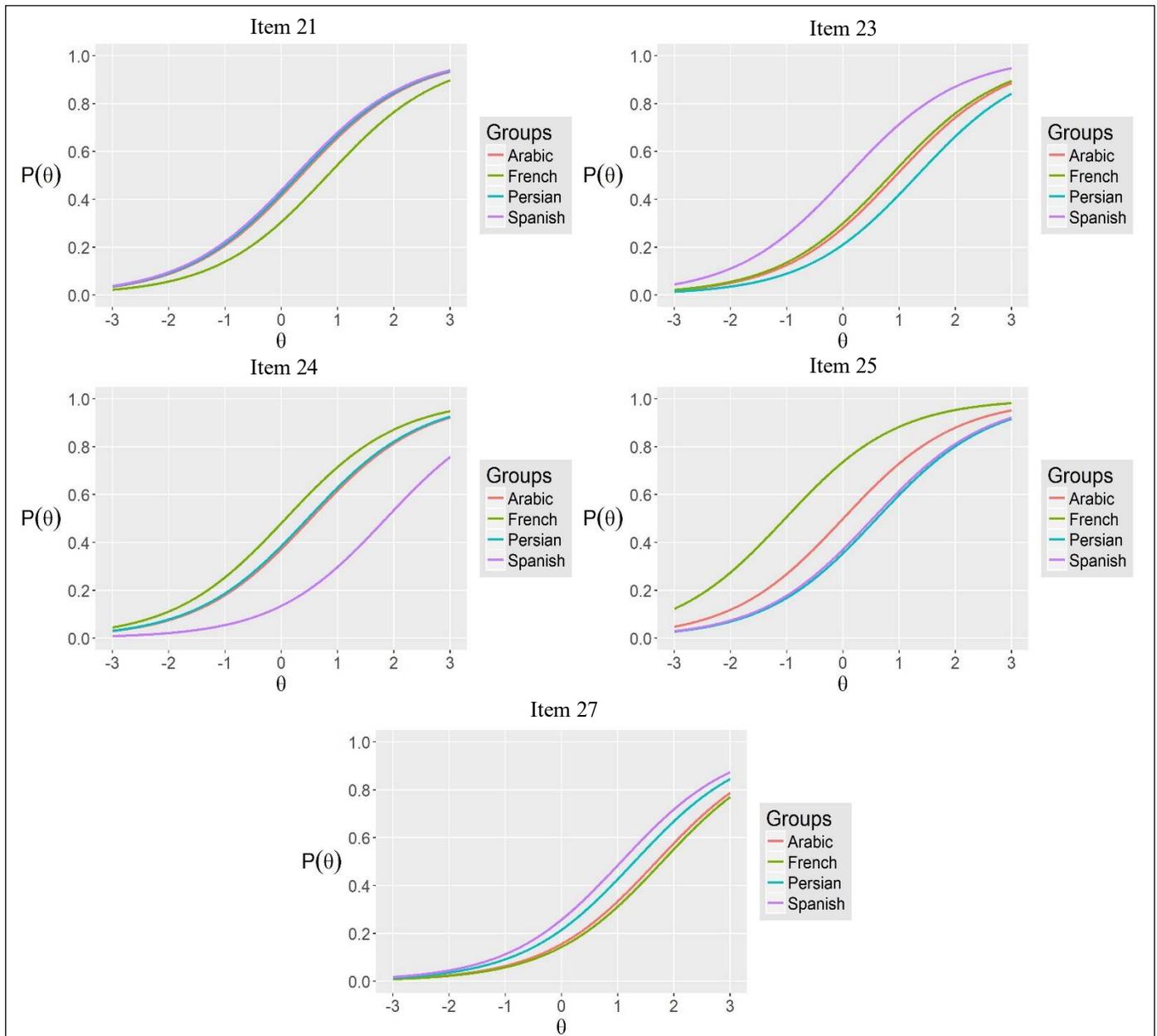


Figure 25 Item Characteristic Curves for Items 21, 23, 24, 25, and 27 in Form B

On a different note, the ICC's for Item 21 did not flag any DIF among the Arabic, Persian and Spanish speakers subgroups as the three ICC's are almost identical. However, the French speakers were clearly outperformed, exhibiting a lower probability of correct response. The ICC's for Item 23 suggest that the Spanish ( $\delta = 0.09$ ) speaking subgroup performed much better on this item than the other subgroups, especially when compared to the Persian ( $\delta = 1.32$ ) speaking subgroup, which obtain a much lower probability of correct response exhibiting moderate to large DIF ( $DIF_{contrast} = 1.24$ , [ $\chi^2 = 27.85$ ,  $p < .001$ ]). Item 27 depicts negligible DIF between the Persian ( $\delta = 1.30$ ) and Spanish ( $\delta = 1.06$ ) speaking subgroups as well as between the French ( $\delta = 1.79$ ) and Arabic ( $\delta = 1.69$ ) speaking subgroups. However, when foreign speakers of French are compared to Arabic or French speakers DIF is present in the data and suggesting that foreign or second language speakers of French outperformed native-like speakers or native speakers of French. As in previous content analysis, items were not flagged as carrying potential content that would trigger bias. This was the case for DIF analysis that looked at test fairness across the first language of the subgroups of interest. A couple of hypotheses were advanced: first, when native speakers or native-like speakers of the target language of the assessment obtained a higher probability of correct response than the other groups, the difference in performance could be attributed to impact and not bias. Second, when presented with very easy items, native and native-like speakers (i.e., French and Arabic subgroups) of the target language of the test might pay little attention to the aural input or overthink the correct answer to the item in question.

#### **4.2.5 Differential item functioning across age for immigration classes**

Differential item functioning across examinees' ages was also examined and the subgroups of interest for the DIF study across age for immigration classes<sup>23</sup> included: Group A ([18 – 29],  $n = 580$ ), Group B ([30 – 35],  $n = 571$ ), and Group C ([36 +],  $n = 344$ ) for Form A. And for Form B the grouping was as follows: Group A ([18 – 29],  $n = 885$ ), Group B ([30 – 35],  $n = 809$ ), and Group C ([36 +],  $n = 423$ ). At the time of the DIF study these age ranges reflect

---

<sup>23</sup> The DIF study was carried out when these age ranges were used by the Ministry of Quebec immigration to award points towards permanent resident applications. The age ranges may not reflect the latest amendments made to the law.

the criteria of the selection process of potential immigrants to Quebec adopted by the *ministère d'Immigration, Diversité et Inclusion* (MIDI).

In Form A, DIF was only flagged in four items. Items 1, 11, 17 and Item 24. Figure 26 below provides the ICCs for the four items. The probability of correct response for Item 1 was higher for the youngest subgroup (i.e., Group A). The DIF contrasts ( $\delta_{\text{age group A}} = -3.40 - \delta_{\text{age group B}} = -2.94$ ,  $\text{DIF}_{\text{contrast}} = -0.46$  [ $\chi^2 = 3.11$ ,  $p < .05$ ];  $\delta_{\text{age group A}} = -3.40 - \delta_{\text{age group C}} = -2.72$ ,  $\text{DIF}_{\text{contrast}} = -0.68$  [ $\chi^2 = 6.88$ ,  $p < .05$ ]) indicated that the item exhibited slight to moderate, and moderate to large DIF against low ability Age Group B and low ability Age Group C, respectively. Item 11 indicated slight to moderate DIF between Age Group A and Age Group C in favor of the younger group ( $\delta_{\text{age group A}} = -0.48 - \delta_{\text{age group C}} = 0.08$ ,  $\text{DIF}_{\text{contrast}} = -0.56$  [ $\chi^2 = 7.99$ ,  $p = .005$ ]). This pattern was also observed in Item 17 ( $\delta_{\text{age group A}} = -0.19 - \delta_{\text{age group C}} = 0.27$ ,  $\text{DIF}_{\text{contrast}} = -0.46$  [ $\chi^2 = 4.75$ ,  $p < .05$ ]).

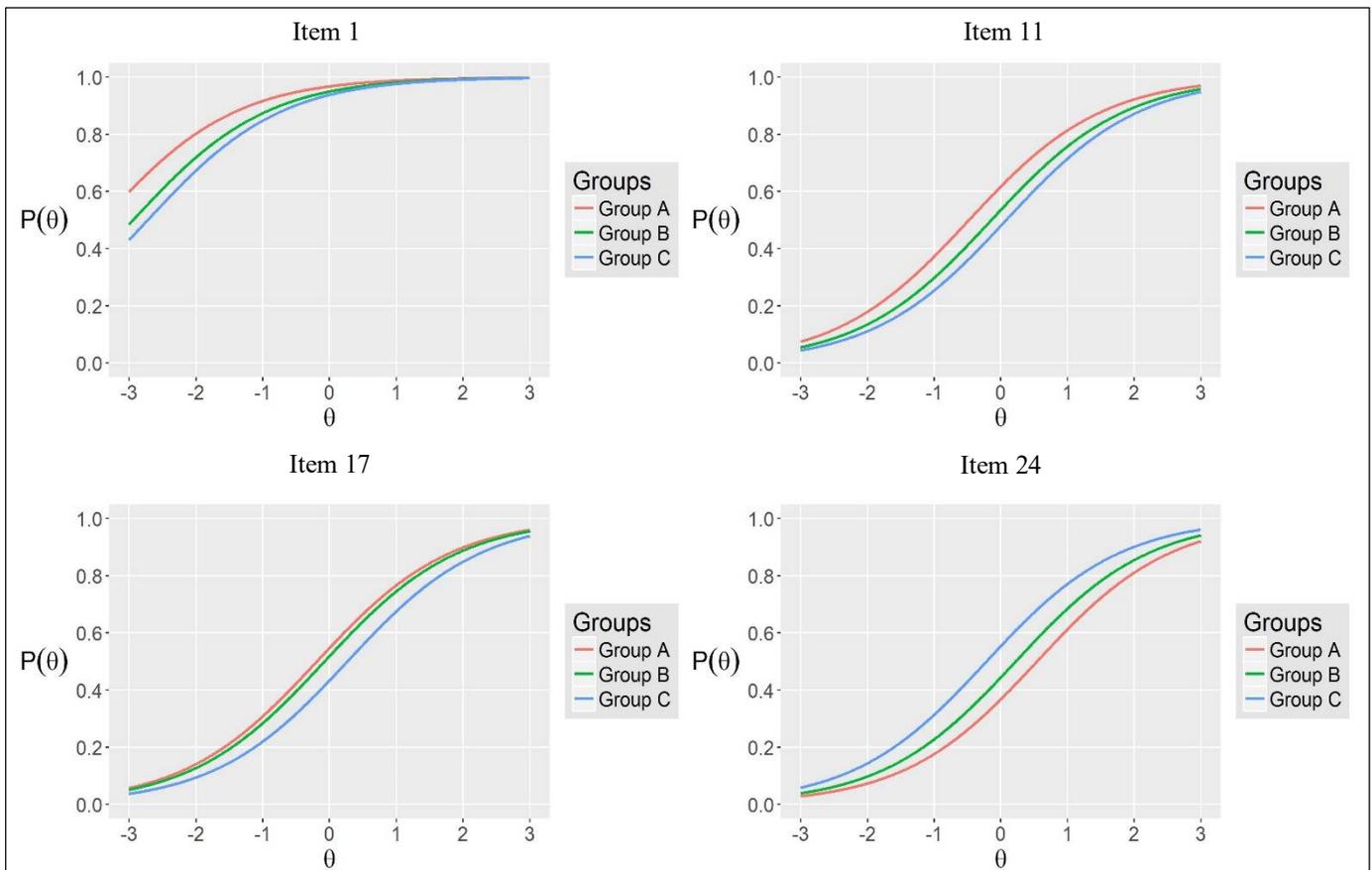


Figure 26 Item Characteristic Curves for Items 1, 11, 17 and 24 in Form A

On Item 24 this pattern was reversed as the item exhibited substantive DIF in favored of the oldest subgroup ( $\delta_{\text{age group A}} = 0.54 - \delta_{\text{age group C}} = -0.21$ ,  $\text{DIF}_{\text{contrast}} = 0.76$  [ $\chi^2 = 15.55$ ,  $p < .001$ ]). A content analysis of the items only revealed a potential explanation of bias for Items 1 and 24. On one hand, the picture recognition format for Item 1 probably eased comprehension for the younger group if visual input enhances comprehension for younger examinees this is a question that could be further explored. On the other hand, the aural input for Item 24 provided a constructive critique of a literary writer, this topic might not have attracted the attention of younger examines and put them in a disadvantage. Again this hypothesis would require more empirical evidence to explain the source of DIF in this item.

In Form B, DIF was only flagged in Item 8. This item indicated slight to moderate DIF between Age Group A and Age Group C in favor of the younger group ( $\delta_{\text{age group A}} = -0.89 - \delta_{\text{age group C}} = -0.43$ ,  $\text{DIF}_{\text{contrast}} = -0.46$  [ $\chi^2 = 5.69$ ,  $p = .017$ ]). A content review of the item revealed a potential explanation: a university representative welcomed new students to campus, then, a student asked a question about renewing the student card as it is needed for the tramway and the university restaurant tickets. Younger audiences as in groups B and C could be more familiar with this type of discourse than Group C (examinees over 36 years of age).

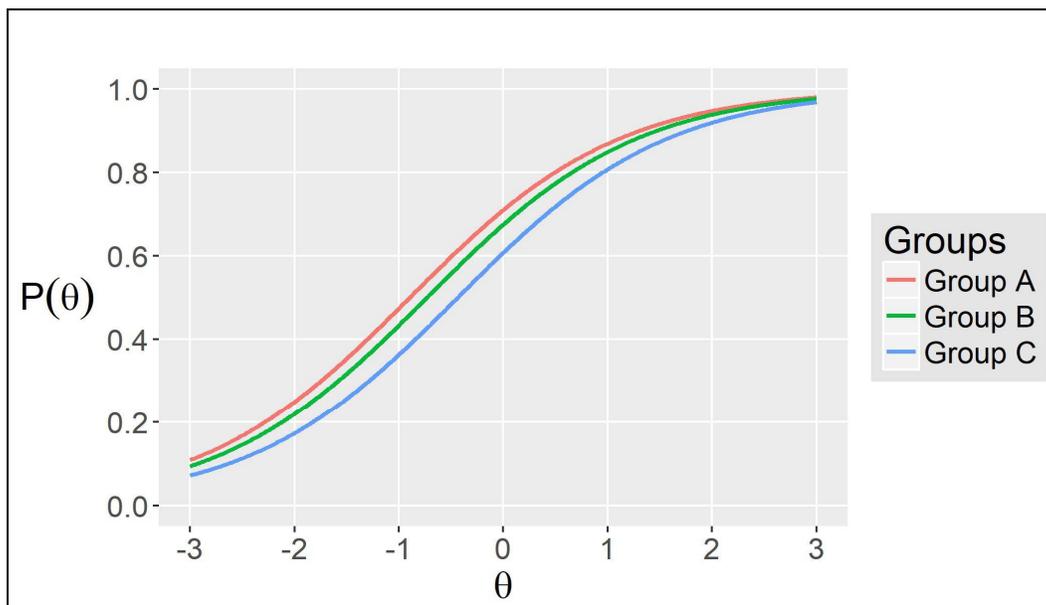


Figure 27 Item Characteristic Curves for Item 8 in Form B

#### 4.2.6 Differential item functioning across age with similar sample sizes

Differential item functioning for similar sample sizes across examinees' age was also examined. This was considered because DIF can be induced when the distribution of examinee abilities differs across groups and since the age grouping criteria used at MIDI appeared arbitrary, conducting DIF analyses using similar sample sizes, large enough to warrant parameter stability, would provide additional empirical evidence to substantiate this MIDI policy. The subgroups of interest for the age DIF analyses across similar sample size classes included: Group D ([17 – 26],  $n = 292$ ), Group E ([27 – 30],  $n = 401$ ), Group F ([31 – 35],  $n = 462$ ), and Group G ([36 +],  $n = 344$ ) for Form A. And for Form B the grouping was as follows: Group D ([15 – 27],  $n = 547$ ), Group E ([28 – 30],  $n = 521$ ), Group F ([31 – 34],  $n = 528$ ), and Group G ([35 +],  $n = 522$ ).

In this iteration, DIF was repeatedly flagged for Items 1, 11, 17 and 24 in Form A as in the age grouping criteria used by MIDI. Two more items were also flagged for DIF (Items 2 and 28). Figure 29 below depicts the ICCs for these items. Items 1 and 24 followed the same trend as in the DIF analyses under the age grouping policy used at MIDI. In this regard, the picture recognition format of Item 1 still favored the youngest examinees at lower levels of listening ability. The DIF contrast ( $DIF_{\text{contrast}} = -.048$ , [ $\chi^2 = 2.17$ ,  $p > .05$ ];  $DIF_{\text{contrast}} = -.64$  [ $\chi^2 = 3.50$ ,  $p < .05$ ]), in relation to the two oldest groups (i.e., Groups F and G, respectively), suggested slight to moderate and moderate to large differential item functioning. Note, however, that the DIF contrast between Age Groups D and F was not statistically significant and the sample-based effect size for this comparison was small (Cohen's  $d = .372$ ). In regards to Item 24, the DIF contrast ( $DIF_{\text{contrast}} = .90$ , [ $\chi^2 = 17.48$ ,  $p > .001$ ]), in relation to the youngest age group suggested substantive DIF. As a result, the content analyses presented previously for Items 1 and 24 appear to hold more consistently for Item 24 with equivalent sample sizes across age groups.

DIF contrast for Item 2 was slight to moderate when comparing Age Group D to age groups E and F, but not statistically significant ( $DIF_{\text{contrast}} = -0.46$  [ $\chi^2 = 0.86$ ,  $p > .05$ ];  $DIF_{\text{contrast}} = -0.58$  [ $\chi^2 = 0.95$ ,  $p > .05$ ], respectively). This was also the trend in Item 11, which DIF contrast was statistical significant under the MIDI age grouping policy, but not consistently so with similar sample sizes between age groups D and G ( $DIF_{\text{contrast}} = -0.56$  [ $\chi^2 = 2.38$ ,  $p > .05$ ]). This

was marginally the same case for Item 17 when contrasting Age Group D and Age Group G ( $DIF_{contrast} = -0.46 [\chi^2 = 4.17, p = .04]$ ).

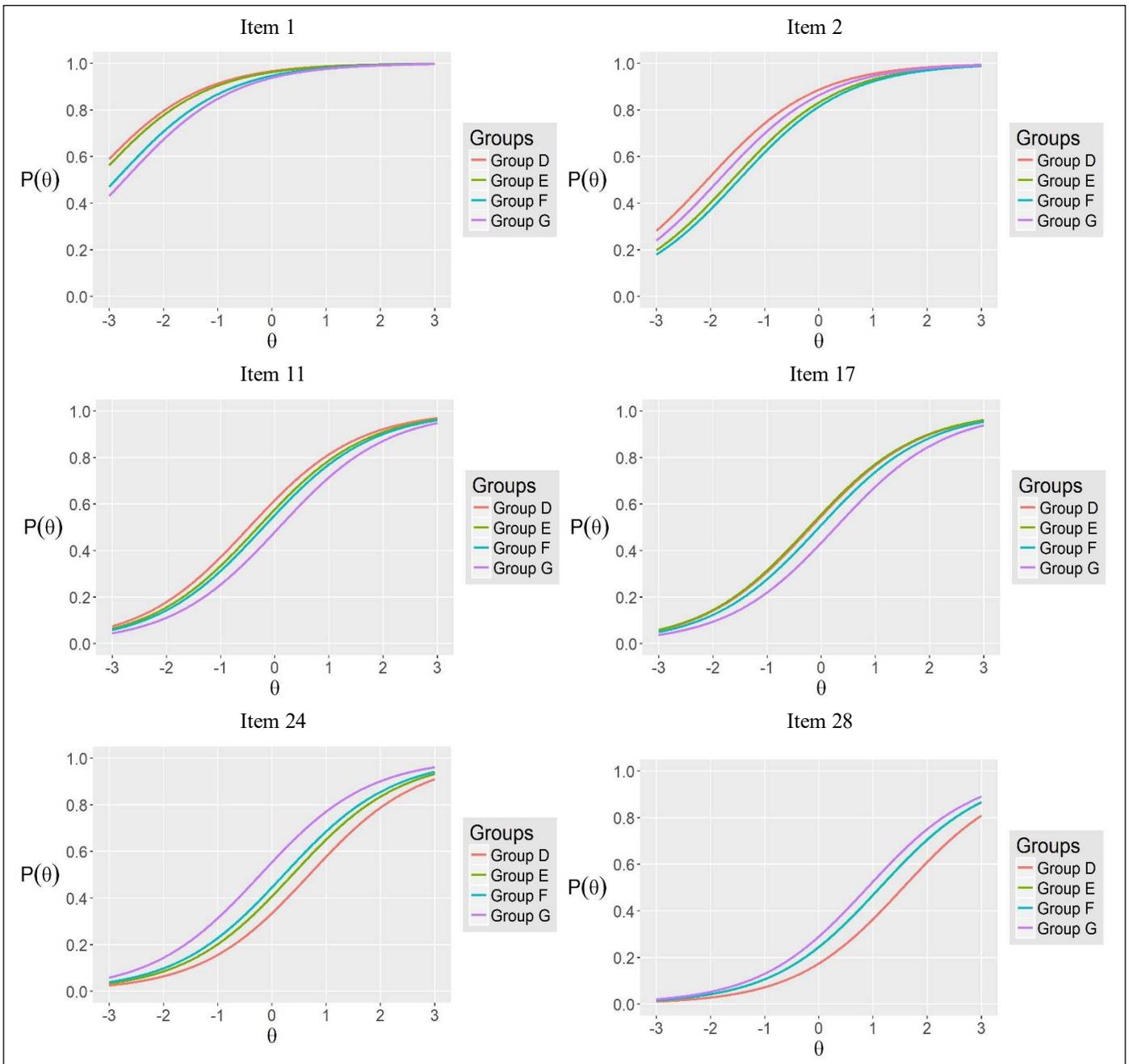


Figure 28 Item Characteristic Curves for Items 1, 2, 11, 17, 24 and 28 in Form A

In regards to Item 28, the DIF contrast was slight in the DIF analysis following the MIDI age grouping policy for Age Groups A and C ( $DIF_{contrast} = 0.43 [\chi^2 = 5.71, p > .05]$ ), favoring the oldest group (i.e., Age Group C). This was heightened when the DIF analysis controlled for

equal sample sizes across age. In this regard, Item 28 favored Group G (the oldest group) over Group D (the youngest group). The DIF contrast suggested moderate to large DIF between Group D and F ( $DIF_{contrast} = 0.66$  [ $\chi^2 = 10.23, p = .001$ ]). The content analysis of the item suggested that a key word to answer the item, *catamaran*, was out of reach of the youngest examinees.

DIF analysis controlling for sample size across age groups in Form B flagged DIF for Item 6 only. Figure 29 below provides the ICCs for the item where the youngest examinees (i.e., Group D) had a higher probability of correct response than the older age groups (i.e., Groups E, F, and G). The DIF contrast suggested slight to moderate DIF ( $DIF_{contrast} = -0.54$  [ $\chi^2 = 5.19, p < .05$ ];  $DIF_{contrast} = -0.59$  [ $\chi^2 = 7.96, p < .05$ ];  $DIF_{contrast} = -0.57$  [ $\chi^2 = 5.40, p < .05$ ], respectively). A content analysis of the item did not provided insightful information regarding this discrepancy in correct response probability. The item was an airport announcement that required listeners to understand explicit aural discourse. This item was one of the easiest on this test form.

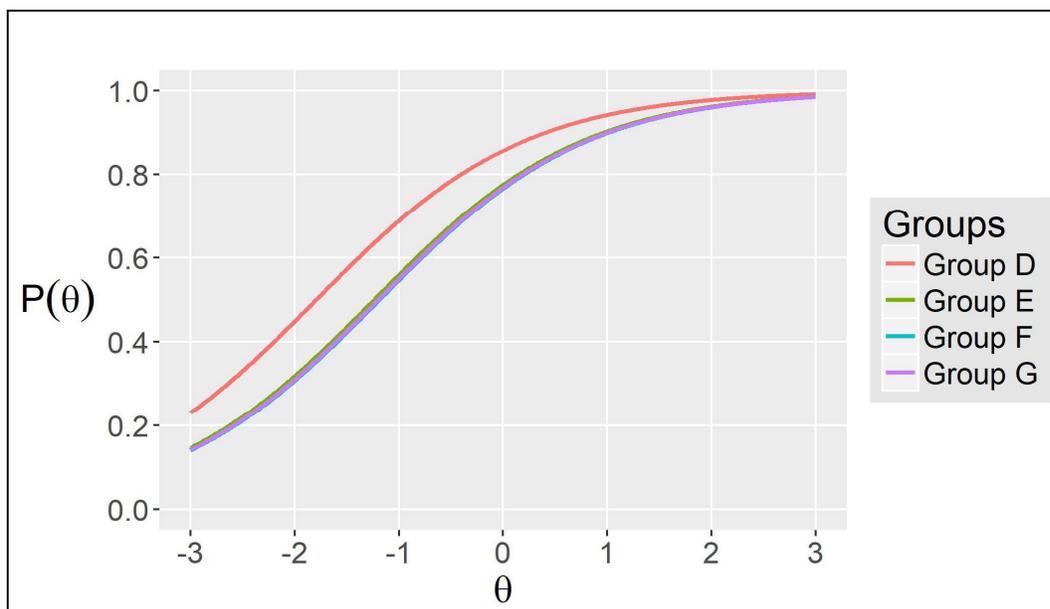


Figure 29 Item Characteristic Curves for Item 6 in Form B

#### 4.2.7 Differential item functioning across North and West African groups

Differential item functioning between North and West African subgroups of interest was also examined. These DIF analyses were enticing because the influence of the French language in these countries dates back to colonization periods and the French language is rooted in foreign and educational policy and it has been historically considered as an assimilation policy in both

North as well as West African states (Balch, 1909; Chumbow & Bobda, 2000). The sample for these subgroups included: North Africa ([Algeria, Morocco, and Tunisia],  $n = 266$ ), and West Africa ([Cameroon, Ivory Coast, and Senegal],  $n = 431$ ). Figure 30 below depicts the approximate geographical location for these countries.

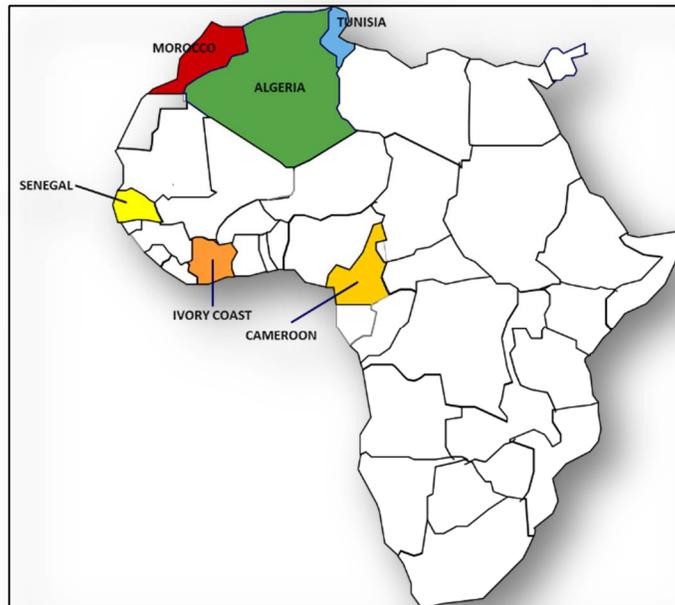


Figure 30 Geographical Location for North and West African Countries in the DIF Study

Figure 31 below depicts the ICCs for North and West African subgroups for Item 1 through 4 as well as Items 8 and 13 in Form A. This set of items favored the North African subgroup and the DIF contrasts ranged from slight to moderate for Items 4 and 13 ( $DIF_{contrast} = -0.63$  [ $\chi^2 = 7.83, p < .05$ ];  $DIF_{contrast} = -0.43$  [ $\chi^2 = 4.93, p > .05$ ], respectively) and from moderate to large for Items 1, 2, 3, and 8 ( $DIF_{contrast} = -1.09$  [ $\chi^2 = 3.87, p = .05$ ];  $DIF_{contrast} = -1.42$  [ $\chi^2 = 0.50, p > .05$ ];  $DIF_{contrast} = -0.70$  [ $\chi^2 = 4.77, p < .05$ ];  $DIF_{contrast} = -1.06$  [ $\chi^2 = 23.76, p < .05$ ] respectively). However, the standard error of measurement for Item 2 ( $SE = 1.00$ ) for the North African subgroup was large, thus affecting the stability of the estimation for the difficulty parameter for these examinees. With this limitation in mind, the picture recognition items in Form A (Items 1, 2, and 3) appeared to favor the North African subgroup. In addition, on Item 2, low ability North African examinees had a higher probability of correct response than the West African examinees. However, from  $-1.00$  logits onwards the probability of correct response for both groups was approximately the same. This item was the second easiest on the

test form and most examinees answered it correctly. A content analysis of this set of items did not reveal potential issues of bias except for the picture recognition items favoring the North African examinees.

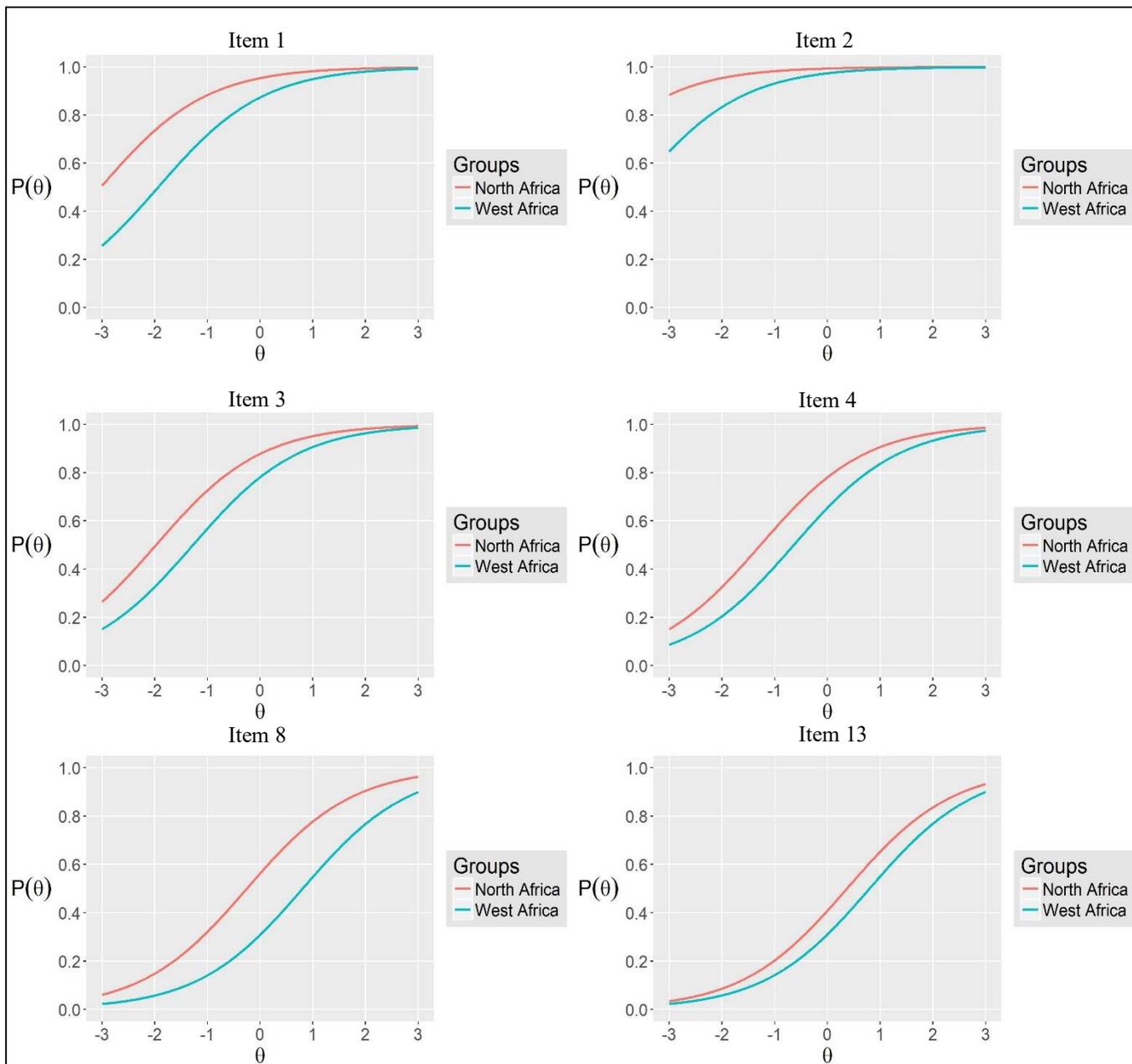


Figure 31 Item Characteristic Curves for Items 1, 2, 3, 4, 8 and 13 in Form A

Figure 32 below provides the ICCs for Items 15, 16, 22, and 24 through 26. In this set of items, Items 15, 16, 24, and 25 favored the West African group ( $DIF_{contrast} = 0.51$  [ $\chi^2 = 8.31, p < .05$ ];  $DIF_{contrast} = 0.80$  [ $\chi^2 = 15.39, p < .05$ ];  $DIF_{contrast} = 0.53$  [ $\chi^2 = 9.42, p < .05$ ];  $DIF_{contrast} =$

1.13 [ $\chi^2 = 31.73, p < .05$ ], respectively). The DIF was slight to moderate for Items 15 and 24 and moderate to large for Items 16 and 25. Conversely, Items 22 and 26 favored the North African examinees DIF contrasts suggested slight to moderate for Item 22 ( $DIF_{contrast} = -0.48$  [ $\chi^2 = 5.44, p < .05$ ] and moderate to large for Item 26 ( $DIF_{contrast} = -0.72$  [ $\chi^2 = 14.57, p < .001$ ])). A content analysis failed to provide potential clues that would clarify the DIF across this set of items.

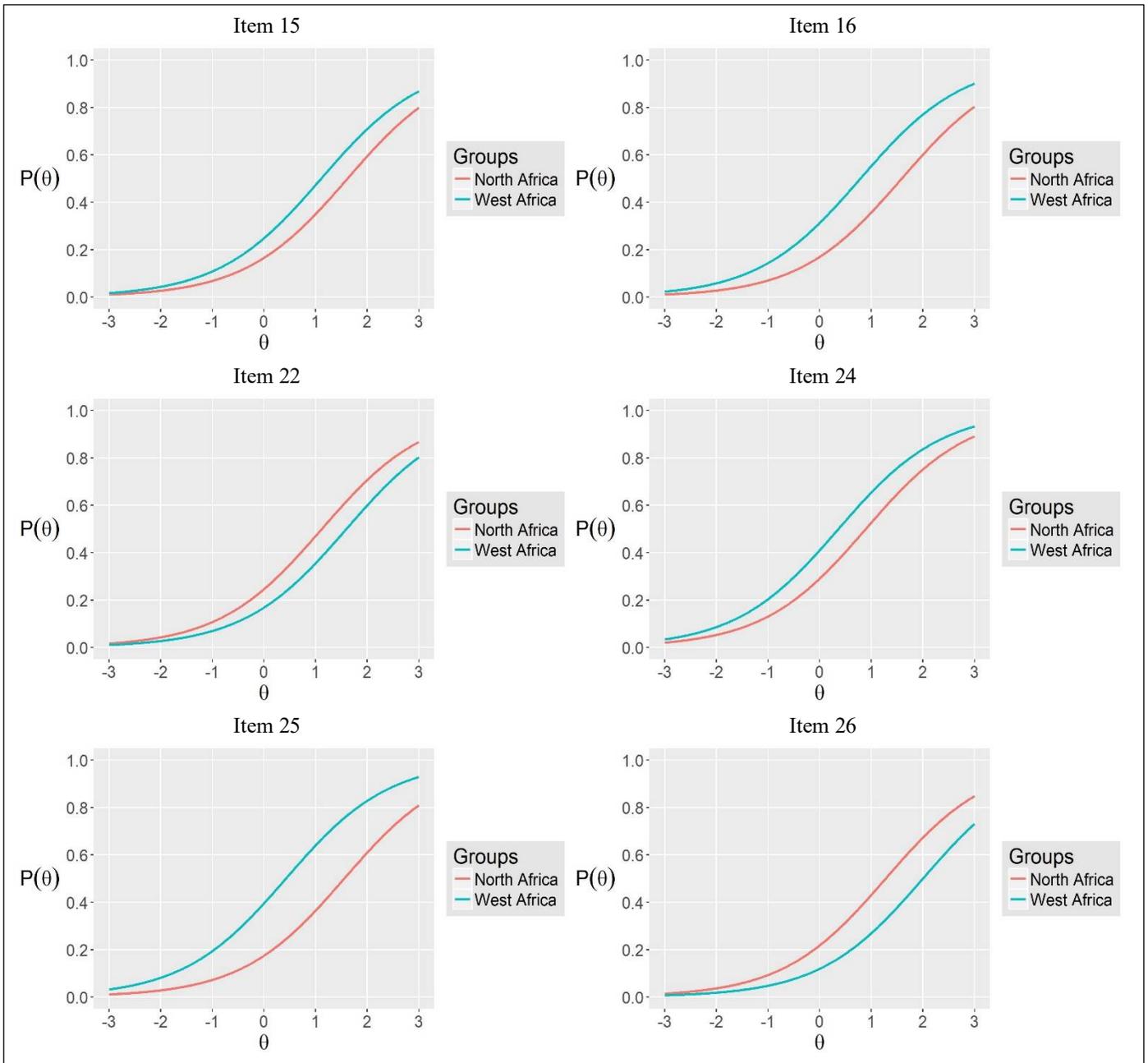


Figure 32 Item Characteristic Curves for Items 15, 16, 22, 24, 25 and 26 in Form A

Figure 33 below depicts the ICCs for Items 1, 4, 5, 6, 7 and 8 in Form B corresponding to the North and West African examinees. Similar to the findings in Form A, the picture recognition item (Item 1) still favored the North African examinees. The DIF contrast for Item 1 suggested moderate to large DIF ( $DIF_{contrast} = -1.35 [\chi^2 = 7.62, p < .05]$ ). Items 5, 7 and 8 also favoured the North African examinees and All the DIF contrasts were statistical significant and suggested moderate to large DIF ( $DIF_{contrast} \geq 0.95$ ). Item 6 exhibited slight to moderate DIF ( $DIF_{contrast} = -0.58$ ) also favoring the North African examinees.

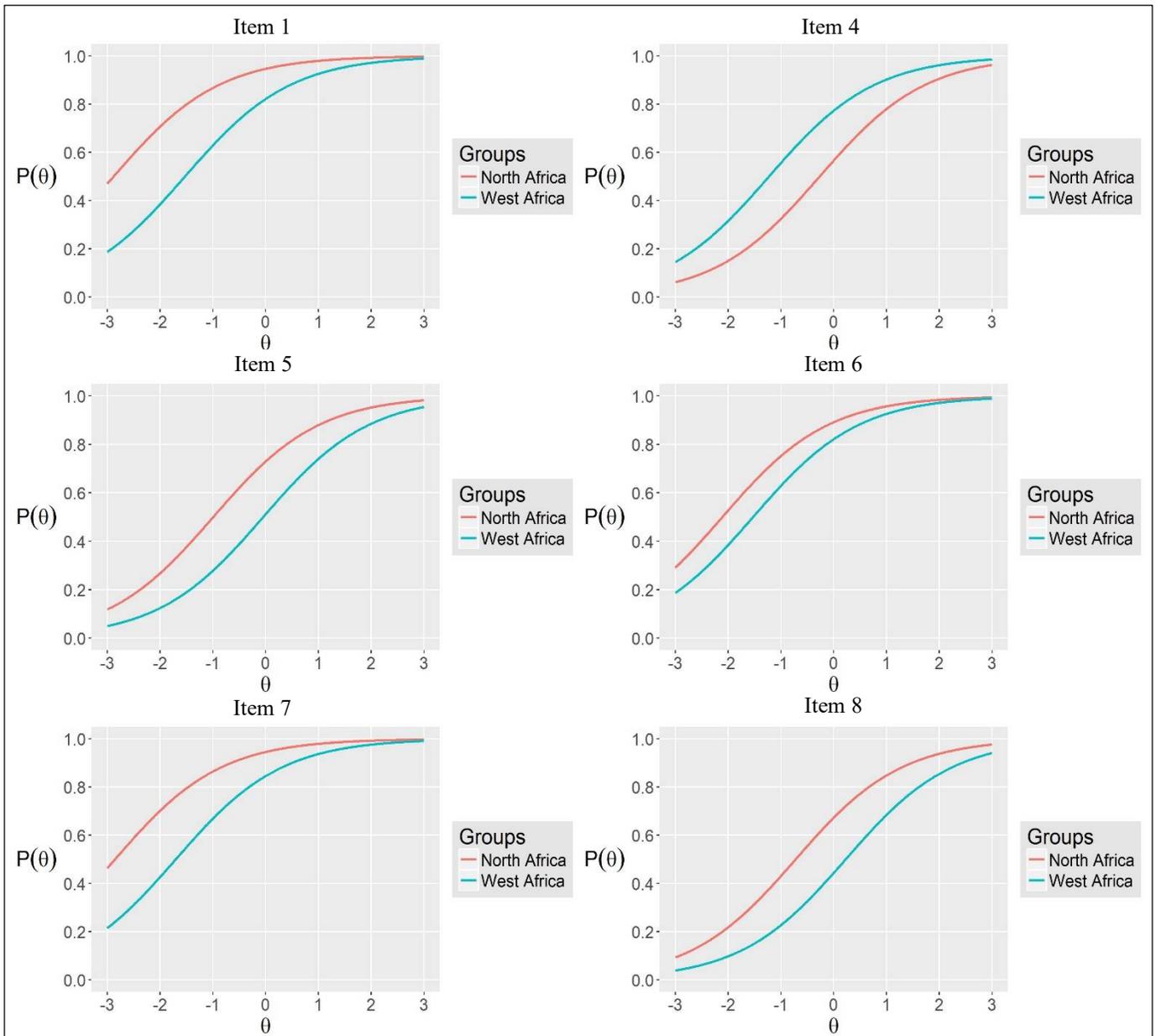


Figure 33 Item Characteristic Curves for Items 1, 2, 3, 4, 5 and 7 in Form B

Item 4 favored the West African examinees and the DIF contrast suggested that this difference in likelihood of correct response was moderate to large across all levels of theta ( $DIF_{contrast} = 0.96, [\chi^2 = 12.41, p < .001]$ ). Except for the item format (Item 1) favoring the North African examinees, a content analysis did not suggest potential problems that would cause the discrepancy in examinees performance in this set of items. Figure 34 below displays the ICCs for Items 9, 10, 11, 21, 23 and 25 in Form B.

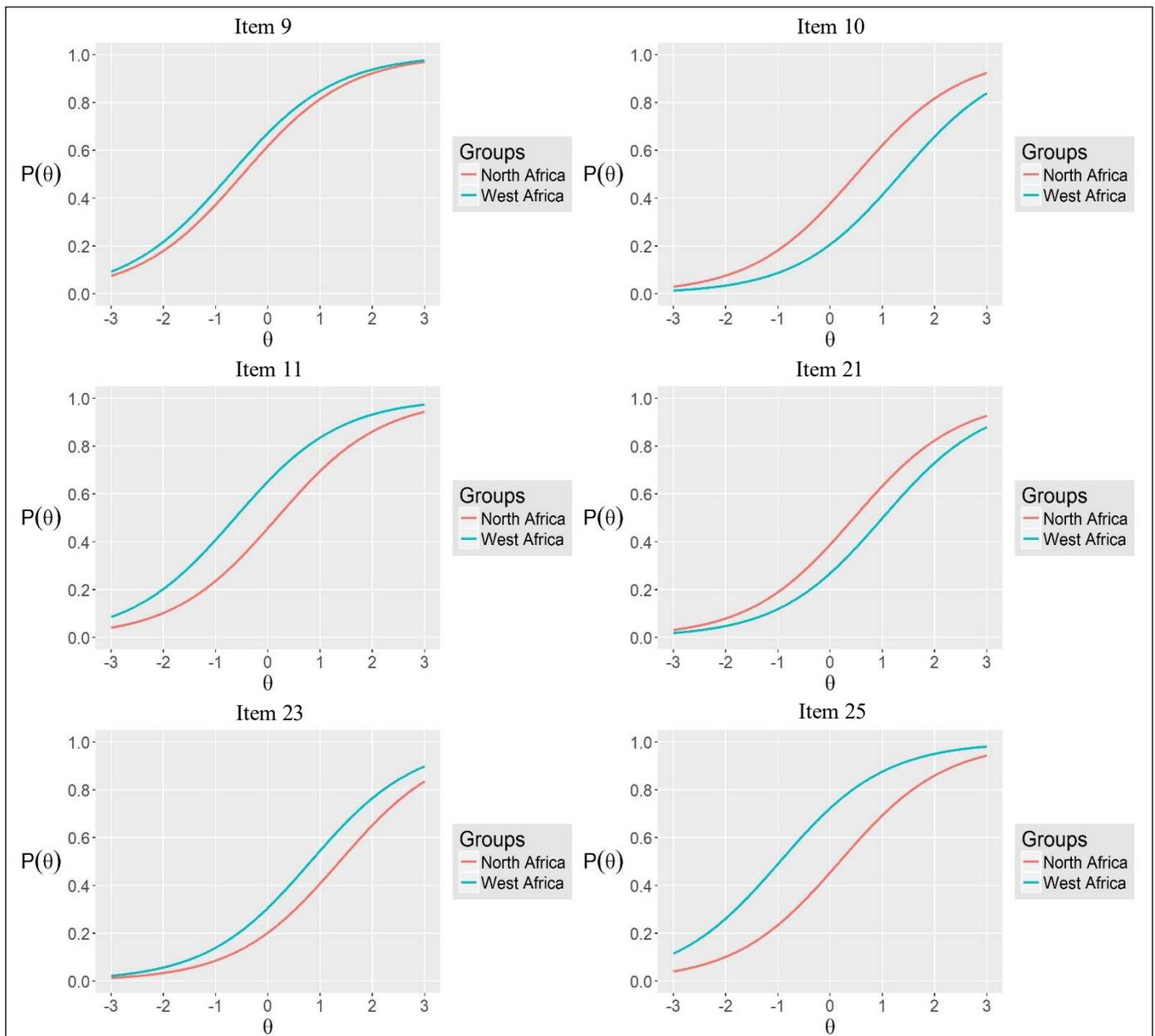


Figure 34 Item Characteristic Curves for Items 9, 10, 11, 21, 23 and 25 in Form B

Item 9 on Figure 34 above depicts no DIF as both subgroups obtained a similar probability of correct response across the ability continuum. Items 10 and 21 favored the North African examinees and the DIF contrasts were statistically significant and moderate to large ( $DIF_{contrast} = -0.85$ , [ $\chi^2 = 19.39$ ,  $p < .001$ ]) and slight to moderate ( $DIF_{contrast} = -0.55$ , [ $\chi^2 = 7.48$ ,  $p < .05$ ], respectively) for each item across all levels of theta or listening ability. West African examinees had a higher chance of correct response probability on Items 11, 23 and 25 and the DIF contrasts were statistical significant, but the absolute magnitude differed. For example, in Items 11 and 25 the magnitude was moderate to large and statistically significant ( $DIF_{contrast} = 0.80$ ,  $p < .001$  and  $1.14$ ,  $p < .001$ , respectively), depicting the ICCs for Items 11 and 25 favoring West African examinees across all levels of theta. Item 23 also favored the West African examinees, but the DIF was slight to moderate ( $DIF_{contrast} = 0.55$ , [ $\chi^2 = 7.27$ ,  $p < .05$ ]). A closer look at the content of the items did not reveal any potential clues that could have explained the DIF in these items.

#### **4.2.8 DIF across first language using the Standardization Index**

The DIF results provided in this section were not generated through Rasch modelling and included all the items while examining DIF using the standardization index (Dorans & Kulick, 1986) across small sample sizes of some first language subgroups: Russian ([Form A],  $n = 104$ ) and Chinese ([Form A],  $n = 87$ ), as well as Portuguese ([Form B],  $n = 97$ ) and Chinese ([Form B],  $n = 96$ ). This method was enticing to explore DIF in small sample sizes because the model assumptions are more relaxed than Rasch-based DIF analysis. In fact, the index is calculated by standardizing the difference between the proportions of examinees who answer a given item correctly (item facility) across score levels with the relative frequency of the reference group at different score levels.

The Standardization Index ranges from -1 to 1 and a standardized difference of 0.10 would indicate that on average, examinees in the reference group who are matched to examinees of the focal group have a 10% greater chance of answering the item correctly. Total scores could be affected if a lot of items exhibit DIF.

Figure 35 provides the standardization indices for the test items in Form A. On the graphic, Items 3, 8, 10, and potentially 27 favored the Russian speaking examinees (i.e., the reference group) whereas Items 2, 5, 6, 9, 23, and potentially 11 and 15 favored the Chinese students (i.e., the focal group).

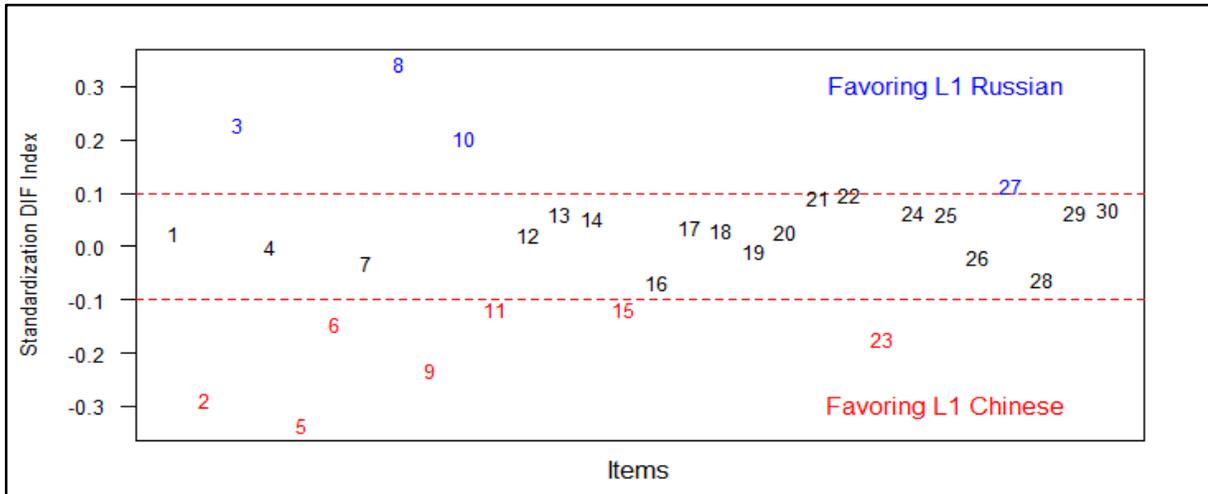


Figure 35 Standardization Indices for Chinese and Russian L1 Speakers

An interesting trend can be observed on the graphic. In the previous result from Rasch analysis, it was noted that items appearing last on the test tended to be, in general, the most difficult items. The difficult items are not a cause of concern here since they are not flag for DIF. Less challenging items appears to work different across Chinese and Russian examinees, more often favoring the former group. A content review of the test did not reveal a source of potential bias, but note however, that Chinese examinees are really adept and are very good at using test taking strategies (Li & Suen, 2015). Surprisingly, as shown in Figure 36 below, the standardization-based DIF analysis in Form B did not flag any DIF as every item fell within the range of -0.10 and 0.10, suggesting negligible DIF for Items 3, 10, 21, 5, 7, 11 and 28.

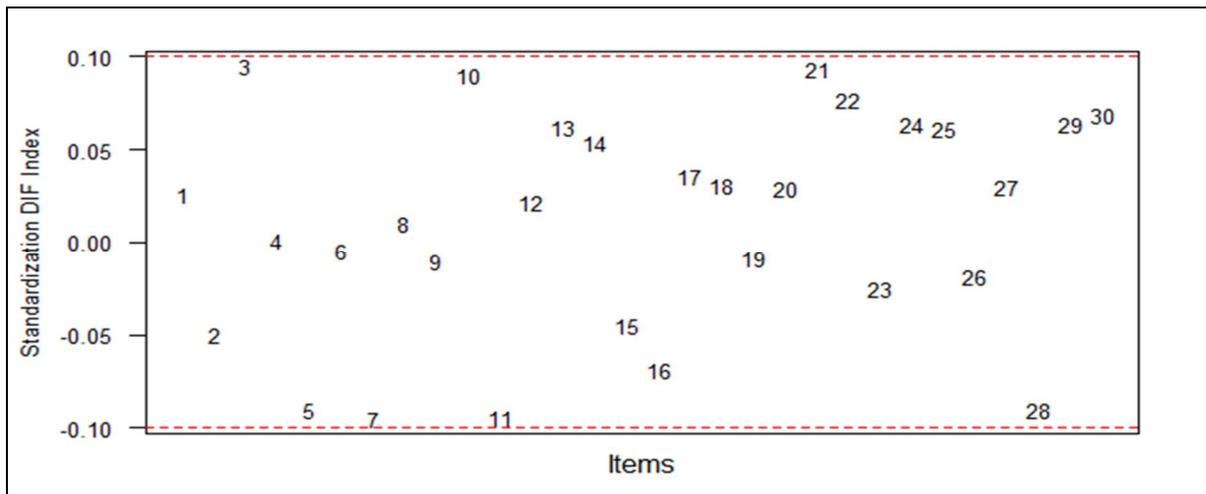


Figure 36 Standardization Indices for Chinese and Portuguese L1 Speakers

#### **4.2.9 Summary of DIF analysis across grouping variables**

The results presented in this section examined potential bias when items were flagged for DIF across gender, first language, age groups and geographical location (North and West Africa) in the TCF test forms under study. Given the assumptions of the model, in the Rasch-based DIF misfitting, noisy items were dropped from the analyses. Potential bias was identified in a few items only and related to speech perception (gender DIF) and discourse genre out of reach for the youngest examinees (literary critique). Another potential source of DIF was attributed to impact when exploring items across first languages. In this regard, native speakers of French were expected to do better on the TCF listening test. Although many items were flagged for DIF, it was challenging to associate the DIF to a potential source of bias related to the content of the items in question. This will be further elaborated in the discussion chapter.

#### **4.3 Research Question Three: Usefulness of Multiple Choice Items**

This section provides information on the effectiveness of multiple choice (MC) items in the assessment of L2 listening comprehension. Unlike traditional item analysis, which draws on classical test theory to examine the quality of the options in a multiple-choice item (usually by calculating item facility and point-biserial correlations on raw scores), this study analyzed the data within an item response theory (IRT) framework. The nominal response model (Bock, 1972) was used to model the options (i.e., the key and distractors) of MC items in both test forms. An attractive feature of this model is that its parameters represent the linear relation between the latent trait (e.g., listening ability) and the log-odds of responding in a given option of a multiple-choice item (Preston & Reise, 2015). The validity evidence gathered in this investigation was used to substantiate various inferences in the interpretation/use argument for the TCF in the context of Quebec immigration. As in previous sections, the results are presented sequentially for both test forms. This section addressed the following research question:

3) What information do the distractors and the keyed option of selected-response (SR) items (multiple-choice) provide to support the validity argument for the use of selected-response items to assess listening comprehension in a second language?

The results are presented in sequential steps. In order to draw valid inferences from IRT-based models, it is of paramount importance to endure adequate model – data fit. Unlike Rasch models and other IRT models (e.g., 2PL and 3PL), fit statistics are not readily available for the

NRM model. One way to address this limitation is to examine parameter invariance across grouping variables (e.g., gender). This study looked at parameter invariance to examine the stability of item parameters across gender. The variable gender was chosen to maintain a relative large sample size for data modeling under an NRM framework. Under this model two parameters are estimated for each option (i.e., eight parameters for a four-option multiple choice item), making sample size an important criterion to warrant the stability of parameter estimation.

### 4.3.1 Descriptive statistics

Table 32 below provides the frequencies in percentages for the options of each item in Form A. The asterisks denote the correct response to the item and as noted in the Rasch analysis the first items on the test are too easy since some of the options are barely chosen. Examining more closely the frequency distribution of the options, two issues can be identified: 1) low frequency will pose a problem for parameter estimation under NRM modeling and 2) some distractors are drawing too much attention, for example, distractor “C” on Item 26 is drawing 39.2% of examinees responses when in fact option “B” (37.8%) is the correct answer. Similarly, distractor “D” on Item 30 is attracting too much attention (40.4%) in relation to the key (15.5%).

Table XXXII Option Frequencies for Items in Form A

Item	<i>n</i>	Item Difficulty	Options			
			A	B	C	D
1	1500	0.945	3.1	1.7	*94.5	0.7
2	1500	0.847	3.1	1.8	*84.8	10.3
3	1496	0.794	4.9	12.8	2.7	*79.6
4	1479	0.801	4.9	*81.3	8.8	5.1
5	1469	0.533	8.8	*54.5	14.6	22.1
6	1486	0.705	7.9	*71.2	6.9	14.1
7	1490	0.824	3.3	6.0	7.7	*83.0
8	1491	0.585	*58.9	12.5	21.2	7.4
9	1481	0.656	7.0	9.5	17.2	*66.4
10	1496	0.720	9.0	15.1	*72.3	3.7
11	1494	0.638	*64.1	15.9	14.1	5.9
12	1490	0.586	16.0	7.2	*59.0	17.7
13	1490	0.564	23.2	10.5	*56.8	9.5
14	1489	0.410	15.2	13.8	*41.4	29.7
15	1498	0.402	33.0	*40.3	21.1	5.7
16	1494	0.406	12.7	20.6	26.0	*40.8
17	1495	0.609	9.6	*61.1	8.8	20.5
18	1489	0.544	20.9	*54.8	12.2	12.1
19	1498	0.514	16.2	22.6	*51.5	9.6
20	1483	0.565	21.5	*57.2	10.0	11.3

Table XXXII continued

21	1494	0.599	7.4	13.5	*60.2	18.9
22	1489	0.414	29.8	13.3	15.2	*41.7
23	1490	0.352	*35.4	18.9	22.8	22.9
24	1493	0.557	10.6	22.0	11.5	*56.0
25	1491	0.472	21.3	12.5	18.8	*47.5
26	1496	0.376	11.7	*37.8	39.2	11.4
27	1478	0.335	34.7	*34.0	10.6	20.7
28	1494	0.400	*40.2	19.0	28.4	12.4
29	1497	0.566	16.4	22.6	*56.8	4.2
30	1496	0.155	*15.5	23.3	20.8	40.4

*Note.* \*denotes the correct response; 304 missing data points (.7) are distributed across the items, this explains the variation of  $n$ .

Some distractors are not drawing enough attention. For example, distractor “D” on Item 1 only attracted 0.7% of the responses, which affects the estimation of parameters under the NRM model. Results should be interpreted with caution for options with low frequency count. For items that have adequate weight of responses across the options, the NRM can yield interesting patterns in regards to option response probability based on examinee ability.

Table 33 below provides the frequency in percentages for each of the options for the items in Form B and the asterisks denote the correct response to the item. Of special interest, Items 17, 26, 27, and 30 contain distractors that draw too much attention.

Table XXXIII Option Frequencies for Items in Form B

Item	$n$	Item Difficulty	Options			
			A	B	C	D
1	2107	0.862	5.9	*86.7	5.4	2.0
2	2097	0.583	14.6	12.3	14.3	*58.8
3	2084	0.687	9.6	10.9	9.7	*69.8
4	2094	0.683	11.9	8.5	*69.1	10.5
5	2100	0.722	6.7	*72.9	5.1	15.4
6	2111	0.806	4.5	*80.8	10.0	4.7
7	2102	0.821	7.9	4.6	4.8	*82.7
8	2110	0.720	5.2	*72.3	15.7	6.8
9	2116	0.734	12.8	8.7	*73.4	5.0
10	2100	0.569	12.3	8.1	*57.4	22.1
11	2109	0.655	*65.8	5.1	21.9	7.3
12	2112	0.730	3.1	14.9	*73.2	8.8
13	2103	0.632	10.9	11.8	*63.7	13.6
14	2112	0.551	*55.2	14.3	12.2	18.4
15	2111	0.540	15.1	17.6	13.2	*54.2
16	2096	0.406	9.7	*41.0	35.7	13.6
17	2110	0.314	10.7	*31.5	27.0	30.9
18	2113	0.477	31.6	14.8	*47.8	5.8

Table XXXIII continued

19	2108	0.593	15.5	14.1	10.9	*59.5
20	2109	0.424	10.1	*42.6	11.8	35.6
21	2103	0.501	11.7	*50.5	22.3	15.5
22	2106	0.224	*22.5	38.1	24.1	15.3
23	2081	0.468	21.6	15.7	15.1	*47.6
24	2111	0.484	16.4	*48.6	25.8	9.2
25	2106	0.589	16.0	*59.2	14.9	9.9
26	2109	0.371	41.9	*37.3	15.4	5.5
27	2110	0.336	11.1	*33.7	43.2	12.0
28	2110	0.260	22.2	13.0	38.7	*26.1
29	2105	0.290	*29.2	32.3	29.9	8.6
30	2108	0.145	*14.6	22.6	23.9	38.9

*Note.* \*denotes the correct response; 387 missing data points (.6) are distributed across the items, this explains the variation of  $n$ .

For example, distractor “D” on Item 17, attracted almost 31% of the examinees’ responses and the key drew 31.5% of the responses. Similar issues are evident in Items 26, 27 and 30. Item 30 was the same item on both test forms and the results in option frequency were very similar. That is, 3619 examinees processed the item in the same way. Option “D” still attracted most of the examinees attention across test forms and the key also was the least chosen option. Items that exhibited these issues are further explored through graphics in the NRM modeling section.

#### 4.3.2 Data modeling with the nominal response model

A set of slope ( $\lambda$ ) and intercept ( $\zeta$ ) parameters were estimated for each response option within each item on both test forms of the TCF under the NRM where the slope parameter can be interpreted as a conventional discrimination parameter (de Ayala, 2009) as in the 2PL model and the intercept parameter reflects the frequencies (or popularity) of the options. To examine parameter invariance, the model was fit to the data, each item at a time, across gender. To correct for type I error, a Bonferroni adjustment was adopted using the conventional alpha level (0.05) and dividing it by the number of items on the test (i.e., 30), which was equivalent to the repeated iterations conducted with the model. This yielded an alpha level of 0.001. This section only includes a selection of graphics that were judged interesting to discuss

Appendix C provides the intercept and slope parameters for both males and females across the test forms under analysis in this study. The zeta and lambda parameters are subscripted from 1 to 4 and these subscripts represent the item options “A” through “D”. The log-likelihood ratio test revealed that parameters were not invariant across gender in Item 19 ( $\chi^2 = 24.090$ ,  $df = 6$ ,  $p$

= .001) and Item 23 ( $\chi^2 = 22.194$ ,  $df = 6$ ,  $p = .001$ ) in Form A. And for Form B, the likelihood ratio test suggested that the NRM parameter invariance across gender for Item 13 ( $\chi^2 = 17.129$ ,  $df = 6$ ,  $p = .009$ ) and Item 17 ( $\chi^2 = 18.847$ ,  $df = 6$ ,  $p = .004$ ) should be interpreted with caution. Since parameter invariance was fairly stable across most of the items in both test forms, the graphics included in this section were generated from the most parsimonious model, which combined both males and females in a single analysis that required less number of parameters (i.e., 8 parameters per item instead of 16 parameters per item due to gender). The parameters from this estimation are also included on Appendix C.

Figure 37 below illustrates how the NRM might be used to understand how examinees along the continuum of the trait (i.e., L2 listening comprehension as defined in the TCF) are interacting with Items 1, 14, 26, and 30. For example, the item category response curves (ICRCs) for Item 1 revealed that the probability of choosing Option C (i.e., the keyed response) increases as the examinee ability increases. The distractors are merely chosen suggesting that the item is too easy. Option “D” is generally flat across the ability continuum and is very unlikely to be chosen even by the least able students. Psychometrically, this item could be discarded because does not provide enough information about the examinees and at least Option D should be revised. However, from a conceptual stand point, language testers usually opt to keep this type of item at the beginning of a test to reduce anxiety and ease examinees into more challenging items.

The ICRCs for Item 14 depict that Option C (the keyed response) increases monotonically as examinee ability increases. Less able examinees have an increasing tendency to choose Option “B”, with a slightly lower probability of choosing Option “D”. Option “A” is the least chosen at lower levels of listening ability, but at around -0.5 ability level this option has a higher probability to be chosen than Option B, which is monotonically decreasing as examinee ability increases. Examinees with an ability around zero had about 40% probability of choosing the correct answer, 30% probability of choosing Option D, 20% of choosing Option A, and 10% of choosing Option B. Looking at the parameter values for the parsimonious NRM model in Appendix C, in conjunction with the ICRCs for each response option for Item 14 is very informative. For example, the intercept ( $\zeta$ ) parameters shows the keyed response “C” is also the most popular response.

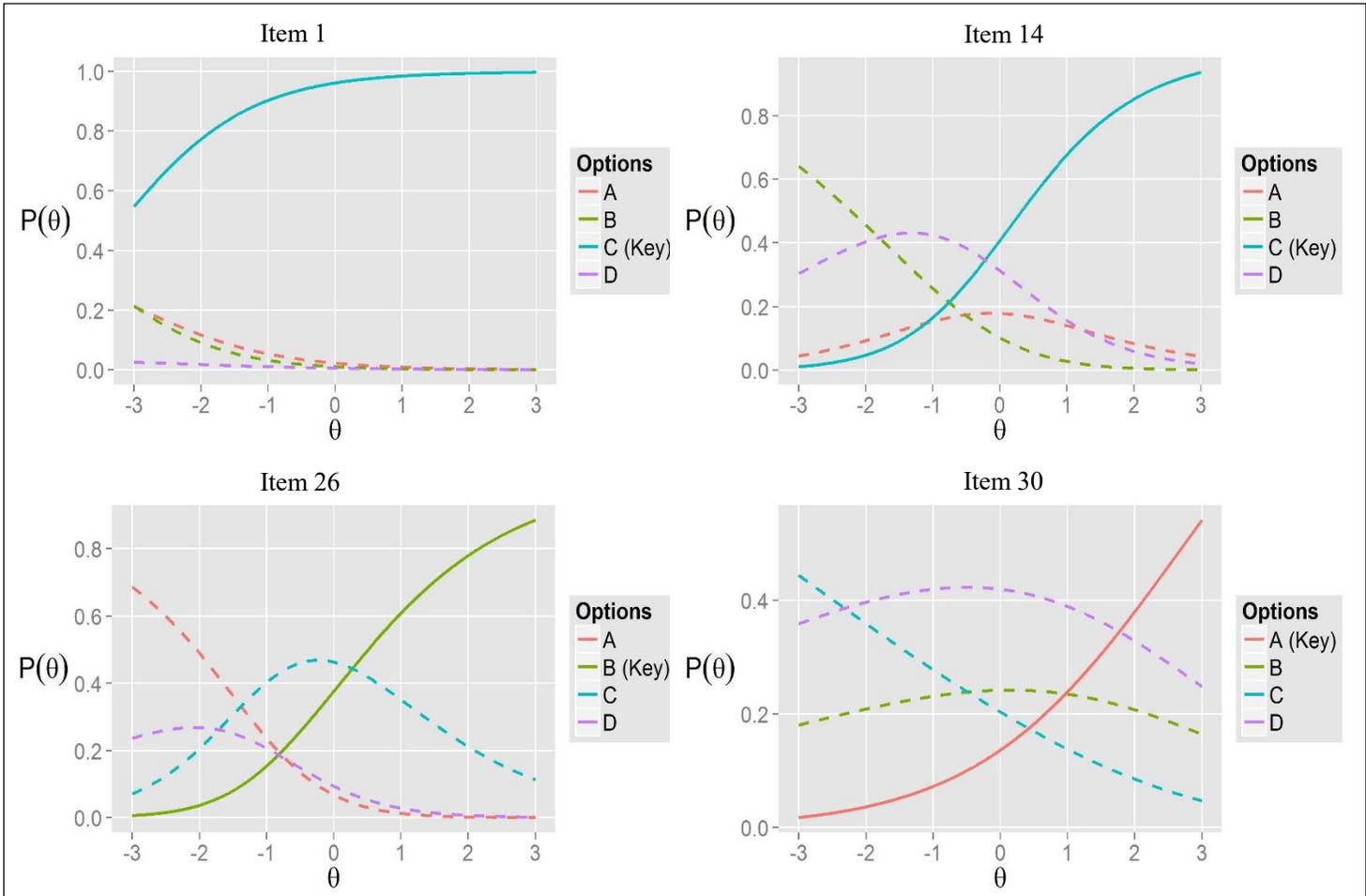


Figure 37 NRM Item Category Response Curves for Items 1, 14, 26, and 30 in Form

In terms of implied category order, the slope ( $\lambda$ ) parameters, suggest that the high ability examinees selected the keyed response “C”, with Option “B” the next most difficult category, followed by Options “D” and “A”. Item 14 provides an example of a desirable item to have on the TCF L2 listening assessment. Conversely, Items 26 and 30 highlight some issues. Option “C” (i.e., a distractor) on Item 26, was the most chosen ( $\zeta_A = -0.990$ ;  $\zeta_B = 0.724$ ;  $\zeta_C = 0.932$ ;  $\zeta_D = -0.666$ ). Examinees with an ability level of zero still have a higher probability of choosing Distractor “C”. It is also counterintuitive that as ability increases, the probability of choosing Distractor “C” increases and this trend stops at about 0.2 logits of ability level. Item 30 also depicts some serious problems as all distractors were more popular than the keyed response ( $\zeta_A = -0.524$ ;  $\zeta_B = 0.048$ ;  $\zeta_C = -0.124$ ;  $\zeta_D = 0.600$ ) and only very high ability examinees ( $\theta = +2$ ) chose the keyed response “A” over the distractors.

Figure 38 below provides the ICRCs for Items 1, 21, 27, and 30 for Form B. The item category response curves (ICRCs) for Item 1 revealed that the probability of choosing Option B (i.e., the keyed response) increases as the examinee ability increases. The ICRCs for the distractors are monotonically decreasing as examinee ability increases. The probability of choosing Option “B” (i.e., the key response) increases rapidly at lower levels of theta ( $\sim -1.9$ ), suggesting that the item is fairly easy. On Item 21, the keyed response “B” also increases monotonically as theta increases and Options A and D were the least chosen ( $\zeta_A = -0.423$ ;  $\zeta_B = 0.931$ ;  $\zeta_C = -0.015$ ;  $\zeta_D = -0.492$ ) displaying similar intercept parameters. However, along the continuum of examinee ability, Option “A” was flatter than Options “C” and “D”.

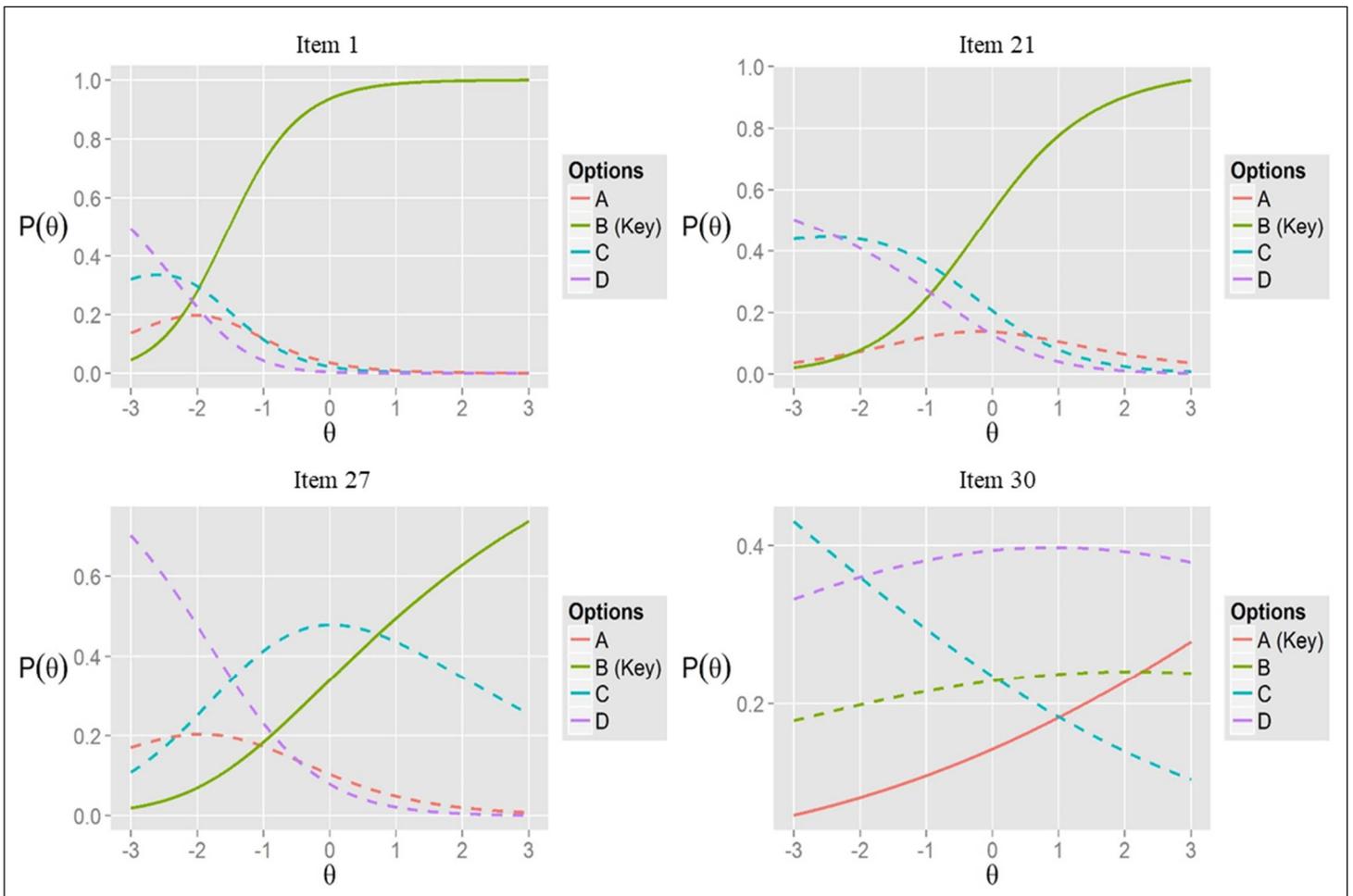


Figure 38 NRM Item Category Response Curves for Items 1, 21, 27, and 30 in Form B

The intercept parameters for Item 27 ( $\zeta_A = -0.613$ ;  $\zeta_B = 0.578$ ;  $\zeta_C = 0.920$ ;  $\zeta_D = -0.885$ ) suggest that Distractor “C” was more frequently chosen than the keyed response (“B”). It is not until higher levels of examinee ability ( $\theta \geq + 0.85$ ) that examinees tended to choose the keyed response over Distractor “C”. Results show that Item 30 is very problematic and although it was repeatedly administered on both test forms, the intercept and slope parameters were very similar as noted in the item option frequency. A content analysis of the options for Item 26 (in Form A) and Item 27 (in Form B) revealed that these items were potentially double keyed. For instance, Item 26 presented a dialogue between two interlocutors, a potential tenant searching for a house to rent and a real estate agent. The potential tenant listed the specifications of the house to rent, but two of those specifications are in the options of the item (Option “B”, the key and Option “C”, a plausible key as well). Similarly, Item 27 presented a dialogue between an interviewer and an interviewee on the topic of music piracy and the deficit this has created to the music industry. The problem here was that the options “B” *deficit in the industry* and “C” *records are copied on the internet* are in fact plausible answers to the stem of the item. Finally, in regard to Item 30, the stem is too long and complex and since the stem are not presented in print to the examinees, this might be a source of cognitive load.

The following chapter discusses and organizes the main findings into a validity argument for the TCF in the context of Quebec immigration. In turn, this seeks to substantiate the claims in the interpretation/use argument made on behalf of the test.

## Chapter 5: Discussion

Different types of analyses were carried out to gather the evidence to support the warrants for the scoring, generalization, and the explanation inferences underlining the interpretation/use argument for the TCF listening test scores in the context of Quebec immigration. This chapter discusses the main findings of these studies and subsequently present the validity argument for the listening component of the TCF using Toulmin's ([1958], 2003) model of inference.

The interpretation of the findings related to listening comprehension in language assessment is done within a rather limited framework, as the literature on CFA and DIF studies in listening comprehension tests is scanty and in the case of the nominal response model is nonexistent. Thus, when applicable, CFA and DIF discussions are connected to findings from general studies that have researched other language skills. For the NRM results, findings are related to the general literature in educational and psychological measurement.

### 5.1 Implications of CFA Results Substantiating Research Question One

This section discusses the findings of confirmatory factor analysis. The CFA study addressed the following research question:

1. What listening sub-skills do the *Test de Connaissance du français* assess? Do test items contribute exclusively to the assessment of the listening construct? Is there supporting evidence indicating that test items are contaminated by construct irrelevant variance or construct underrepresentation?

Although listening comprehension is a complex process involving both bottom up and top down processes (Field, 1999) as well as the ability to understand inferred meanings, interpret illocutionary intentions, and draw conclusions (Buck, 2001; Field, 2013), according to the panel of experts and the confirmatory factor analyses, the listening component of the TCF as operationalized in the test forms under study assesses primarily the understanding of explicit aural discourse. In this vein, before item deletion, 24 items in Form A as well as 19 items in Form B targeted predominantly the ability to *comprehend explicit aural discourse*. The panel also suggested that both test forms targeted the ability to infer implicit ideas and both factors (*inferring implicit ideas*) were comprised of 6 indicators before purification (Items 8, 12, 19, 25, 28 and 29 in Form A and Items 10, 14, 23, 24, 25, 26 in Form B). In addition, the panel

judged that five items (Items 11, 18, 19, 20, and 27) targeted the ability to *understanding the main idea or general topic* in Form B, but considered that neither test form targeted the ability to *understand specific details*, which potentially undermines acoustic lexical recognition, an important feature of listening comprehension (Field, 2003). The distribution of the identified subskills was not equivalent across the two test forms, questioning the representation of the L2 listening construct and posing a threat to the validity of test score-based interpretations and uses of the test (Messick, 1989). This is problematic because ultimately the interpretations of test scores generally assume that the construct is represented adequately across test forms and that important cognitive processes of the construct under assessment are covered. In this regard, based on the recommendations of the panel and the results from the CFA, equivalent construct representation is not warranted in the test forms under study. The findings presented here are also consistent with previous research which has found that some test forms of high-stakes L2 listening assessments rely heavily on items that mostly target the understanding explicitly stated information (Aryadoust, 2013; Geranpayeh & Taylor, 2008). The four-subskill listening comprehension model identified in the literature (Buck, 2001; Field, 2013; Ockey, 2013; Rost, 1990) was partially operationalized in the test forms, underrepresenting the complexity of the mental processes involved in listening comprehension.

In the CFA analyses, all items loaded significantly in their specified factors, but according to the adopted rule of thumb of structural loading (Thompson, 2004), some factor loadings (i.e., items 1, 5, 27, 29 and 30 in Form A and items 2, 3, 22, 26, 28, 29, 30 in Form B) were weak ( $\lambda < .400$ ), suggesting a source of construct irrelevant variance. This was substantiated in the Rasch analysis as the data from these items did not conform to the Rasch model expectations. Similar to previous CFA studies in L2 listening assessment (Aryadoust, 2013) and language assessment in general (Bae & Bachman, 1998; Sawaki, 2007; Yoo & Manna, 2015), the inter-factor correlations were high ( $r = \geq .925$ ), indicating a strong relationship between the subskills of L2 listening comprehension. In this regard, the correlated factor models have important implications regarding the listening construct operationalized in the TCF. First, the high correlations among the hypothesized latent listening subskills support reporting a single score for the listening dimension of the test. Second, after more equivalent test forms are assembled, inter-factor correlations should be monitored to ensure that score reports are consistent with the factor structure of the test.

Conceptually, it can be argued that listening comprehension as operationalized in the TCF is a divisible construct into constituent parts which has implications for test specifications and test development. This is similar to previous studies that have sought to explore the factor structure of language tests (Bae & Bachman, 1998; Sawaki, Stricker & Oranje, 2009) and have found that language subskills are actually divisible. A higher order model was specified to explore the high correlation among the L2 listening subskills in the TCF. However, the higher order CFA model did not converge. This resonates with In'nami & Koizumi (2011) who specified higher order CFA models, but the models were not superior to correlated CFA models.

In terms of methodology, the tetrachoric correlation matrix and the robust WLS estimator used in the estimation of the CFA models yielded stable parameters across both test forms. This is consistent with previous research that has modelled binary data with relative large sample sizes and have found that the stability of parameters is warranted under such conditions (Moshagen & Much, 2014). This has implication for users of CFA methodology in the field of language assessment.

The findings from the CFA analysis have implications for language schools that offer preparation courses for the TCF in the context of Quebec immigration. Given that the test emphasizes the understanding of explicit aural discourse, language schools and personal language tutors might opt to emphasize this skill and neglect the complexity of the listening construct. This can be considered a negative side effect or *washback* (Alderson & Wall, 1993) stemming from the test.

## **5.2 Complementarity of Confirmatory Factor Analysis and Rasch Modeling**

Although measurement models are very instrumental for data modeling in validation research, researchers are ultimately responsible for making important decisions that can impact the results. As an example, in the CFA section, a rather conservative rule of thumb was adopted to flag potentially problematic items (Thompson, 2004, [ $\lambda \geq 0.400$ ]), which resulted in the deletion of four items in Form A (Items 1, 5, 27, 29 and 30) and seven items in Form B (Items 2, 3, 22, 26, 28, 29 and 30). These items were also problematic in the Rasch modeling study since most items were flagged as underfitting or noisy, however, Item 1 ( $\lambda = 0.377$ ), which has been flagged as problematic in the CFA analyses actually fit the Rasch model (Outfit mean-square = 0.91) given the adopted rule of thumb for item deletion (Wright & Linacre, 1994, [fit

range = 0.80 - 1.20]). With the exemption of Item 1, the adopted guidelines for item screening were coherent in both CFA and Rasch when considering outfit mean-square statistics. Rule of thumbs have implications for both factor modification and data-model fit in CFA and Rasch modelling respectively because depending on how conservative or liberal the guidelines are, researchers can arrive at quite different conclusions. Thus, careful attention must be paid when adopting arbitrary rules of thumbs to screen problematic items in language measures.

Furthermore, in CFA, factor loadings can be used as an indication of how much variance in one item can be attributed to the latent variable (Brown, 2015) and in Rasch modeling fit statistics indicate how accurately or predictably the data fit the model (Linacre, 2002). If the adopted guidelines for factor purification in CFA suggest that an item should be eliminated, but the Rasch analysis only yields muted or overfitting fit information for the item, perhaps the item should be kept in the analysis because it does not distort measurement. When Rasch measurement is used jointly with CFA, Rasch fit statistics provide a lower and upper bound to identify muted or noisy data which can be informative for CFA. It is also worth noting that the complementarity of CFA and the classical Rasch model highly depends on the dimensionality of the data. If both analysis suggest a single factor model (or high inter-factor correlations), which was the case in this study, the Rasch model can be seen as a single factor model and comparison can be made more easily.

### **5.3 Implications of DIF Results for Research Question Two**

This section discusses the findings of the Rasch-based and the standardization index differential item functioning analyses related to the fairness of the TCF across gender, first language, age, and geographical region (African examinees only) of the examinees who took either of the test form under analysis. This study addressed the following research question:

2. Does the Test de Connaissance du français exhibit harmful differential item functioning (DIF), leading to test bias against gender, first language, age, and geographical location of examinees, threatening the validity of test score-based interpretations and uses of the test in the context of Quebec immigration?

DIF analysis across grouping variables (i.e., gender, L1, age, and geographical location of African examinees) yielded a number of test items that exhibited potential bias, especially under the L1 and geographical grouping variables. These results have important implications for the

*Centre International d'Études Pédagogique (CIEP)* in order to monitor potential threats related to test fairness.

### **5.3.1 Unidimensionality and local independence in Rasch-based DIF**

Ensuring model fit before conducting IRT-based or Rasch-Based DIF study is a condition that must be satisfied in order to draw valid inferences from the results (Sireci & Rios, 2014). There have been discussions on the precondition of unidimensionality in the context of language assessment and the conclusion seem to point towards multidimensionality (e.g., Buck, 1994). However, these discussions, by the most part, refer to conceptual dimensionality and not to psychometric dimensionality. Conceptually, it is useful to consider language constructs as multidimensional because authenticity is enhanced by conceptualizing test items that reflect real life language conditions and various mental processes in the case of the receptive skills. However, if data are found to be unidimensional from psychometric analyses, one can gain insight into the assessment of language skills. Moreover, the high inter factor correlation also provided support of unidimensional data. Furthermore, the Rasch model has been robust to minor violation of dimensionality (Boadé, 2013). In addition, previous research has suggested that after the Rasch dimension has been extracted, in order for dimensionality to hold, the first contrast of the principal component analysis (PCA) of the standardized Rasch residuals should not be greater than 2 (Raïche, 2005). This was met in Form A ( $\lambda = 2.057$ ) and marginally violated in Form B ( $\lambda = 2.169$ ), before underfitting items were excluded from further Rasch analyses. Given that misfit can have an adverse impact on DIF analyses, underfitting items were sequentially dropped, which also caused the eigenvalues of the first contrast on both test forms to decrease, suggesting that the data were even more appropriate for Rasch analysis. However, even after underfitting items had been dropped, misfit was induced in other items, and as noted earlier, this calls for the judgement of the researcher who needs to decide when to stop deleting items to warrant construct representation.

In regard to the assumption of local independence the correlation matrix of the standardized residuals did not yield absolute inter-item correlations either before or after dropping noisy items that would require further attention. Thus, the assumptions of the unidimensional the Rasch model were met, which is important for the DIF analysis.

### **5.3.2 DIF investigation implications for gender**

The political discourse regarding Quebec immigration promotes diversity and inclusion (MIDI, 2017); therefore, the TCF should yield results that are fair for women and men seeking to settle permanently in Quebec. In Form A, two items favored each subgroup: Item 16 favored women and Item 23 favored men. A content analysis revealed that, potentially, Item 16 favored women because of speech perception. In this regard, the listening script presented a dialogue where a female voice was the prevalent interlocutor and this could have triggered inattention in the male subgroup. Item 23, favoring men, was also attributed to an exchange between two interlocutors where the male voice was predominant. In this vein, Cramer (1977) suggested that speech perception can play a key role especially in initial encounters between individuals and can affect comprehension. However, these are just hypotheses that require further investigation. Perhaps it might be important to have male and female voices fairly represented across listening language tests. An alternative explanation can be attributed to inattention, or guessing by low ability examinees on both items (Aryadoust, Goh & Lee, 2011). Form B did not exhibit any DIF that required attention. Another discernable pattern across gender was that the picture recognition items did not favor any gender subgroups.

### **5.3.3 DIF investigation implications for first language (L1)**

Given the plethora of nationalities seeking to immigrate to Canada through the province of Quebec, it is obvious that examinee populations can exhibit various language backgrounds. The results from the L1 DIF across both test forms yielded results that were difficult to associate to potential source of bias. Similar to previous research in L1 DIF studies (e.g., Abbott, 2007; Allalouf & Abranzon, 2008) the findings in this section of the DIF investigation are only hypotheses. For example, for items where French speaking examinees had a higher probability of correct response, it stands to say that this was probably a consequence of impact and not bias. The TCF is a French proficiency test and it is expected that native or native like speakers of French as is often the case of Arabic speaking examinees in Africa have a higher probability of correct response than L2 speakers. A question that follows is why, in Items 1, 12, 13, 18, 19, 24, and 25 in Form A, was this trend reversed? A probable explanation to this is twofold: 1) in general these items were not too difficult for native speakers of native like speakers of French and this might have prompted some inattention when answering these items. 2) The L2 speakers

of Spanish used listening strategies (Goh, 1997, 2000) that either favored their performance (Lundsteen, 1993) or successfully guessed (Aryadoust *et al.*, 2011) the correct answer to the items that challenged L2 speakers of French. Moreover, native or native like speakers of the target language of the assessment might not be used to L2 assessments and this might hinder their performance. These are hypotheses that were not addressed in this study, but suggestions are provided in the conclusion chapter.

#### **5.3.4 DIF investigation implications for age groups**

DIF analysis under the matching variable age under the classification of *ministère d'Immigration, Diversité et Inclusion* (MIDI) flagged four items for Form A and only one item for Form B. The picture recognition feature of some of the items on both test forms tended to favor the youngest examinees. This has implications for testing programs whose target populations are younger audiences. The aural input for Item 24 in Form A required understanding a constructive critique of a literary writer, which can be considered to some extent a required specialized background knowledge needed to respond correctly to the item, and divorced from the test construct. This might have disadvantaged the youngest (18 to 29 years of age) group (Banerjee & Papageorgiou, 2016). This has implications for test development as items should be carefully crafted to ensure that specialized background knowledge is not required to respond to test items.

In order to collect further evidence regarding the consistency of the findings for the items flagged for DIF under the age grouping policy implemented at MIDI, a different grouping criterion that considered sample size across age subgroups was in place. This has implications for DIF studies that oftentimes make use of unequal or small sample sizes across group comparisons. Following Sireci and Rios' (2013) as well Zumbo's (1999) recommendations for the design of DIF studies the age grouping policy adopted at MIDI was partially supported. In this regard, the items that were flagged for DIF under Form A were also flagged under the DIF design that prioritized sample size instead of arbitrary grouping criteria. However, two additional items were also flagged under these analyses. Similarly, DIF on Item 2 for Form B disappeared, and a new item was flagged for DIF. Thus, it should be noted that unequal and small sample sizes across age groups affect the precision of the person ability and item difficulty estimates of the Rasch model and consequently the DIF estimates can be unstable.

Research into the listening processes (Geranpayeh & Kunnan, 2007) have suggested to examine DIF in terms of the cognitive processes that the subskill demands. However, the content analysis of the DIF items under age groups did not revealed such patterns. This calls for rigorous mixed methods methodologies when conducting DIF studies as this can improve our understanding underlying cognitive processes and sources of bias. For instance, examinees' verbal protocols or free recall protocols (Buck, 1991; Vandergrift, 2007) can shed light onto the potential item bias, by offering insights on how examinees of different age groups process listening comprehension items that exhibit DIF.

### **5.3.5 DIF investigation implications for geographical location**

A large number of North (Algeria, Morocco, and Tunisia) and West (Cameroon, Ivory Coast, and Senegal) African examinees took the TCF for potential immigration to Quebec. The analyses flagged very large DIF across several items in both test forms, but a source of potential bias could not be associated based on item content. An interesting trend, however, suggested that the first and easiest items of the test in both test forms tended to favor the North African examinees and the last, say, ten items, tended to favor the West African examinees. And as noted in Pae (2004) large value of DIF may also indicates that an item measures an additional construct since items under DIF analysis can measure an auxiliary dimension differently across test taker groups (Geranpayeh & Kunnan, 2007). This suggests that the first items on both test forms generally tended to exhibit DIF against West African examinees because they assessed an auxiliary dimension that the other items did not. Similarly, the last items on both test forms generally tended to exhibit DIF against North African examinees because they assessed an auxiliary dimension that the other items did not. Additional work doing verbal protocols with examinees who have North and West African backgrounds can shed light onto this perplexing pattern.

Another implication from the regional DIF analyses can be associated to the accent because strong accentedness can impact the difficulty level of the items (Ockey, Papageorgiou & French, 2016). A French accent (from France) was used across some of the longer listening items. In this vein, items that consisted of longer listening input tended to be more difficult than picture recognition, short dialogues, and mini-conversation items. West African examinees tended to be more successful on these items. Note, however, that we cannot make strong claims

based on the aural input on these test forms as accentedness was not assessed with an accent strength scale (e.g., Ockey & French, 2016). An implication that follows from this is in line with the idea of paying careful attention to strength of accents and look for neutral ones when designing listening comprehension assessments (Ockey, Papageorgiou & French, 2016).

### **5.3.6 Implications for the interpretation of DIF analyses**

The results from the DIF analysis suggested potential issues regarding speech perception (Kramer, 1977), specialized content (Banerjee & Papageorgiou, 2016) and accent (Ockey, Papageorgiou & French, 2016). However, it was difficult to associate a source of potential bias to all the items that were flagged for DIF. This is consistent with the pervasive statistical limitation of DIF methodologies – the use of a measure of ability that is internal to the test itself, results in measures of DIF that are ipsitive (Penfield & Camili, 2006), circular (Camilli, 1993) or artificial (Andrich & Hagquist, 2012, 2015). For example, the presence of bias in the focal group (e.g., West African examinees) yields a situation whereby reference (e.g., North African examinees) group members having the same test score will not share the same expected ability level. In this regard, the focal group members will have a higher expected ability than the matched reference group members. As a result, the non-biased items of the test will display a small level of DIF favoring the focal group. In other words, DIF observed against the reference group depends upon the impact of the biased items on the focal group test score distribution. There have been efforts to address the circularity problem of DIF methodologies by using external covariates to match the test takers and explain causal DIF (Liu *et al.* 2016), or dissolving items into two separate items (Andrich & Hagquist, 2012, 2015). However, further research is required using these procedures with language assessment data to gain more insight into the ipsitivity issue of DIF analysis in language assessment.

Other alternatives can include a focus mixed-method DIF methodology that uses both statistical modeling and verbal protocols elicited from the subgroup of interest, but this requires that advances in DIF analysis (e.g., techniques used to flag artificial DIF) be used in tandem with the targeted focus groups doing the verbal protocols.

## 5.4 Implications of NRM Results for Research Question Three

This section discusses the findings nominal response model analysis related to the usefulness of multiple choice items in the assessment of listening comprehension in a second language. This study addressed the following research question:

3) What information do the distractors and the keyed option of selected-response (SR) items (multiple-choice) provide to support the validity argument for the use of selected-response items to assess listening comprehension in a second language?

In this study the nominal response model ([NRM], Bock, 1972) was fit to the data across gender to gain insight into the stability of item parameters. Eventually this warranted the interpretation of the results stemming from this measurement model. These results have important implications for the *Centre International d'Études Pédagogique* in the area of test development –namely, the quality of multiple choice (MC) items (including both the keyed response and the distractors). Given that research applying the NRM model to language data is essentially nonexistent, the results are discussed in relation to the small body of research that exist on data modeling within an NRM framework.

### 5.4.1 Implications of the option frequency of MC items

There were a few items in both test forms where the distractors were highly attractive. For example, on Item 30, which was the same item across the two test forms, the key (Option A) was the least chosen option. This item yielded very weak factor loadings, did not fit the Rasch model, and the item category response curves (ICRCs) were very problematic. Although less pronounced, this was also the case for Item 27 on both test forms. These results are informative for the development stage of TCF items at the CIEP. The descriptive analyses also have implication for data modelling, data screening and parameter interpretation under the NRM framework. For instance, parameters for options with very low frequency should be interpreted with caution in NRM modelling or excluded from the analysis to warrant model parameter stability. Option frequency can be of great help interpreting the model parameters since the intercept parameter ( $\zeta$ ) in the NRM model essentially represents the popularity of the response category (Ostini & Nering, 2006).

### 5.4.2 Implications of the NRM modelling

Unlike item analysis in classical test theory the NRM provides person parameter estimates (or person thetas) and the probability of choosing an option at different levels of the targeted construct can be readily observed in graphics. This often reveals what options are most attractive to examinees with low, moderate, and high levels of the target trait (Penfield, 2014). This can be instrumental in getting insight about the type of distractors that draw examinees' attention based on their level of ability. Drawing on this information, the quality of the distractors can be improved during test development – by avoiding distractors that are unlikely to be chosen across the ability continuum. In addition, when the intercept parameter of the keyed response and a distractor or distractors approximate each other in value, it is convenient to verify this information along the trait continuum as this can shed light onto options-trait interactions and double-keyed items.

Broader implications are also applicable to the diagnostic assessment of communicative competence of potential immigrants. In this vein, language assessments, crafted for diagnostic purposes, can be analyzed within an NRM framework to shed light onto the implied option ordering based on ability level. This is revealed in the slope ( $\lambda$ ) parameter and as a result, the options can be carefully developed to provide information about the language development stage of examinees. However, this would require a close collaboration between the disciplines of language testing and second language acquisition.

In this study, model parameters recovery was assessed by examining the normality of the latent trait distribution or person parameter estimates (Johnson & Reise, 2014). The distribution of the person parameters for Form A was close to normal, but the distribution of person parameter estimates for Form B was positively skewed. This can bias the estimation of the intercept and slope parameters and caution in their interpretation is advised.

The following section incorporates the results from the CFA, DIF, and NRM studies into the validity argument of the *Test de connaissance du français* in the context of immigration in the province of Quebec.

#### 5.4 Validity Argument of the Test de connaissance du français

Argument-based approaches to test score validation has been criticized (e.g., Borsboom & Markus, 2013), and the dominant view of validity has been referred to as an incompatible theory concerned with *interpretation* and *uses* of test scores (Cizek, 2012). However, from a pragmatic point of view, argument-based validity is an enticing approach to validate the interpretation and uses of test scores. In fact, major testing programs (e.g., *Educational Testing Service*) have framed their validation work within argument-based approaches to test score validation (e.g., Chapelle *et al.*, 2008).

The validity argument for the TCF evaluates a global statement or claim: “The TCF assesses important listening comprehension skills across all the examinees that opt to take the test. The test is fair and do not favor examinees in regard gender, first language, age, geographical location, or nationality. Finally, the method of the test format ensures that all the important skills of L2 listening comprehension are captured and that the format of the test does not create problems in test score generation”. This statement is assumed in the description of the test provided in the websites of the *ministère d’Immigration, Diversité et Inclusion* (MIDI) and the *Centre International d’Études Pédagogique* (CIEP). For instance:

In addition to optimizing the *Ministère’s* selection practices, standardized tests and diplomas recognized by the *Ministère* ensure that the evaluation of candidates for their French knowledge is accurate, consistent and fair (MIDI, 2018).

Similarly, the CIEP advocates that:

The TCF is developed for anyone, regardless of their nationality and native language, who wishes to begin permanent immigration procedures with the Quebec Ministry of Immigration, Diversity and Inclusion (CIEP, 2018).

Building a validity argument for the interpretation and uses of test scores requires providing warrants and backings for all the inferences in the interpretation/use argument (Chapelle *et al.*, 2008; Kane, 2006, 2013). The inferences are considered as bridges that link one inference to the next until the validity argument is complete. This normally culminates with the decision inference or conclusion of the argument. The number of inferences in a complete validity argument can include: scoring, generalization, explanation, extrapolation and utilization inferences. This study provided warrants and backings for the scoring, generalization, and

explanation inferences supporting the global statement or claim made on behalf the TCF L2 listening scores. The extrapolation and decision inferences need to be addressed to complete the validity argument. Recent research in language testing has exclusively examined, using corpus linguistics, the extrapolation inference in the validity argument of high stakes language test (Laflair & Staples, 2017). Thus, the presentation of the first three inferences for the validity argument of the TCF only means that more research is needed to complete the validity argument for the interpretation and uses of test scores in the context of Quebec immigration.

Figure 39 provides the inferences, warrants and backings of the validity argument for TCF in the context of Quebec immigration. As noted earlier, this study provides backing exclusively for the scoring, generalization and explanation inferences. In later sections of this chapter ideas are presented on how data may be collected to address the extrapolation and decision inferences to complete the argument and although consequences are not generally treated as an inference, they are important to include whenever test scores can significantly impact the lives of individuals involved in the testing process. On Figure 39 below, each inference calls for a certain kind of data analysis and the conclusion of each inference serves as a bridge to the next inference (Kane, 2001). Each inference is supported by warrants, which in turn relies on assumptions that need to be verified and supported by evidence (Chapelle *et al.*, 2008; Kane, 2013). For instance, the warrant for the scoring inference is stated as follows: “test performance data for the TCF is standardized and the scoring criteria is accurate. This warrant assumes that: a) test development and administration have been standardized, b) important listening comprehension skills are targeted in the assessment and c) answer keys are accurate and applied appropriately. Each assumption is supported by one or more backings, which strengthens the plausibility of the warrants. Warrants support the inferences underlying the global claims or statements about the test scores. Rebuttals challenge the plausibility of the warrants and can weaken the underlying statements or claims of the validity argument. As noted in Aryadoust (2013) the assumptions reflected in the scoring inference would be supported by backings stating that the CIEP has properly collected these data, experts have judged the test items as tapping into important L2 listening skills, and that the scoring rules are accurate and applied appropriately (Kane, 2006).

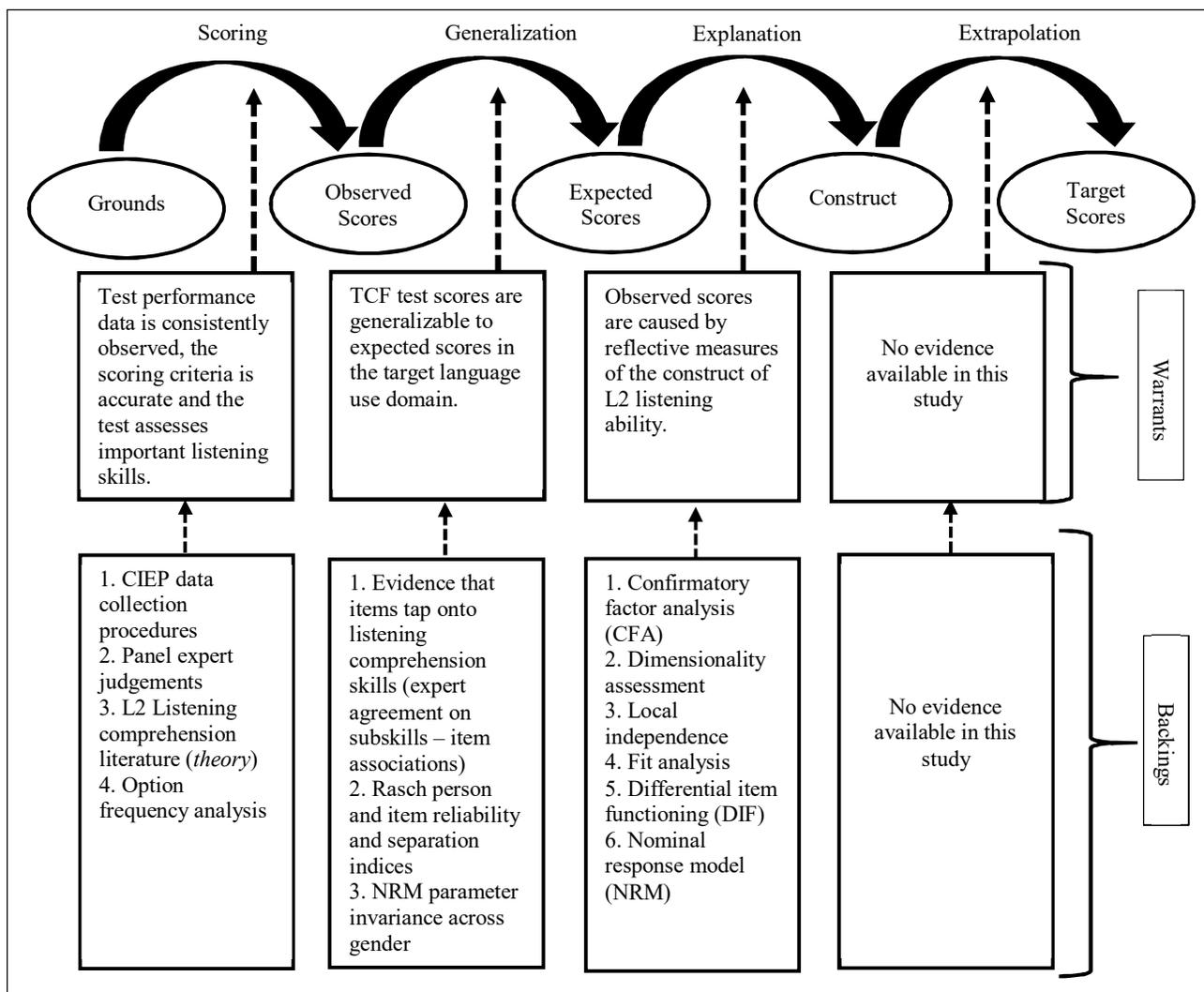


Figure 39 Validity Argument of the *Test de connaissance du français* Listening Test

The following sections provide the warrants, assumptions and backings, and rebuttals for each of the inferences underlying the validity argument for the TCF drawing on the findings of the studies carried out in this study.

#### 5.4.1 Scoring inference

Figure 40 provides the warrants, assumptions, backings, and rebuttals for the scoring inference. Two important assumptions are addressed here:

- i) The TCF assesses listening skills that are important for comprehending real life aural input in French in the province of Quebec.

ii) Answer keys are accurate and applied appropriately.

The scoring evidence reported in the findings in Chapter four is key to provide the backings and rebuttals that underline the assumptions of the warrants for the scoring inference. This inference must be warranted by evidence suggesting that the listening subskills operationalized in the TCF are important features of L2 listening comprehension and that the scoring criteria is accurate and used appropriately across test administrations.

The panel of experts thoroughly analyzed the content of the test items and made judgements regarding the listening subskill each test item appeared to assess. To our knowledge, the *Centre International d'Études Pédagogiques* (CIEP) does not have documented procedures available to the public regarding this evidence. Thus, the item-subskills associations in the validity argument for the TCF depend heavily on the judgement of the panel, which in fact was supported in the CFA analysis.

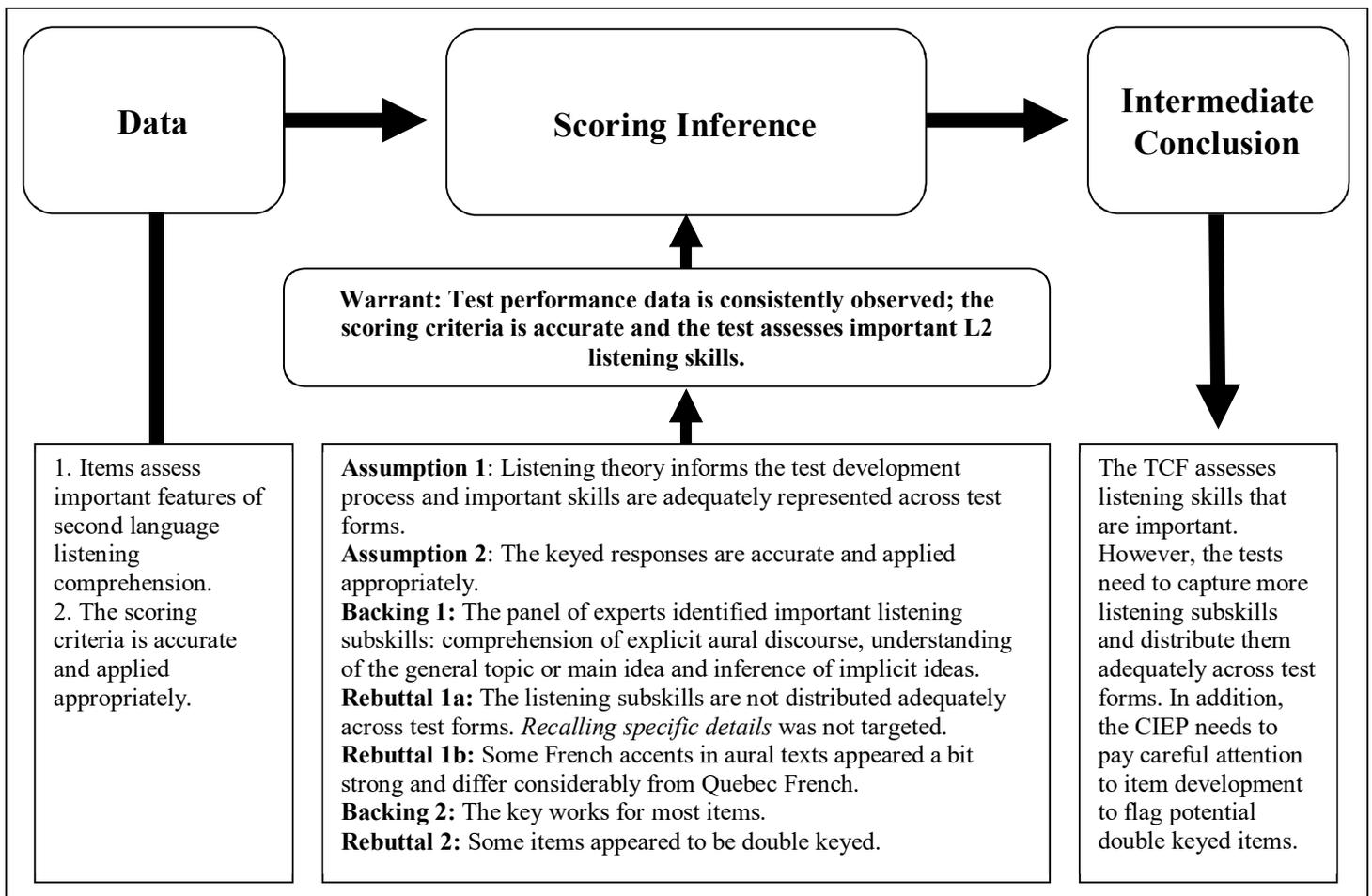


Figure 40 Scoring Inference of the TCF Listening Test with Warrants, Assumptions, Rebuttals, and Backings

The panel of experts also judged that the test assesses important features of listening comprehension, but mostly associated with explicit aural discourse and this can potentially underrepresent the targeted construct of the test (Messick, 1989). A rebuttal on the same figure refers to the accentedness (Ockey, *et al.*, 2016) of the aural texts in the TCF. Most of these texts are tainted with a French (from France) accent. In turn, this might undermine the actual variety of accents of the test taker population that generally takes the TCF and the accents from Montreal and rural areas of Quebec. This has implication at two levels: 1) it is desirable to reflect the variety of accents of the target population to enhance the fidelity of the test, but 2) it is also important to include neutral Quebec accents since examinees are seeking to immigrate to Quebec and it is in the best of interests of the immigration authorities to obtain valuable information regarding the language ability of potential immigrants. The keyed responses are accurately used since the TCF listening test is machine scored, but as noted in one of the rebuttals on Figure 41, two items (Item 26 in Form A and Item 27 in Form B) appeared to be double keyed.

#### **5.4.2 Generalization inference**

The generalization inference is based on the warrant that TCF scores are estimates of expected scores that examinees would obtain on similar tasks and test forms (Chapelle *et al.*, 2008). Figure 41 below illustrates the warrants, assumptions, backings, and rebuttals for the generalization inference. The warrant assumes that sufficient items have been included in the test, that examinees can be classified into distinguishable strata of ability levels; that items are reliable and their hierarchy was successfully estimated with the targeted sample of examinees, and that the parametrization for multiple choice items is stable across female and male examinees. Backing for these assumptions was obtained through Rasch person and item separation indices as well as parameter invariance across gender through the nominal response model. Chapter four provides further details for the backings of these assumptions. In a nutshell, the Rasch person and item separation index suggested that regardless of the items that were dropped, both test forms were able to classify examinees into different ability levels. Similarly the item difficulty hierarchy was warranted by the high item separation indices. The parametrization of the NRM model yielded stable intercept and slope parameters across gender.

This provided the evidence needed to support the assumption that the options in multiple choice items across TCF test forms perform similarly across gender.

The generalization inference interpret test scores as samples from a universe of observations (Markus & Borsboom, 2013). Thus, it is of paramount importance to define the type of items that underline the *theory* of the construct under assessment. It stands to say that if the assumption of the generalization inference strongly relies on samples from a universe of observations, when the test is operational, the samples (in this case items) from the universe of observations should be coherent with, and represent adequately, the mental processes contained in the delineated universe. This has implication for task and test specifications prior to the development of tests. That is, specification documents should be sufficiently defined to create parallel tasks and test forms that reflect a representative sample of the universe of observations.

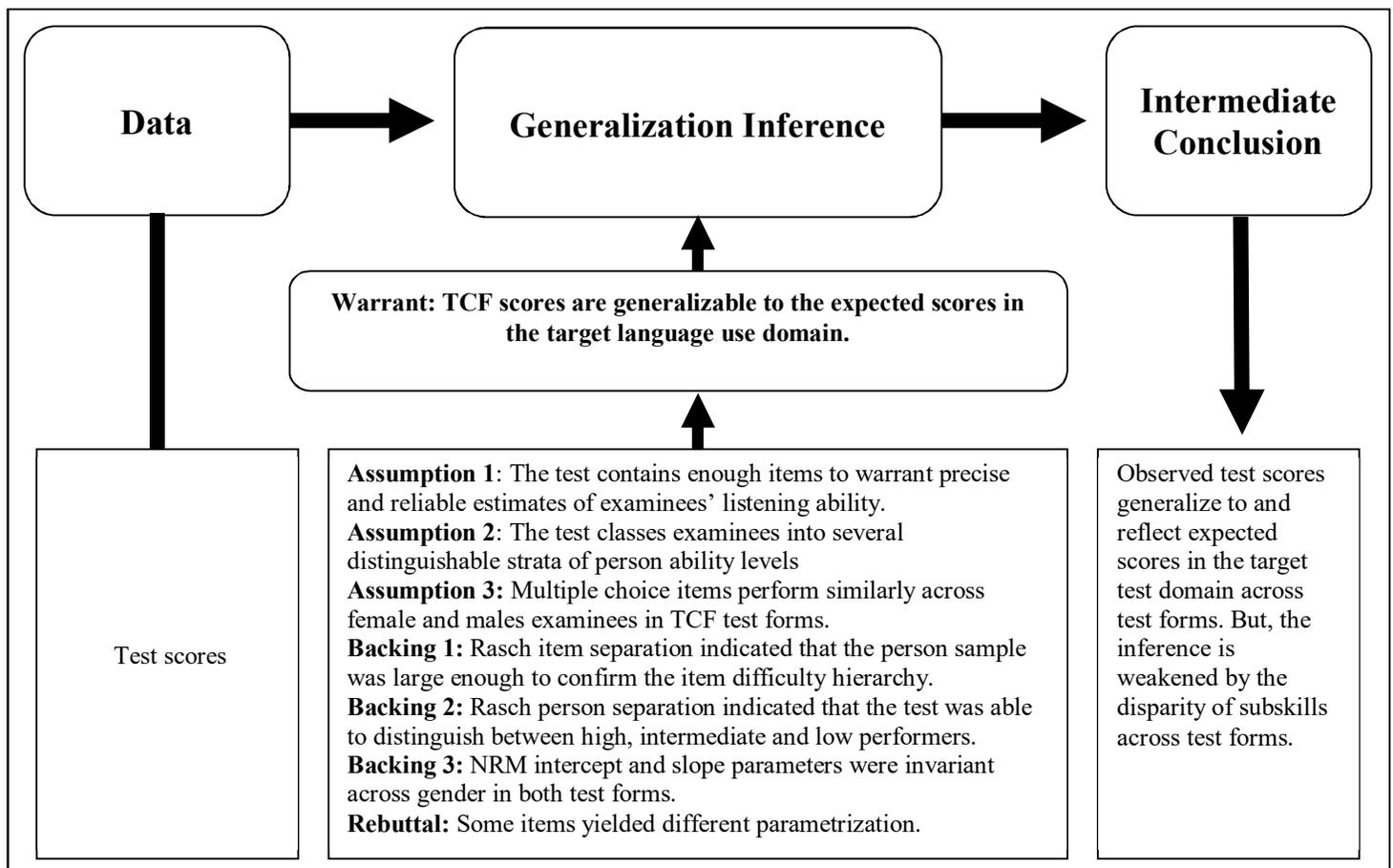


Figure 41 Generalization Inference of the TCF Listening Test with Warrants, Assumptions, Rebuttals, and Backings

Although the person and item reliability and separation indices were acceptable, they are only applicable to the samples from the universe that were captured in the TCF test forms. As the results from Chapter four showed, the samples items from the universe of observation is not adequately distributed within or between test forms. This weakens a bit the generalization inference.

### **5.4.3 Explanation inference**

The explanation inference concludes that expected scores are attributable to the theorized latent trait under assessment (Chapelle *et al.*, 2008). In this vein, the explanation inference (Figure 42 below) assumes that test scores are attributable to the listening construct under assessment, that the TCF is free from sources of construct irrelevant variance (CIV) and construct underrepresentation, and that the test is fair across test takers subgroups.

The warrants supporting this inference also assume that observed scores are associated with listening comprehension subskills that define the L2 listening construct and that construct is operationalized similarly across test takers subgroups of interest. The backings for these assumptions were collected through CFA and DIF analyses. The evidence supporting these assumptions yielded mixed results that might weaken the warrant of the inference. In this regard, the panel members ascertained that items in both test forms primarily targeted the comprehension of explicit aural discourse and to a lesser extent both forms assessed the ability to make inferences. However, the panel also judged that only Form B contained items that assessed the understanding of the general topic or main idea and that none of the test forms targeted the understanding of specific details. In addition, the identified subskills were disproportional distributed in both test forms. As noted in the discussion, this resonates with previous research regarding the explanation inference of a high stakes test of academic listening (Aryadoust, 2013; Geranpayeh & Taylor, 2008), thus attenuating its strength.

In terms of construct irrelevant variance, the evidence for this assumption comes from different analyses and it is bit attenuating. The CFA yielded item loadings (but not many) that were weak, suggesting the targeting of other construct. In Rasch modelling, a variety of items did not conform to the model expectation and to a lesser extent the assumption of dimensionality was slightly violated. This suggest sources of irrelevant variance, but to pinpoint the sources is more challenging. A potential hypothesis for Rasch analysis is that since native speakers of

French also took the test, items that are too easy for these examinees might induce inattention. Then, when high performance test takers show erratic Guttman patterns, the Rasch model signals this as noise.

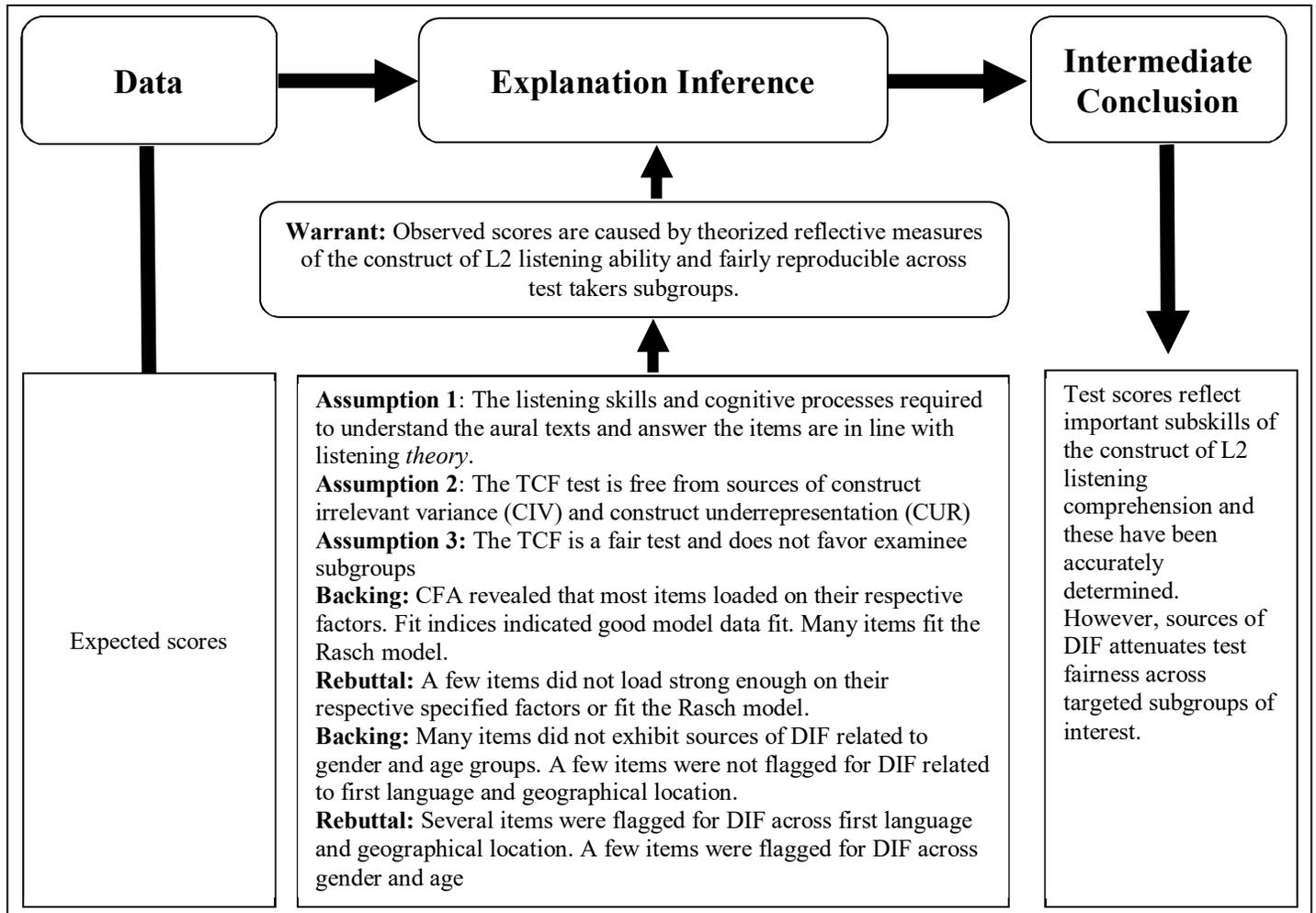


Figure 42 Explanation Inference of the TCF Listening Test with Warrants, Assumptions, Rebuttals, and Backings

In addition, the DIF analyses flagged a few items for gender and age DIF, but many items were flag for DIF across first language and geographical location of African examinees. The explanation inference is weakened by these findings because the intended listening subskills underlying the TCF could not be invariant across the population of examinees. This assertion might be relaxed when comparing native or native like speakers of French to the L2 examinees. We tend to attribute the difference in performance to impact and not to item bias. However, this assertion of impact becomes problematic when L2 examinees depict higher probability of

correct response to a reasonably simple item. There are inherent issues in DIF methodologies (see the ipsitivity problem described in Chapter 4), but this evidence weakens the explanation inference of the validity argument of the TCF

### **5.4.3 Remarks of the extrapolation and decision inferences**

Although this study did not include the extrapolation or the decision inferences due to logistical limitations, these two components are necessary to complete the validity argument for the TCF and some directions are provided in the next chapter on how to collect the required evidence to support the warrants for the extrapolation and decision inferences. This study has presented the initial steps to validate the interpretation and uses of TCF listening scores in the context of Quebec immigration, and ideas on how to complete it are detailed in the conclusion chapter. The argument-based approach to validation appears to be an enticing approach that can provide rich and structured information regarding the language ability of prospective immigrants to key stakeholders involved in immigration decisions.

## Chapter 6: Conclusion

### 6.1 Summing up of the Validity Argument

This study examined the validity of the interpretation and uses of test scores from the listening component of the *Test de connaissance du français* used for immigration purposes in the province of Quebec. In the journey to Quebec immigration, other factors play a pivotal role in the selection of potential immigrants. The most important and determinant factors for successful applications include: *level of education, work and/or professional experience, age, and language ability (connaissance linguistiques)*. Although language is not the sole criterion, it does play an important role in high stakes decisions made on behalf of applicants who seek to settle permanently in the province. The consequences of testing can actually affect people's lives and this merits careful attention by the key stakeholders who are responsible, or involved either directly or indirectly, in the decision making process of immigrant applications for permanent residence.

Validity evidence is rather limited in languages other than English and in fact there is a constant need to validate the interpretation and uses of test scores across several languages in various contexts. The validity argument for the TCF suggests that it would be appropriate to work on the distribution of listening subskills operationalized in the test to enhance the quality of the data collected with this instrument. In addition, the contextual uses of TCF scores – used across several backgrounds – prompted the study of the test fairness and this signaled specific recommendations to ensure test fairness (e.g., script accentedness and speech perception). Finally, the method data collection (i.e., multiple choice items) suggested that at least two items were potentially double keyed and needed further attention.

### 6.2 Overview and Appraisal of the Research Questions

This study evaluated three research questions, and each of these questions required a different set of methodologies. The questions were stated as follows:

1. What listening sub-skills do the *Test de Connaissance du français* assess? Do test items contribute exclusively to the assessment of the listening construct? Is there supporting evidence indicating that test items are contaminated by construct irrelevant variance or construct underrepresentation?

The TCF listening comprehension test from two operational test forms was found to underrepresent the construct of L2 listening by emphasizing the assessment of the understanding of explicit aural discourse. Important elements of L2 listening such as inferring implicit ideas were present in the test forms, but minimally represented. Drawing on the item-subskill associations recommended by the panel of experts, the identified subskills were disproportional distributed across test forms. Some items were judged to be double keyed and this caused problems in data modelling procedures (e.g., CFA, DIF, and NRM). In addition, some test were suspect of construct irrelevancies (i.e., misfitting items and weak factor loadings). These issues can be readily addressed in the test development process and throughout the stages of testing in order to collect evidence to strengthen the current validity argument of the test. Unlike previous CFA research into the construct of L2 listening comprehension (e.g., Aryadoust, 2013), the CFA models were specified based on the recommendation of twelve experts in the field of applied linguistics, this has implication for CFA validity research which need to move away from simplistic methods when assigning mental processes to items in reading or listening assessments. Moreover, the construct of L2 listening in the TCF could be better represented by developing clear guidelines in terms of item development and test assembly so that relevant subskills are targeted and distributed adequately across test forms.

2. Does the *Test de Connaissance du français* exhibit harmful differential item functioning (DIF), leading to test bias against gender, first language, age, and geographical location of examinees, threatening the validity of test score-based interpretations and uses of the test in the context of Quebec immigration?

The TCF was found to exhibit DIF across very few items in terms of gender and age of the examinees. But, a vast amount of DIF items were flagged across first language and geographical background. For the most part, it was difficult to associate a source of bias to the DIF items and this constitutes one of the limitations of the DIF study. Some DIF could be easily attributed to impact and circularity problems of DIF methodologies, but the question requires further research.

3. What information do the distractors and the keyed option of selected-response (SR) items (multiple-choice) provide to support or attenuate the validity argument for the use of selected-response items to assess listening comprehension in a second language?

Under NRM modeling, multiple choice items were found to yield useful information regarding the ability level of the examinee and their tendencies to choose a given option in the item. This measurement model, thus far, constitutes a pioneering effort in data modelling in the field of language testing and assessment and in applied linguistics in general.

In summary, the findings served to provide backings for the assumptions underlying the warrants supporting the inferences for the validity argument of the TCF. The following section elaborates on the conclusions for the validity argument of the TCF.

### **6.3 Validity Argument of the Test de connaissance du français**

The TCF validity argument proceeds across connected validity inferences, and seeks to arrive at a conclusion regarding the interpretation and uses of test scores (Chapelle *et al.*, 2008; Kane, 1992, 2001, 2006, 2013) in the context of Quebec immigration. In this study, evidence was provided to substantiate the scoring, generalization and explanation inferences in the validity argument.

The scoring inference provided evidence for the use of the scoring criteria and the examination of item content to determine the sub-skill underlying the item. Agreement coefficients and reliability indices provided the backings for the assumptions underlying the warrants under this inference. Rebuttals were also identified and were judged to provide counter evidence against the scoring inference (the potential double keyed items for example). The panel data also served to specify the CFA models in the explanation inference.

The generalization inference provided evidence for the stability of item difficulty hierarchy, person class stratification, and NRM parameters across gender. Results suggested that similar results should be expected if conditions were similar in other testing situations. Nevertheless, when rebuttals are identified in previous inferences they weaken the bridges that link one step of the argument to the next. For instance, the panel members identified that the items did not tap onto all subskills proportionally across test forms. Thus, it follows that in the generalization inference, those test forms, if supported by reliability analyses, only generalize to the same conditions they represent – generalization can only be made to similar situation and in this case are contaminated with construct underrepresentation.

The Explanation inference provided evidence for the reproducibility of the theorized listening construct in the data. This inference connects the expected scores to the construct under

assessment. In the context of language assessment and in educational in general, where content is very important, the explanation inferences becomes the most significant element of the validity argument (Kane, 2012). Many items loaded onto their specified factors and adequate model-data fit was achieved, thus supporting the inference. However, the Rasch model flagged a few items as underfitting or noisy, which distort the measurement system (Wright & Linacre, 1994). In addition, the rebuttals voiced in the scoring inference, which carried through the generalization inference, hinder the explanation inference because the under representation of listening subskills across test forms weakens it.

#### **6.4 Limitations of the Study**

The limitations of the study include caveats that apply to the argument-based validity framework, the measurement models and the absence of qualitative analysis of important aspects of the listening test. Following this order, since mostly secondary data were used to build the validity argument, indispensable (data such as test taker criterion measures) were not available to address the extrapolation inference in the argument. And although validity argument frameworks can be instrumental to build a case for the use of test scores, they require an intensive research program, which sometimes may require longitudinal data to substantiate the key inferences in the argument. This can be unfeasible in certain contexts where it is challenging to follow research participants for an adequate period of time. In addition, this study was conducted a posteriori or post-hoc, which is not ideal because decisions had been already made on behalf of examinees who applied to immigrate to the province of Quebec. In a real situation, validation studies take place a priori test use.

In terms of the measurement models, the CFA analyses were specified following the suggestion of a panel of experts and heavily rely on the judgement of the panel when perhaps a different panel could have recommended other subskills-item associations. In addition, the study only examined two test forms and given the results, it is unwarranted to generalize the results to other TCF forms. Person fit was ignored in the Rasch analyses preceding DIF. The DIF analysis also presented limitations related to the circularity problems of the methodology and given the multi-cultural diversity of the examinees the difficulty to pinpoint potential sources of bias was heighten and more complex. Also, non-uniform DIF was not considered, which could have yielded interesting results. Furthermore, the NRM data-modelling was potentially affected by

small frequency counts in some of the item options, yielding parameters that are unreliable. In addition, other researchers could have chosen and have preferred different measurement models to analyze the data, which can lead to a different, but not necessarily contradictory way of structuring the validity argument. For example, a researcher that favors cognitive diagnostic measurement models would have built a Q-matrix with the experts and considered that one item tapped onto more than one listening subskill, making the argument from a multidimensional point of view.

In addition, the listening scripts were not analyzed qualitatively, thus neglecting important key issues such as accentedness, vocabulary, and appropriateness of the test register in the context of Quebec. The use of secondary data also hampered other potential data collection such as examinee interviews.

### **6.5 Recommendation for Future Research**

The scantiness of validity arguments for language tests other than English calls for validation studies that justify the interpretation and uses of test scores in other languages. It is important to examine the results provided in this study and collect more evidence for the replicability of results across other test forms of the TCF for immigration purposes to Quebec. In addition, evidence to support the extrapolation and decision inferences is required to culminate the inferential steps in the validity argument of the TCF. For the extrapolation inference, criterion evidence is needed to establish the relationship or power of prediction of TCF listening scores. Even though it is almost impossible to obtain criterion measures when using secondary data, there are other ways when the extrapolation inference can be substantiated. For instance, potential immigrants are required to submit test scores when applying for specialized jobs, in this regard, potential employers could be interviewed or surveyed about the adequacy of the test for the job in question. The decision inference is more political and would require the collaboration of the MIDI, CIEP, and language testing educators to establish defensible cut scores that are linked to the weight attributed to language proficiency in Quebec's immigration laws.

Content analysis of the test items was neglected in this study, but such studies can yield important information about the linguistic elements that might contribute to the difficulty of items in the assessment of listening in French. Listening comprehension has gained more

attention in the past few decades and should continue gathering momentum in the future. Thus, more research is needed to better understand the underlying mental processes that interlocutors or passive listeners experience when decoding aural discourse and stream of sounds. Other research can relate to the TCF particularly. The stems of the listening items that are not in a picture recognition format are not presented in print in the test booklets. Some research could examine the effects of this modality as longer listening texts can be harder to process.

In order to enhance the DIF studies, sensitive review panels who share cultural and demographic information similar to the DIF groups under study can be more effective in identifying item bias. On a methodological level, given the circularity problems of DIF methodologies, splitting the item with the largest DIF and rerunning the analysis can help flag artificial DIF that is induced by the item with largest DIF (Andrich & Hagquist, 2015). Non-uniform DIF can also provide other interesting perspectives in this line of research. Finally, The NRM model can provide insightful results in language assessment, but when using this model it is important to use large sample sizes to obtain robust model parameters.

## References

- Abbott, M. L. (2007). A confirmatory approach to differential item functioning on an ESL reading assessment. *Language Testing*, 24(1), 7-36. doi:10.1177/0265532207071510
- Alderson, J. C., & Wall, D. (1993). Does washback exist? *Applied Linguistics*, 14(2), 115-129. doi:10.1093/applin/14.2.115
- Allalouf, A., & Abramzon, A. (2008). Constructing better second language assessments based on differential item functioning analysis. *Language Assessment Quarterly*, 5(2), 120-141. doi:10.1080/15434300801934710
- Altman, D. G. (1991). *Practical statistics for medical research*. London, UK: Chapman and Hall/CRC.
- American Educational Research Association., American Psychological Association., & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.
- Anderson, A., & Lynch, T. (1988). *Listening*. New York, NY: Oxford University Press.
- Andrich, D. (1978). Scaling attitude items constructed and scored in the Likert tradition. *Educational and Psychological Measurement*, 38(3), 665-680. doi:10.1177/001316447803800308
- Andrich, D., & Hagquist, C. (2012). Real and artificial differential item functioning. *Journal of Educational and Behavioral Statistics*, 37(3), 387-416. doi:10.3102/1076998611411913
- Andrich, D., & Hagquist, C. (2015). Real and artificial differential item functioning in polytomous items. *Educational and Psychological Measurement*, 75(2), 185-207. doi:10.1177/0013164414534258
- Angoff, W. H. (1972). Use of difficulty and discrimination indices for detecting item bias. In R. A. Berk (Ed.), *Handbook of methods for detecting test bias* (pp. 96-116). Baltimore, MD: Johns Hopkins University Press.
- Angoff, W. H. (1988). Validity: An evolving concept. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 19-32). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Aryadoust, V. (2009). Mapping Rasch-based measurement onto the argument-based validity framework. *Rasch Measurement Transactions*, 23(1), 1192-1193.

- Aryadoust, V. (2012). Differential item functioning in while-listening performance tests: The case of the international English language testing system (IELTS) listening module. *International Journal of Listening*, 26(1), 40-60. doi:10.1080/10904018.2012.639649
- Aryadoust, V. (2013). *Building a validity argument for a listening test of academic proficiency*. Newcastle, UK: Cambridge Scholars Publishing.
- Aryadoust, V., Goh, C. M., & Kim, L. O. (2011). An investigation of differential item functioning in the MELAB listening test. *Language Assessment Quarterly*, 8(4), 361-385. doi:10.1080/15434303.2011.628632
- Aryadoust, V., Goh, C. M., & Lee, K. O. (2012). An investigation of differential item functioning in the MELAB listening test. *Language Assessment Quarterly*, 8(4), 361-385.
- Bachman, L., & Palmer, A. (1996). *Language testing in practice*. New York, NY: Oxford University Press.
- Bachman, L., & Palmer, A. (2010). *Language assessment in practice*. New York, NY: Oxford University Press.
- Bachman, L. F. (2005). Building and supporting a case for test use. *Language Assessment Quarterly*, 2(1), 1-34. doi:10.1207/s15434311laq0201\_1
- Baddeley, A. D. (1986). *Working memory*. Oxford, UK: Oxford University Press.
- Bae, J., & Bachman, L. F. (1998). A latent variable approach to listening and reading: testing factorial invariance across two groups of children in the Korean/English Two-Way Immersion Program. *Language Testing*, 15(3), 380-414. doi:10.1177/026553229801500304
- Bae, J., & Bachman, L. F. (2010). An investigation of four writing traits and two tasks across two languages. *Language Testing*, 27(2), 213-234. doi:10.1177/0265532209349470
- Baghaei, P. (2008). Local dependency and Rasch measures. *Rasch Measurement Transactions*, 21(3), 1105-1106.
- Baker, F. B. (1993). Equating tests under the nominal response model. *Applied Psychological Measurement*, 17(3), 239-251. doi:10.1177/014662169301700305
- Balch, T. W. (1909). French colonization in North Africa. *The American Political Science Review*, 3(4), 539-551. doi:10.2307/1944685

- Banerjee, J., & Papageorgiou, S. (2016). What's in a topic? Exploring the interaction between test-taker age and item content in high-stakes testing. *International Journal of Listening*, 30(1-2), 8-24. doi:10.1080/10904018.2015.1056876
- Batty, A. O. (2015). A comparison of video- and audio-mediated listening tests with many-facet Rasch modeling and differential distractor functioning. *Language Testing*, 32(1), 3-20. doi:10.1177/0265532214531254
- Bertrand, R., & Blais, J. G. (2004). Du concept de validité. In R. Bertrand & J. G. Blais (Eds.), *Modèles de mesure: L'apport de la théorie des réponses aux items* (pp. 237-276). Sainte Foy, QC: Presses de l'Université du Québec.
- Birnbaum. (1968). Some latent trait models and their use in inferring an examinee ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 395-479). Reading, MA: Addison-Wesley.
- Blackledge, A. (2009). "As a country we do expect": The further extension of language testing regimes in the United Kingdom. *Language Assessment Quarterly*, 6(1), 6-16. doi:10.1080/15434300802606465
- Boadé, G. (2013). *Robustesse du modèle de Rasch unidimensionnel*: International Book Market Service Limited.
- Bock, D. R. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, 37(1), 29-51. doi:10.1007/bf02291411
- Bollen, K. A. (1984). Multiple indicators: Internal consistency or no necessary relationship? *Quality and Quantity*, 17(4), 377-385. doi:10.1007/BF00227593
- Bollen, K. A., & Lennox, R. (1991). Conventional wisdom on measurement: A structural perspective. *Psychological Bulletin*, 110(2), 305-314. doi:10.1037/0033-2909.110.2.305
- Bolli, G. G. (2013). Migration policies in Italy in relation to language requirements. The project Italiano, lingua nostra: Impact and limitations. In E. D. Galaczi & C. Weir (Eds.), *Exploring language frameworks: Proceedings of the ALTE Kraków conference, July 2011* (pp. 45-61). Cambridge, UK: Cambridge University Press.
- Bond, T. V., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences* (2 ed.). New York, NY: Taylor & Francis Group.

- Bond, T., & Fox, C. M. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences* (3 ed.). New York, NY: Routledge.
- Borsboom, D. (2016). Zen and the art of validity theory. *Assessment in Education: Principles, Policy & Practice*, 23(3), 415-421. doi:10.1080/0969594X.2015.1073479
- Borsboom, D., Cramer, A. O. J., Kievit, R. A., Scholten, A., & Franic, S. (2009). The end of construct validity. In R. W. Lissitz (Ed.), *The concept of validity: Revisions, new directions, and applications* (pp. 135-170). Charlotte, NC: Information Age Publishing.
- Borsboom, D., & Markus, K. A. (2013). Truth and evidence in validity theory. *Journal of Educational Measurement*, 50(1), 110-114.
- Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The concept of validity. *Psychological Review*, 111(4), 1061-1071.
- Brennan, R. L. (2001). An essay on the history and future of reliability from the perspective of replications. *Journal of Educational Measurement*, 38(4), 295-317. doi:10.1111/j.1745-3984.2001.tb01129.x
- Brennan, R. L. (Ed.). (2006). *Educational measurement* (4 ed.). Westport, CT: American Council of Education and Praeger.
- Brindley, G. (1998). Assessing listening abilities. *Annual Review of Applied Linguistics*, 18, 171-191. doi:10.1017/S0267190500003536
- Brindley, G. (2008). Educational reform and language testing. In E. Shohamy & N. H. Hornberger (Eds.), *Language testing and assessment: Encyclopedia of language and education* (Vol. 7). New York, NY: Springer Science.
- Brown, G. (1977). *Listening to spoken English*. London, UK: Longman.
- Brown, G., & Yule, G. (1983). *Teaching the spoken language*. Cambridge, UK: Cambridge University Press.
- Brown, T. A. (2015). *Confirmatory factor analysis for applied research* (2 ed.). New York, NY: Guilford Press.
- Browne, M. V., & Cudek, R. (1993). Alternative ways of assessing model fit. In K. Bollen & J. Long (Eds.), *Testing equation structural models* (pp. 136-162). Newbury Park, CA: Sage.
- Buck, G. (1990). *The testing of second language listening comprehension* (Unpublished doctoral thesis). University of Lancaster, Lancaster, UK.

- Buck, G. (1991). The testing of listening comprehension: An introspective study. *Language Testing*, 8(1), 67-91. doi:10.1177/026553229100800105
- Buck, G. (1992). Listening comprehension: Construct validity and trait characteristics. *Language Learning*, 42(3), 313-357. doi:10.1111/j.1467-1770.1992.tb01339.x
- Buck, G. (1994). The appropriacy of psychometric measurement models for testing second language listening comprehension. *Language Testing*, 11(2), 145-170. doi:10.1177/026553229401100204
- Buck, G. (2001). *Assessing listening*. Cambridge, UK: Cambridge University Press.
- Cai, H. (2013). Partial dictation as a measure of EFL listening proficiency: Evidence from confirmatory factor analysis. *Language Testing*, 30(2), 177-199. doi:10.1177/0265532212456833
- Camilli, G. (1993). The case against item bias detection techniques based on internal criteria: Do item bias procedures obscure test fairness issues? In P. W. Holland & H. Wainer (Eds.), *Differential item functioning*. Hillsdale, NJ: Lawrence Erlbaum.
- Camilli, G. (2006). Test fairness. In R. Brennan (Ed.), *Educational measurement* (4 ed., pp. 221-256). Westport, CT: American Council of Education and Praeger.
- Carrol, J. B. (1983). Psychometric theory and language testing. In J. W. Oller (Ed.), *Issues in language testing research* (pp. 80-107). Rowley, MA: Newbury House Publishing.
- Celce-Murcia, M., & Olshtain, E. (2000). *Discourse and context in language teaching: A guide for language teachers*. Cambridge, UK: Cambridge University Press.
- Chapelle, C. A. (1999). Validity in language assessment. *Annual Review of Applied Linguistics*, 19, 254-272. doi:10.1017/S0267190599190135
- Chapelle, C. A. (2012). Validity argument for language assessment: The framework is simple.... *Language Testing*, 29(1), 19-27. doi:10.1177/0265532211417211
- Chapelle, C. A., Cotos, E., & Lee, J. (2015). Validity arguments for diagnostic assessment using automated writing evaluation. *Language Testing*, 32(3), 385-405. doi:10.1177/0265532214565386
- Chapelle, C. A., Enright, M. A., & Jamieson, J. M. (2008). *Building a validity argument for the test of English as a foreign language*. New York, NY: Routledge.
- Chapelle, C. A., Jamieson, J., & Hegelheimer, V. (2003). Validation of a web-based ESL test. *Language Testing*, 20(4), 409-439. doi:10.1191/0265532203lt266oa

- Chen, Z., & Henning, G. (1985). Linguistic and cultural bias in language proficiency tests. *Language Testing*, 2(2), 155-163. doi:10.1177/026553228500200204
- Cheng, L., & Sun, Y. (2015). Interpreting the impact of the Ontario Secondary School Literacy Test on second language students within an argument-based validation framework. *Language Assessment Quarterly*, 12(1), 50-66. doi:10.1080/15434303.2014.981334
- Chi, Y. (2011). *Validation of an academic listening test: Effects of breakdown tests and test takers' cognitive awareness of listening processes* (Unpublished doctoral thesis). University of Illinois at Urbana-Champaign, Urbana-Champaign, IL.
- Chou, Y. T., & Wang, W. C. (2010). Checking dimensionality in item response theory models with principal component analysis on standardized residuals. *Educational and Psychological Measurement*, 70(5), 717-731. doi:10.1177/0013164410379322
- Chumbow, S. B., & Bobda, S. A. (2000). French in West Africa: a sociolinguistic perspective. In *International Journal of the Sociology of Language* (Vol. 141, pp. 39-60).
- Cizek, G. (2012). Defining and distinguishing validity: Interpretations of score meaning and justification of test use. *Psychological Methods*, 17(1), 31-43. doi:10.1037/a0026975
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2 ed.). Hillsdale, NJ: Erlbaum Associates.
- Colman, A. M. (2006). *Dictionary of psychology*. Oxford, UK: Oxford University Press.
- Cooke, M. (2009). Barrier or entitlement? The language and citizenship agenda in the United Kingdom. *Language Assessment Quarterly*, 6(1), 71-77. doi:10.1080/15434300802606580
- Cronbach, L. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational measurement* (2 ed., pp. 443-507). Washington, DC: American Council of Education.
- Cronbach, L. (1988). Five perspectives on validity argument. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 3-17). Hillsdale, NJ: Lawrence Erlbaum.
- Cronbach, L. (1989). Construct validation after 30 years. In R. L. Linn (Ed.), *Intelligence: Measurement theory, theory and public policy* (pp. 147-171). Urbana, IL: University of Illinois Press.
- Cronbach, L., & Meehl, P. (1955). Construct validity in psychological tests. *Psychological Bulletin*, 52(4), 281-302. doi:10.1037/h0040957

- Cudek, R. (1989). The analysis of correlation matrices using covariance structure models. *Psychological Bulletin*, 105(2), 317-327. doi:10.1037/0033-2909.105.2.317
- Cumming, A. (2013). Validation of language assessments. In C. A. Chapelle (Ed.), *The encyclopedia of applied linguistics* (pp. 1-10): Blackwell Publishing.
- Cureton, E. E. (1951). Validity. In E. F. Lindquist (Ed.), *Educational measurement*. Wanshington, DC American Council of Education.
- Davies, A. (1984). Validating three tests of English language proficiency. *Language Testing*, 1(1), 50-69. doi:10.1177/026553228400100105
- De Ayala, R. J., & Sava-Bolesta, M. (1999). Item parameter recovery for the nominal response model. *Applied Psychological Measurement*, 23(1), 3-19. doi:10.1177/01466219922031130
- DeMars, C. E. (2003). Sample size and the recovery of nominal response model item parameters. *Applied Psychological Measurement*, 27(4), 275-288. doi:10.1177/0146621603027004003
- DeMars, C. E. (2011). An analytic comparison of effect sizes for differential item functioning. *Applied Measurement in Education*, 24(3), 189-209. doi:10.1080/08957347.2011.580255
- Desbiens, D., Laurier, M. D., & Leroux, J. (2011). *Échelle québécoise des niveaux de compétence en français des personnes immigrantes adultes*. Montreal, QC: Canada.
- Desroches, F., Crendal, A., Renaud, F., Casanova, D., Demeuse, M., & Artus, F. (2006). *Les biais culturels dans les tests internationaux de français langue étrangère: Diagnostic et analyse*. Paper presented at the 19th conference entitled "Vers de nouvelles formes, modélisations et pratiques d'évaluation": Association pour le Développement des Méthodologies d'Évaluation en Éducation (ADMEE), Luxembourg: Luxembourg.
- DeVellis, R. F. (1991). *Scale development: Theories and applications*. Newbury Park, CA: Sage.
- Diamantopoulos, A., & Siguaw, J. A. (2006). Formative versus reflective indicators in organizational measure development: a comparison and empirical illustration. *British Journal Management*, 17(4), 263-282. doi:10.1111/j.1467-8551.2006.00500.x

- DiStefano, C. (2002). The impact of categorization with confirmatory factor analysis. *Structural Equation Modeling: A Multidisciplinary Journal*, 9(3), 327-346. doi:10.1207/S15328007SEM0903\_2
- DiStefano, C., & Hess, B. (2005). Using confirmatory factor analysis for construct validation: An empirical review. *Journal of Psychoeducational Assessment*, 23(3), 225-241. doi:10.1177/073428290502300303
- Doe, C. (2015). Student Interpretations of Diagnostic Feedback. *Language Assessment Quarterly*, 12(1), 110-135. doi:10.1080/15434303.2014.1002925
- Dorans, N. J., & Kulick, E. (1986). Demonstrating the utility of the standardization approach to assessing unexpected differential item performance on the Scholastic Aptitude Test. *Journal of Educational Measurement*, 23(4), 355-368. doi:10.1111/j.1745-3984.1986.tb00255.x
- Dörnyei, Z., & Katona, L. (1992). Validation of the C-test amongst Hungarian EFL learners. *Language Testing*, 9(2), 187-206. doi:10.1177/026553229200900206
- Downing, T. (2002). Threats to the validity of locally developed multiple choice tests in medical education: Construct irrelevant variance and construct underrepresentation. *Advances in Health Sciences Education*, 7(3), 235-241. doi:10.1023/A:1021112514626
- Dunkel, P. (1991). Listening in the native and second/foreign language: Toward an integration of research and practice. *TESOL Quarterly*, 25(3), 431-457. doi:10.2307/3586979
- Ebel, R. L. (1965). *Measuring educational achievement*. Englewood Cliffs, NJ: Prentice Hall.
- Eckes, T. (2015). *Introduction to many-facet Rasch measurement* (2 ed.). New York, NY: Peterlang.
- Edwards, J. R. (2011). The fallacy of formative measurement. *Organizational Research Methods*, 14(2), 370-388. doi:10.1177/1094428110378369
- Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum.
- Engelhard, G., Kobrin, J. L., & Wind, S. A. (2014). Exploring differential subgroup functioning on SAT writing items: What happens when English is not a test taker's best language? *International Journal of Testing*, 14(4), 339-359. doi:10.1080/15305058.2014.931281

- Enright, M. K., & Quinlan, T. (2010). Complementing human judgment of essays written by English language learners with e-rater® scoring. *Language Testing*, 27(3), 317-334. doi:10.1177/0265532210363144
- Farhady, H. (1983). On the plausibility of the unitary language proficiency factor. In J. W. Oller (Ed.), *Issues in language testing research* (pp. 11-28). Rowley, MA: Newbury House Publishing.
- Farnsworth, T. L. (2013). An investigation into the validity of the TOEFL iBT speaking test for international teaching assistant certification. *Language Assessment Quarterly*, 10(3), 274-291. doi:10.1080/15434303.2013.769548
- Ferne, T., & Rupp, A. A. (2007). A synthesis of 15 years of research on DIF in language testing: Methodological advances, challenges, and recommendations. *Language Assessment Quarterly*, 4(2), 113-148. doi:10.1080/15434300701375923
- Feyten, C. (1991). The power of listening ability: An overlooked dimension in language acquisition. *The Modern Language Journal*, 75(2), 173-180. doi:10.1111/j.1540-4781.1991.tb05348.x
- Fidalgo, A. M., Alavi, S., M., & Amirian, S. M. (2014). Strategies for testing statistical and practical significance in detecting DIF with logistic regression models. *Language Testing*, 31(4), 433-451. doi:10.1177/0265532214526748
- Field, J. (1999). Key concepts in ELT: Bottom up and top down. *ELT Journal*, 53(4), 338-339. doi:10.1093/elt/53.4.338
- Field, J. (2003). Promoting perception: Lexical segmentation in L2 listening. *ELT Journal*, 57(4). doi:10.1093/elt/57.4.325
- Field, J. (2004). An insight into listeners' problems: Too much bottom-up or too much top-down? *System*, 32(3), 363-377. doi:10.1016/j.system.2004.05.002
- Field, J. (2008a). Bricks or mortar: Which parts of the input does a second language listener rely on? *TESOL Quarterly*, 42(3), 411-432. doi:10.1002/j.1545-7249.2008.tb00139.x
- Field, J. (2008b). Guest editor's introduction emergent and divergent: A view of second language listening. *System*, 36(1), 2-9. doi:10.1016/j.system.2008.01.001
- Field, J. (2008c). Revisiting segmentation hypotheses in first and second language listening. *System*, 36(1), 35-51. doi:10.1016/j.system.2007.10.003

- Field, J. (2013). Cognitive validity. In A. Geranpayeh & L. Taylor (Eds.), *Examining listening: Research and practice in assessing second language listening* (pp. 77-151). Cambridge, UK: Cambridge University Press.
- Flavell, J. H. (1979). Metacognition and cognitive memory: A new area of cognitive developmental inquiry. *American Psychologist*, 34(10), 906-911. doi:10.1037/0003-066X.34.10.906
- Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5), 378-382. doi:10.1037/h0031619
- Flowerdew, J. (1994). Research of relevance to second language lecture comprehension: An overview. In J. Flowerdew (Ed.), *Academic listening: Research perspectives* (pp. 7-29). Cambridge, UK: Cambridge University Press.
- Fox, J. (2004). Test decision over time: Tracking validity. *Language Testing*, 21(4), 437-465. doi:10.1191/0265532204lt292oa
- Fox, J., Pychyl, T., & Zumbo, B. (1997). An investigation of background knowledge in the assessment of language proficiency. In A. Huhta, V. Kohonen, L. Kurki-Suonio, & S. Luoma (Eds.), *Current developments and alternatives in language assessment: Proceedings of LTRC 96*. Jyväskylä, FIN: University of Jyväskylä.
- Fulcher, G. (1997). An English language placement test: issues in reliability and validity. *Language Testing*, 14(2), 113-139. doi:10.1177/026553229701400201
- Fulcher, G. (2010). *Practical language testing*. New York, NY: Routledge
- Fulcher, G., & Davidson, F. (2007). *Language testing and assessment: An advanced resource book*. New York, NY: Routledge.
- Fulcher, G., & Davidson, F. (2009). Test architecture, test retrofit. *Language Testing*, 26(1), 123-144. doi:10.1177/0265532208097339
- Garrett, H. (1937). *Statistics in psychology and education*. New York, NY: Longmans.
- Geranpayeh, A., & Kunnan, A. (2007). Differential item functioning in terms of age in the certificate in advanced English examination. *Language Assessment Quarterly*, 4(2), 190-222. doi:10.1080/15434300701375758
- Geranpayeh, A., & Taylor, L. (Eds.). (2013). *Examining listening: Research and practice in assessing second language listening*. Cambridge, UK: Cambridge University Press.

- Glas, C. A. W. (1999). Modification indices for the 2-PL and the nominal response model. *Psychometrika*, 64(3), 273-294. doi:10.1007/bf02294296
- Goh, C. M. (2000). A cognitive perspective on language learners' listening comprehension problems. *System*, 28(1), 55-75. doi:[https://doi.org/10.1016/S0346-251X\(99\)00060-3](https://doi.org/10.1016/S0346-251X(99)00060-3)
- Goh, C. M. (1997). Metacognitive awareness and second language listeners. *ELT Journal*, 51(4), 361-369. doi:10.1093/elt/51.4.361
- Goh, C. M., & Taib, Y. (2006). Metacognitive instruction in listening for young learners. *ELT Journal*, 60(3), 222-232. doi:10.1093/elt/ccl002
- Gordon, M. (1981). Models of pluralism: The new America dilemma. *Annals of the American Academy of Political and Social Science*, 454, 178-188. doi:10.1177/000271628145400115
- Gouvernement du Québec. (2018). *Règlement sur la pondération applicable à la sélection des ressortissants étrangers: Loi sur l'immigration au Québec*. Retrieved from [http://www2.publicationsduquebec.gouv.qc.ca/dynamicSearch/telecharge.php?type=2&file=/I\\_0\\_2/I0\\_2R2.htm](http://www2.publicationsduquebec.gouv.qc.ca/dynamicSearch/telecharge.php?type=2&file=/I_0_2/I0_2R2.htm).
- Government of Quebec. (1977, 2017). *Charte de la langue française*. Québec, Canada: Éditeur Officiel du Québec.
- Government of Quebec. (2017). *Loi sur l'immigration au Québec*. Québec, Canada: Éditeur Officiel du Québec.
- Granger, C. (2008). Rasch analysis is important to understand and use for measurement. *Rasch Measurement Transactions*, 21(3), 1122-1123.
- Grin, F. (2006). Economic considerations in language policy. In T. Ricento (Ed.), *An introduction to language policy: Theory and method* (pp. 77-94). Malden, MA: Blackwell Publishing.
- Guerrero, M. D. (2000). The unified validity of the Four Skills Exam: applying Messick's framework. *Language Testing*, 17(4), 397-421. doi:10.1177/026553220001700402
- Guilford, J. P. (1946). New standards for test evaluation. *Educational and Psychological Measurement*, 6(4), 427-439.
- Gwet, K. L. (2012). *Hanbook of inter-rater reliability* (3 ed.). Gaithersburg, MD: Advanced Analytics.

- Gwet, K. L. (2015). AgreeStat: Chance-corrected agreement and Intraclass correlation coefficients with excel (Version 2015.6).
- Haladyna, T. M., & Rodriguez, M. C. (2013). *Developing and validating test items*. New York, NY: Routledge.
- Hansen, C., & Jensen, C. (1994). Evaluating lecture comprehension. In J. Flowerdew (Ed.), *Academic listening: Research perspectives* (pp. 241-268). Cambridge, UK: Cambridge University Press.
- Harding, L. (2012). Accent, listening assessment and the potential for a shared-L1 advantage: A DIF perspective. *Language Testing*, 29(2), 163-180. doi:10.1177/0265532211421161
- Heubert, J. P., & Hauser, R. M. (Eds.). (1999). *High stakes: Testing for tracking, promotion and graduation*. Washington, DC: National Academy Press.
- Hogan-Brun, G., Mar-Molinero, C., & Stevenson, P. (2009). Testing regimes: Introducing cross-national perspectives on language, migration and citizenship. In G. Hogan-Brun, C. Mar-Molinero, & P. Stevenson (Eds.), *Discourses on language and integration: Critical perspectives on language testing regimes in Europe* (pp. 1-14). Philadelphia, PA: John Benjamins Publishing.
- Holland, P. W., & Thayer, D. T. (1988). Differential item functioning and the Mantel-Haenszel procedure. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 129-145). Hillsdale, NJ: Lawrence Erlbaum.
- Holland, P. W., & Wainer, H. (Eds.). (1993). *Differential item functioning*. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Hu, L. t., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, 6(1), 1-55. doi:10.1080/10705519909540118
- Hubley, A. M., & Zumbo, B. (2011). Validity and the consequences of test interpretation and use. *Social Indicators Research*, 103(2), 219-230. doi:10.1007/s11205-011-9843-4
- In'nami, Y., & Koizumi, R. (2012). Factor structure of the revised TOEIC® test: A multiple-sample analysis. *Language Testing*, 29(1), 131-152. doi:10.1177/0265532211413444
- International Language Testing Association. (2000). *ILTA code of ethics*. Vancouver, BC: ILTA.

- Jackson, D. L., Gillaspay Jr, J. A., & Purc-Stephenson, R. (2009). Reporting practices in confirmatory factor analysis: An overview and some recommendations. *Psychological Methods, 14*(1), 6-23. doi:10.1037/a0014694
- Jamal, A. (2007). Nationalizing states and the constitution of hollow citizenship: Israel and its palestinians citizens. *Ethnopolitics, 6*(4), 471-493. doi:10.1080/17449050701448647
- Jensen, C., & Hansen, C. (1995). The effect of prior knowledge on EAP listening test performance. *Language Testing, 12*(1), 99-119. doi:10.1177/026553229501200106
- Jerrard, R., & Jerrard, M. (1998). *The gradschool handbook: An insider's guide to getting in and succeeding*. New York, NY: The Berkley Publishing Group.
- Jia, Y. (2013). *Justifying the use of a second language oral test as an exit test in hong kong: An application of assessment use argument framework*. (Unpublished doctoral thesis), University of Los Angeles, Los Angeles, CA.
- Joint Advisory Committee. (1993). *Principles for fair student assessment practices for education in Canada*. Edmonton, AB.
- Jones, D. (1999). *Saying for cynics*. Calgary, CA: Detselig Enterprises.
- Jöreskog, K. G. (1971). Statistical analysis of sets of congeneric tests. *Psychometrika, 36*(2), 109-133. doi:10.1007/BF02291393
- Jun, H. S. (2014). *A validity argument for the use of scores from a web-search-permitted and web-source-based integrated writing test*. (Unpublished doctoral thesis), Iowa State University, Ames, IA.
- Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin, 112*(3), 527-535. doi:10.1037/0033-2909.112.3.527
- Kane, M. T. (2001). Current concerns in validity theory. *Journal of Educational Measurement, 38*(4), 319-342. doi:10.1111/j.1745-3984.2001.tb01130.x
- Kane, M. T. (2006). Validation. In R. Brennan (Ed.), *Educational measurement* (4 ed., pp. 17-64). Westport, CT: American Council of Education and Praeger.
- Kane, M. T. (2009). Validating the interpretations and uses of test scores. In R. W. Lissitz (Ed.), *The concept of validity: Revisions, new directions, and applications* (pp. 39-64). Charlotte, NC: Information Age Publishing.
- Kane, M. T. (2012). Validating score interpretations and uses. *Language Testing, 29*(1), 3-17. doi:10.1177/0265532211417210

- Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, 50(1), 1-73.
- Kane, M. T., Crooks, T., & Cohen, A. (1999). Validating measures of performance. *Educational Measurement: Issues and Practice*, 18(2), 5-17. doi:10.1111/j.1745-3992.1999.tb00010.x
- Kenny, D. A., & Milan, S. (2012). Identification: A nontechnical discussion of a technical issue. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 145-163). New York, NY: Guilford Press.
- Kim, H.-J. (2010). Investigating the construct validity of a speaking performance test. *Spain Fellow Working Papers in Second or Foreign Language Assessment*, 8, 1-30.
- Kim, M. (2001). Detecting DIF across the different language groups in a speaking test. *Language Testing*, 18(1), 89-114. doi:10.1177/026553220101800104
- Kim, Y.-H. (2010). *An argument-based validity inquiry into the empirically-derived descriptor-based diagnostic (EDD) assessment in ESL academic writing*. (Unpublished doctoral thesis), University of Toronto, Toronto, ON.
- Knoch, U., & Chapelle, C. A. Validation of rating processes within an argument-based framework. *Language Testing*, 00(0), 1-23. doi:10.1177/0265532217710049
- Kramer, C. (1977). Perceptions of female and male speech. *Language and Speech*, 20(2), 151-161. doi:10.1177/002383097702000207
- Kunnan, A. (1997). Connecting fairness and validation. In A. Huhta., V. Kohonen., L. Kurti-Suomo., & S. Luoma. (Eds.), *Current developments and alternatives in language assessment* (pp. 85-105). Jyväskylä, Finland: University of Jyväskylä.
- Kunnan, A. (2009). Testing for citizenship: The U.S. naturalization test. *Language Assessment Quarterly*, 6(1), 89-97. doi:10.1080/15434300802606630
- LaFlair, G. T., & Staples, S. (2017). Using corpus linguistics to examine the extrapolation inference in the validity argument for a high-stakes speaking assessment. *Language Testing*, 34(4), 451-475. doi:10.1177/0265532217713951
- Lallmamode, S. P., Mat Daud, N., & Abu Kassim, N. L. (2016). Development and initial argument-based validation of a scoring rubric used in the assessment of L2 writing electronic portfolios. *Assessing Writing*, 30, 44-62. doi:10.1016/j.asw.2016.06.001

- Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1), 159-174. doi:10.2307/2529310
- Lawley, D. N., & Maxwell, A. E. (1963). *Factor analysis as a statistical method*. London, UK: Butterworth.
- Lazarsfeld, P. F. (1959). Latent structure analysis. In S. Koch (Ed.), *Psychology: A study of science* (pp. 476-543). New York, NY: McGraw Hill.
- Le, H. (2010). *Developing a validity argument for the English placement listening Fall 2010 test at Iowa State University*. (Unpublished masters thesis), Iowa State University, Ames, IA.
- Levelt, W. J. M. (1993). The architecture of normal spoken language use. In G. Blanken, J. Dittmann, H. Grimm, J. C. Marshall, & C. W. Wallesch (Eds.), *Linguistic disorders and pathologies: An international handbook* (pp. 1-15). New York, NY: de Gruyter.
- Li, H., & Suen, H. K. (2015). How do Chinese ESL learners recognize English words during a reading test? A comparison with romance-language-speaking ESL learners. *International Multilingual Research Journal*, 9(2), 93-107. doi:10.1080/19313152.2014.995013
- Li, Z. (2015). *An argument-based validation study of the English placement Test (EPT) – Focusing on the inferences of extrapolation and ramification*. (Unpublished doctoral thesis), Iowa State University, Ames, IA.
- Linacre, J. M. (1998a). Rasch first or factor first? *Rasch Measurement Transactions*, 11(4), 603.
- Linacre, J. M. (1998b). Structure in Rasch residuals: Why principal component analysis. *Rasch Measurement Transactions*, 12(2), 636.
- Linacre, J. M. (2002). What do infit and outfit, mean-square and standardized mean? *Rasch Measurement Transactions*, 16(2), 878.
- Linacre, J. M. (2008). Variance in data explained by Rasch measures. *Rasch Measurement Transactions*, 22(1), 1164.
- Linacre, J. M. (2010). Differential item functioning DIF sample size nomogram. *Rasch Measurement Transactions*, 26(4), 1391.
- Linacre, J. M. (2013). Differential item functioning DIF sample size nomogram. *Rasch Measurement Transactions*, 26(4), 1391.

- Linacre, J. M. (2017). Winsteps (Version 4.0.1).
- Lissitz, R., & Samuelsen, K. (2007). Dialogue on validity: A suggested change in terminology and emphasis regarding validity and education. *Educational Researcher*, 36(8), 437-448.
- Liu, Y., Zumbo, B., Gustafson, P., Huang, Y., Kroc, E., & Wu, A. (2016). Investigating causal DIF via propensity score methods. *Practical Assessment, Research & Evaluation*, 21(13), 1-24.
- Llosa, L. (2007). Validating a standards-based classroom assessment of English proficiency: A multitrait-multimethod approach. *Language Testing*, 24(4), 489-515.  
doi:10.1177/0265532207080770
- Lord, F. M. (1980). *Application of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.
- Lund, R. J. (1991). A comparison of second language listening and reading comprehension. *The Modern Language Journal*, 75(2), 196-204. doi:10.1111/j.1540-4781.1991.tb05350.x
- Lundsteen, S. W. (1993). Metacognitive listening. In A. D. Wolvin & C. G. Coakley (Eds.), *Perspectives on listening* (pp. 106-123). Norwood, NJ: Ablex Publishing.
- Lynch, T. (1998). Theoretical perspectives on listening. *Annual Review of Applied Linguistics*, 18, 3-19. doi:10.1017/S0267190500003457
- Lynch, T. (2009). *Teaching second language listening: A guide to evaluating, adapting, and creating tasks for listening in the language classroom*. Oxford, UK: Oxford University Press.
- MacCallum, R. C., & Austin, J. T. (2000). applications of structural equation modeling in psychological research. *Annual Review of Psychology*, 51(1), 201-226.  
doi:10.1146/annurev.psych.51.1.201
- Mackenzie, S. B., Posakoff, P. M., & Jarvis, C. B. (2005). The problem of measurement model misspecification in behavioral and organizational research and some recommended solutions. *Journal of Applied Psychology*, 90(4), 710-730.  
doi:10.1037/0021-9010.90.4.710
- Mann, W., & Marshall, C. R. (2010). Building an assessment use argument for sign language: the BSL Nonsense Sign Repetition Test. *International Journal of Bilingual Education and Bilingualism*, 13(2), 243-258. doi:10.1080/13670050903474127

- Markus, K. A., & Borsboom, D. (2013). *Frontiers of test validity theory: Measurement, causation, and meaning*. New York, NY: Routledge.
- Marslen-Wilson, W. D. (1989). *Lexical representation and process*. Cambridge, MA: MIT Press.
- Masters, G. (1982). A Rasch model for partial credit scoring. *Psychometrika*, 42(2), 149-174. doi:10.1007/BF02296272
- May, S. (2008). Language education, pluralism and citizenship. In S. May & N. Hornberger (Eds.), *Encyclopedia of language education: Language policy and political issues in education* (2 ed., Vol. 1, pp. 15-29). New York, NY: Springer.
- McNamara, T. (1991). Test dimensionality: IRT analysis of an ESP listening test. *Language Testing*, 8(2), 139-159. doi:10.1177/026553229100800204
- McNamara, T. (2005). 21st century shibboleth: Language tests, identity and intergroup conflict. *language Policy* 4(4), 351-370. doi:10.1007/s10993-005-2886-0
- McNamara, T. (2009). Australia: The dictation test redux? *Language Assessment Quarterly*, 6(1), 106-111. doi:10.1080/15434300802606663
- McNamara, T., & Roever, C. (2006). *Language testing: The social dimension*. Oxford, UK: Oxford University Press.
- McNamara, T., & Shohamy, E. (2008). Language tests and human rights. *International Journal of Applied Linguistics*, 18(1), 89-95.
- Mehrens, W. A. (1997). The consequences of consequential validity. *Educational Measurement: Issues and Practice*, 16(2), 16-18. doi:10.1111/j.1745-3992.1997.tb00588.x
- Mendelsohn, D. J. (1994). *Learning to listen: A strategy-based approach for the second language learner*. San Diego, CA: Dominic Press.
- Messick, S. (1981). Evidence and ethics in the evaluation of tests. *Educational Researcher*, 10(9), 9-20. doi:10.3102/0013189X010009000
- Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3 ed., pp. 13-103). New York, NY: American Council on Education and Macmillan.
- Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher*, 23(2), 13-23. doi:10.3102/0013189x023002013

- Messick, S. (1996). Validity and washback in language testing. *Language Testing*, 13(3), 156-241.
- Messick, S. (1998). Test validity: A matter of consequence. *Social Indicators Research*, 45, 35-44.
- Michell, J. (2009). Invalidity in validity. In R. W. Lissitz (Ed.), *The concept of validity: Revisions, new directions, and applications* (pp. 111-133). Charlotte, NC: Information Age Publishing.
- Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits to out capacity for information processing. *Psychological Review*, 63(2), 81-97.  
doi:10.1037/h0043158
- Ministère de l'Immigration Diversité et Inclusion (MIDI). (2015). *Évaluation des connaissances du français et de l'anglais*. Retrieved from <http://www.immigration-quebec.gouv.qc.ca/fr/immigrer-installer/travailleurs-permanents/demande-immigration-general/conditions-requises/connaissances-linguistiques.html>
- Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives*, 1(1), 3-62.  
doi:10.1207/S15366359MEA0101\_02
- Moshagen, M., & Musch, J. (2014). Sample size requirements of the robust weighted least squares estimator. *Methodology*, 10(2), 60-70. doi:10.1027/1614-2241/a000068
- Moss, P. (1998). The role of consequences in validity theory. *Educational Measurement: Issues and Practice*, 17(2), 6-12. doi:10.1111/j.1745-3992.1998.tb00826.x
- Muniz, J., Hambleton, R. K., & Xing, D. (2001). Small sample studies to detect flaws in item translations. *International Journal of Testing*, 1(2), 115-135.  
doi:10.1207/S15327574IJT0102\_2
- Muthén, B., & Kaplan, D. (1985). A comparison of some methodologies for the factor analysis of non-normal Likert variables. *British Journal of Mathematical and Statistical Psychology*, 38(2), 171-189. doi:10.1111/j.2044-8317.1985.tb00832.x
- Muthén, B., & Muthén, L. (2017). Mplus (Version 8).
- Nation, P. (2006). How large a vocabulary is needed for reading and listening? *Canadian Modern Language Review*, 63(1), 59-82. doi:10.3138/cmlr.63.1.59

- Nation, P., & Newton, J. (2009). *Teaching ESL/EFL listening and speaking*. New York, NY: Routledge.
- Newton, P., & Shaw, S. D. (2014). *Validity in educational and psychological assessment*. Thousand Oaks, CA: Sage.
- Ockey, G. J. (2007). Investigating the validity of math word problems for English language learners with DIF. *Language Assessment Quarterly*, 4(2), 149-164.  
doi:10.1080/15434300701375717
- Ockey, G. J. (2013). Assessment of listening. In C. A. Chapelle (Ed.), *Encyclopedia of Applied Linguistics* (pp. 212-218). Malden, MA: Blackwell Publishing.
- Ockey, G. J., & French, R. (2016). From one to multiple accents on a test of L2 listening comprehension. *Applied Linguistics*, 37(5), 693-715. doi:10.1093/applin/amu060
- Ockey, G. J., Papageorgiou, S., & French, R. (2016). Effects of strength of accent on an L2 interactive lecture listening comprehension test. *International Journal of Listening*, 30(1-2), 84-98. doi:10.1080/10904018.2015.1056877
- Oller, J. W. (1979). The factorial structure of language proficiency: Divisible or not? In J. J. W. Oller (Ed.), *Language test at school: A pragmatic approach* (pp. 423-458).
- Oller, J. W. (Ed.) (1983). *Issues in language testing research*. Rowley, MA: Newbury House Publishers.
- Ostini, R., & Nering, M. L. (2006). *Polytomous item response theory models*. Thousand Oaks, CA: Sage.
- Pae, T.-I. (2004). DIF for examinees with different academic backgrounds. *Language Testing*, 21(1), 53-73. doi:10.1191/0265532204lt274oa
- Pae, T.-I. (2012). Causes of gender DIF on an EFL language test: A multiple-data analysis over nine years. *Language Testing*, 29(4), 533-554. doi:10.1177/0265532211434027
- Pagé, M. (2011). *Politiques d'intégration et cohésion sociale*. Montréal, QC: Bibliothèque et Archives Nationales du Québec.
- Pardo-Ballester, C. (2010). The validity argument of a web-based Spanish listening exam: Test usefulness evaluation. *Language Assessment Quarterly*, 7(2), 137-159.  
doi:10.1080/15434301003664188
- Parkers, J. (2007). Reliability as argument. *Educational Measurement: Issues and Practice*, 26(4), 2-10. doi:10.1111/j.1745-3992.2007.00103.x

- Penfield, R. D. (2014). An NCME instructional module on polytomous item response theory models. *Educational Measurement: Issues and Practice*, 33(1), 36-48.  
doi:10.1111/emip.12023
- Penfield, R. D., & Camilli, G. (2006). Differential item functioning and item bias. In C. R. Rao & S. Sinharay (Eds.), *Handbook of Statistics* (Vol. 26, pp. 125-167): Elsevier.
- Popham, W. J. (1997). Consequential validity: Right concern-wrong concept. *Educational Measurement: Issues and Practice*, 16(2), 9-13. doi:10.1111/j.1745-3992.1997.tb00586.x
- Preston, K. S. J., Parral, S. N., Gottfried, A. W., Oliver, P. H., Gottfried, A. E., Ibrahim, S. M., & Delany, D. (2015). Applying the nominal response model within a longitudinal framework to construct the positive family relationships Scale. *Educational and Psychological Measurement*, 75(6), 901-930. doi:10.1177/0013164414568717
- Preston, K. S. J., & Reise, S. P. (2014). Estimating the nominal response model under nonnormal conditions. *Educational and Psychological Measurement*, 74(3), 377-399. doi:10.1177/0013164413507063
- Preston, K. S. J., & Reise, S. P. (2015). Detecting faulty within-item category functioning with the nominal response model. In S. P. Reise & D. A. Revicki (Eds.), *Handbook of item response theory modeling: Applications to typical performance assessment* (pp. 386-405).
- R Core Team. (2015). A language and environment for statistical computing: R foundation for statistical computing (Version version 3.4.1). Vienna, Austria.
- Raïche. (2005). Critical eigenvalue sizes in standardized residual principal component analysis. *Rasch Measurement Transactions*, 19(1), 1012.
- Raju, N. S. (1988). The area between two item characteristic curves. *Psychometrika*, 53(4), 495-502. doi:10.1007/BF02294403
- Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Den: Danish Institute of Educational Research.
- Rasch, G. (1977). On specific objectivity: An attempt at formalizing the request for generality and validity of scientific statements. *Danish Yearbook of Philosophy*, 14, 58-94.  
doi:<http://www.rasch.org/memo18.htm>
- Reif, M. (2015). mcIRT: R package (Version 0.41).

- Revuelta, J. (2005). An item response theory model for nominal data based on the rising selection ratios criterion. *Psychometrika*, 70(2), 305-324. doi:10.1007/s11336-002-0975-y
- Ricento, T. (2006). Theoretical perspectives in language policy: An overview. In T. Ricento (Ed.), *An introduction to language policy: Theory and method* (pp. 3-9). Malden, MA: Blackwell Publishing.
- Robin, F. (2001). STDIF: Standardization DIF analysis program. Amherst, MA: University of Massachusetts, School of Education.
- Roever, C. (2007). DIF in the assessment of second language pragmatics. *Language Assessment Quarterly*, 4(2), 165-189. doi:10.1080/15434300701375733
- Rost, M. (1990). *Listening in language learning*. New York, NY: Longman.
- Rost, M. (1994). On line summaries as representations of lecture understanding. In J. Flowerdew (Ed.), *Academic listening: Research perspectives* (pp. 93-130). Cambridge, UK: Cambridge University Press.
- Rost, M. (2011). *Teaching and researching listening*. Harlow, UK: Longman.
- Rubin, J. (1994). A review of second language listening comprehension research. *The Modern Language Journal*, 78(2), 199-221. doi:10.1111/j.1540-4781.1994.tb02034.x
- Ryan, K. E., & Bachman, L. F. (1992). Differential item functioning on two tests of EFL proficiency. *Language Testing*, 9(1), 12-29. doi:10.1177/026553229200900103
- Samejima, F. (1969). *Estimation of latent ability using a response pattern of graded scores* (Psychometric Monograph No. 17). Richmond, VA: Psychometric Society. Retrieved from <http://www.psychometrika.org/journal/online/MN17.pdf>
- Samejima, F. (1979). *A new family of models for the multiple choice item*. ONR Research Report 79-4. Knoxville, TN: University of Tennessee.
- Saville, N. (2009). Language assessment in the management of international migration: A framework for considering the issues. *Language Assessment Quarterly*, 6(1), 17-29. doi.org/10.1080/15434300802606499
- Sawaki, Y. (2007). Construct validation of analytic rating scales in a speaking assessment: Reporting a score profile and a composite. *Language Testing*, 24(3), 355-390. doi:10.1177/0265532207077205

- Sawaki, Y., Stricker, L. J., & Oranje, A. H. (2009). Factor structure of the TOEFL Internet-based test. *Language Testing*, 26(1), 005-030. doi:10.1177/0265532208097335
- Schiffman. (2006). Language policy and linguistic culture. In T. Ricento (Ed.), *An introduction to language policy: Theory and method* (pp. 111-125). Malden, MA: Blackwell Publishing.
- Schmittmann, V. D., Cramer, A. O. J., Waldorp, L. J., Epskamp, S., Kievit, R. A., & Borsboom, D. (2013). Deconstructing the construct: A network perspective on psychological phenomena. *New Ideas in Psychology*, 31(1), 43-53. doi:10.1016/j.newideapsych.2011.02.007
- Schramm, W. (Ed.) (1971). *The process and effects of mass communication*. Urbana Champaign, IL: University of Illinois Press.
- Seo, D., Taherbhai, H., & Frantz, R. (2016). Psychometric evaluation and discussions of English language learners' listening comprehension. *International Journal of Listening*, 30(1-2), 47-66. doi:10.1080/10904018.2015.1065747
- Shank, R. C., & Abelson, R. P. (1977). Scripts, plans, and knowledge. In P. N. Johnson-Laird & P. C. Wason (Eds.), *Thinking: Readings in cognitive science* (pp. 421-434). Cambridge, UK: Cambridge University Press.
- Shaw, S., Crisp, V., & Johnson, N. (2012). A framework for evidencing assessment validity in large-scale, high-stakes international examinations. *Assessment in Education: Principles, Policy & Practice*, 19(2), 159-176. doi:10.1080/0969594X.2011.563356
- Shealy, R., & Stout, W. (1993). A model-based standardization differences and detects test bias/DTF as well as item bias/DIF. *Psychometrika*, 58(2), 159-194. doi:10.1007/BF02294572
- Shepard, L. A. (1993). Evaluating test validity. *Review of Research in Education*, 19, 405-450.
- Shohamy, E. (2001). *The power of tests: A critical perspective of the uses of language tests*. Harlow, UK: Pearson Education.
- Shohamy, E. (2006). *Language policy: Hidden agendas and new approaches*. New York, NY: Routledge.
- Shohamy, E. (2007). The power of language tests, the power of the English language and the role of ELT. In J. Cummins & C. Davison (Eds.), *International handbook of English language teaching* (Vol. 11, pp. 521-532). New York, NY: Springer.

- Shohamy, E., & Kanza, T. (2009). Language and citizenship in Israel. *Language Assessment Quarterly*, 6(1), 83-88. doi:10.1080/15434300802606622
- Sireci, S. G. (2007). On validity theory and test validation. *Educational Researcher*, 36(8), 477-481.
- Sireci, S. G. (2009). Packing and unpacking sources of validity evidence: History repeats itself again. In R. W. Lissitz (Ed.), *The concept of validity: Revisions, new directions, and applications* (pp. 19-37). Charlotte, NC: Information Age Publishing.
- Sireci, S. G. (2011). Evaluating test and survey items for bias across languages and cultures. In D. Matsumoto & F. VanDeVijver (Eds.), *Cross-cultural research methods in psychology* (pp. 216-241). Cambridge, UK: Cambridge University Press.
- Sireci, S. G., & Rios, J. A. (2013). Decisions that make a difference in detecting differential item functioning. *Educational Research and Evaluation*, 19(2-3), 170-187. doi:10.1080/13803611.2013.767621
- Smith Jr, E. V. (2002). Understanding Rasch measurement: Detecting and evaluating the impact of multidimensionality using item fit statistics and principal component analysis of residuals. *Journal of Applied Measurement*, 3(2), 205-231.
- Smith, R. M. (1991). The distributional properties of Rasch item fit statistics. *Educational and Psychological Measurement*, 51(3), 541-565. doi:10.1177/0013164491513003
- Smith, R. M. (2004). Fit analysis in latent trait measurement models. In E. V. Smith & R. M. Smith (Eds.), *Introduction to Rasch measurement: Theory, models, and applications* (pp. 73-92). Maple Grove, MN: JAM Press.
- Spearman, C. (1904). General intelligence, objectively determined and measured. *The American Journal of Psychology*, 15(2), 201-292. doi:10.2307/1412107
- Spolsky, B. (1995). *Measured words: The development of objective language testing*. Oxford, NY: Oxford University Press.
- Spolsky, B. (1997). The ethics of gatekeeping tests: what have we learned in a hundred years? *Language Testing*, 14(3), 242-247. doi:10.1177/026553229701400302
- Swain, M., Brooks, L., & Tocalli-Beller, A. (2002). Peer-peer dialogue as a means of second language learning. *Annual Review of Applied Linguistics*, 22, 171-185. doi:10.1017/S0267190502000090

- Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement*, 27(4), 361-370. doi:10.1111/j.1745-3984.1990.tb00754.x
- Tennant, A., & Conaghan, P. (2007). The Rasch measurement model in rheumatology: What is it and why use it? When should it be applied, and what should we look for in a Rasch paper? *Arthritis & Rheumatism*, 57(8), 1358-1362. doi:10.1002/art.23108
- Thissen, D., & Steinberg, L. (1984). A response model for multiple choice items. *Psychometrika*, 49(4), 501-519. doi:10.1007/BF02302588
- Thissen, D., Steinberg, L., & Fitzpatrick, A. R. (1989). Multiple-choice models: The distractors are also part of the item. *Journal of Educational Measurement*, 26(2), 161-176. doi: <http://www.jstor.org/stable/1434863>
- Thissen, D., Steinberg, L., & Wainer, H. (1988). Use of item response theory in the study of group differences in trace lines. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 147-169). Hillsdale, NJ: Erlbaum.
- Thompson, B. (2004). *Exploratory and confirmatory factor analysis: Understanding concepts and applications*. Washington, DC: American Psychological Association.
- Tobia, V., Ciancaleoni, M., & Bonifacci, P. (2016). Theoretical models of comprehension skills tested through a comprehension assessment battery for primary school children. *Language Testing*, 34(2), 223-239. doi:10.1177/0265532215625705
- Toulmin, S. E. (1958, 2003). *The uses of argument*. New York, NY: Cambridge University Press.
- Van Avermaet, P. (2009). Fortress Europe? Language policy regimes for immigration and citizenship. In G. Hogan-Brun, C. Mar-Molinero, & P. Stevenson (Eds.), *Discourses on language and integration: Critical perspectives on language testing regimes in Europe* (pp. 15-44). Amsterdam, The Netherlands: John Benjamins Publishing.
- Van Avermaet, P., & Rocca, L. (2013). Language testing and access. In E. D. Galaczi & C. Weir (Eds.), *Exploring language frameworks: Proceedings of the ALTE Kraków conference, July 2011* (pp. 11-44). Cambridge, UK: Cambridge University Press.
- Vandergrift, L. (1997a). The Cinderella of communication strategies: Reception strategies in interactive listening. *The Modern Language Journal*, 81(4), 494-505. doi: <http://www.jstor.org/stable/328892>

- Vandergrift, L. (1997b). The comprehension strategies of second language French listeners: A descriptive study. *Foreign Language Annals*, 30(3), 387-409. doi:10.1111/j.1944-9720.1997.tb02362.x
- Vandergrift, L. (2002). It was nice to see that our predictions were right: Developing metacognition in L2 listening comprehension. *Canadian Modern Language Review*, 58(4), 555-575. doi:10.3138/cmlr.58.4.555
- Vandergrift, L. (2004). Listening to learn or learning to listen? *Annual Review of Applied Linguistics*, 24, 3-25. doi:10.1017/S0267190504000017
- Vandergrift, L. (2007). Recent developments in second and foreign language listening comprehension research. *Language Teaching*, 40(3), 191-210. doi:10.1017/S0261444807004338
- Vandergrift, L., & Baker, S. (2015). Learner variables in second language listening comprehension: An exploratory path analysis. *Language Learning*, 65(2), 390-416. doi:10.1111/lang.12105
- Vandergrift, L., Goh, C. M., Mareschal, C. J., & Tafaghodtari, M. H. (2006). The metacognitive awareness listening questionnaire: Development and validation. *Language Learning*, 56(3), 431-462. doi:10.1111/j.1467-9922.2006.00373.x
- Voss, E. (2012). *A validity argument for score meaning of a computer-based ESL academic collocational ability test based on a corpus-driven approach to test design*. (Unpublished doctoral thesis), Iowa State University, Ames, IA.
- Wang, H. (2010). *Investigating the justifiability of an additional test use: An application of assessment use argument to an english as a foreign language test*. (Unpublished doctoral thesis), University of California, Los Angeles, CA.
- Wang, H., Choi, I., Schmidgall, J., & Bachman, L. F. (2012). Review of Pearson Test of English Academic: Building an assessment use argument. *Language Testing*, 29(4), 603-619. doi:10.1177/0265532212448619
- Weir, C. (2005). *Language testing and validation: An evidence-based approach*. Hampshire, UK: Palgrave MacMillan.
- Weir, C. J., & Vidaković, I. (2013). The measurement of listening ability 1913-2012. In C. Weir, I. Vidaković, & E. Galaczi (Eds.), *Measured constructs: A history of cambridge*

- English language examinations 1913-2012* (pp. 347-414). Cambridge, UK: Cambridge University Press.
- Wenden, A. (1987). Metacognition: An expanded view on the cognitive abilities of L2 learners. *Language Learning*, 37(4), 573-597. doi:10.1111/j.1467-1770.1987.tb00585.x
- Wenden, A. (1998). Metacognitive knowledge and language learning. *Applied Linguistics*, 19(4), 515-537. doi:10.1093/applin/19.4.515
- Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis* (2 ed.). Houston, TX: Springer.
- Wilson, M. (2003). Discovery listening - improving perceptual understanding. *ELT Journal*, 57(4), 335-343. doi:10.1093/elt/57.4.335
- Wright, B. (1994). Local dependency, correlations, and principal components. *Rasch Measurement Transactions*, 10(3), 509-511.
- Wright, B. (1996). Comparing Rasch measures and factor analysis. *Structural Equation Modeling*, 3(1), 3-24. doi:10.1080/10705519609540026
- Wright, B., & Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions*, 8(3), 370.
- Wright, B., & Mok, M. M. (2004). An overview of the family of Rasch measurement models. In E. V. Smith & R. M. Smith (Eds.), *Introduction to Rasch measurement: Theory, models, and applications* (pp. 1-24). Maple Grove, MN: JAM Press.
- Wright, S. (2008). Citizenship tests in Europe: Editorial introduction. *International Journal on Multicultural Societies*, 10(1), 1-9.
- Wu, M., & Adams, R. (2013). Properties of Rasch residual fit statistics. *Journal of Applied Measurement*, 14(4), 339-355.
- Xi, X. (2010). How do we go about investigating test fairness? *Language Testing*, 27(2), 147-170. doi:10.1177/0265532209349465
- Yoo, H., & Manna, V. F. (2015). Measuring English language workplace proficiency across subgroups: Using CFA models to validate test score interpretation. *Language Testing*, 34(1), 101-126. doi:10.1177/0265532215618987
- Youn, S. J. (2015). Validity argument for assessing L2 pragmatics in interaction using mixed methods. *Language Testing*, 32(2), 199-225. doi:10.1177/0265532214557113

- Zumbo, B. (1999). *A handbook on the theory and methods of differential item functioning (DIF): logistic regression modeling as a unitary framework for binary and Likert-type (ordinal) item scores*. Ottawa, ON: Directorate of Human Resources Research and Evaluation, Department of National Defense.
- Zumbo, B. (2003). Does item-level DIF manifest itself in scale-level analyses? Implications for translating language tests. *Language Testing*, 20(2), 136-147.  
doi:10.1191/0265532203lt248oa
- Zumbo, B. (2007). Three generations of DIF analyses: Considering where it has been, where it is now, and where it is going. *Language Assessment Quarterly*, 4(2), 223-233.  
doi:10.1080/15434300701375832
- Zumbo, B. (2014). What role does, and should, the test standards play outside of the United States of America? *Educational Measurement: Issues and Practice*, 33(4), 31-33.  
doi:10.1111/emip.12052
- Zwick, R., Thayer, D. T., & Lewis, C. (1999). An empirical bayes approach to Mantel-Haenszel DIF analysis. *Journal of Educational Measurement*, 36(1), 1-28.  
doi:10.1111/j.1745-3984.1999.tb00543.x

## Appendix A: Descriptive Statistics for Nationalities and First Languages

Table XXXIV Descriptive Statistics for Nationalities in Form A

Nationality	<i>n</i>	<i>M</i>	<i>SD</i>
Ivorian	209	18.82	4.48
Colombian	159	9.55	3.52
Cameroonian	154	20.10	4.64
French	149	25.48	3.26
Moroccan	115	17.33	5.01
Tunisian	102	18.81	4.86
Chinese	89	11.96	4.81
Israeli	72	10.14	3.26
Senegalese	68	17.66	4.73
Bulgarian	49	11.98	4.53
Algerian	49	22.33	3.70
Lebanese	46	17.85	5.68
Guinean	36	15.25	5.05
Russian	29	10.17	5.12
Iranian	16	11.31	4.38
Indian	11	10.91	4.57
Togolese	9	14.78	7.07
Egyptian	9	12.33	5.12
Congolese	9	20.22	5.02
Ukrainian	7	10.43	3.82
Syrian	7	10.57	6.60
Italian	7	19.29	7.57
Beninese	7	23.29	3.73
Peruvian	6	15.50	5.61
Spanish	6	19.83	3.87
Belgium	5	25.40	2.30
American	5	11.60	3.91
Pakistani	4	14.50	3.70
Mexican	4	17.00	5.60
Djiboutian	4	21.75	3.69
Rwandan	3	19.67	7.57
Portuguese	3	18.33	7.09
Haitian	3	19.00	1.73
Korean	3	11.33	1.53
Bolivian	3	15.33	6.03
Vietnamese	2	17.00	4.24
Tajik	2	13.00	8.49
Serbian	2	11.50	0.71
Romanian	2	17.50	6.36
Palestinian	2	6.50	0.71
Nigerian	2	18.50	9.19
Moldavian	2	17.50	0.71
Mauritian	2	20.50	6.36
Croatian	2	9.50	0.71
Brazilian	2	16.00	5.66
Belarussian	2	16.00	1.41
Argentinian	2	11.00	5.66

Table XXXIV continued

German	2	25.50	0.71
Zimbabwean	1	15.00	NA
Singaporean	1	16.00	NA
Ugandan	1	4.00	NA
Nepalese	1	9.00	NA
Mauritanian	1	22.00	NA
Malian	1	23.00	NA
Madagascan	1	22.00	NA
Jordanian	1	9.00	NA
Irish	1	25.00	NA
Honduran	1	7.00	NA
Ecuadorian	1	21.00	NA
Cuban	1	14.00	NA
Chilean	1	23.00	NA
Cape Verdean	1	27.00	NA
Canadian	1	23.00	NA
Austrian	1	15.00	NA
Angolan	1	3.00	NA
Albanian	1	18.00	NA

Table XXXV Descriptive Statistics for First Language in Form A

First Language	<i>n</i>	<i>M</i>	<i>SD</i>
French	546	21.35	4.98
Arabic	281	17.63	5.38
Spanish	184	10.52	4.55
Russian	104	10.36	3.94
Mandarin	87	11.75	4.57
Bulgarian	49	11.98	4.53
Wolof	33	17.42	4.01
English	33	15.24	6.38
Fulani	22	15.09	5.88
Missing data	20	17.50	4.97
Kabyle	18	21.72	4.71
Persian	13	11.00	3.89
Serere	10	16.50	5.15
Italian	9	17.67	7.57
Bamileke	7	18.00	4.04
Portuguese	6	15.00	7.90
Malinke	6	13.83	3.06
Berber	6	21.50	5.54
Armenian	5	21.20	8.23
Ukrainian	4	11.50	3.42
Urdu	4	14.50	3.70
Gujarati	4	10.25	2.75
Diola	4	16.50	1.91
Korean	4	15.75	8.92
Romanian	3	17.33	4.51
German	3	22.00	6.08
Vietnamese	2	17.00	4.24
Tajik	2	13.00	8.49
Somali	2	21.50	6.36
Rwandan	2	17.00	8.49
Punjabi	2	8.00	0.00
Malagasy	2	22.50	0.71
Hebrew	2	8.00	2.83
Croatian	2	9.50	0.71
Bengali	2	15.00	8.49
Afar	2	22.00	0.00
Zulu	1	15.00	NA
Turkish	1	20.00	NA
Tamil	1	10.00	NA
Serbian	1	11.00	NA
Nepalese	1	9.00	NA
Dutch	1	22.00	NA
Moldavian	1	18.00	NA
Lingala	1	19.00	NA
Igbo	1	12.00	NA
Hindi	1	17.00	NA
Multiple languages	1	13.00	NA
Creole	1	16.00	NA
Bantu	1	12.00	NA
Bambara	1	23.00	NA
Azeri	1	6.00	NA
Albanian	1	18.00	NA

Table XXXVI Descriptive Statistics for Nationalities in Form B

Nationality	<i>n</i>	<i>M</i>	<i>SD</i>
Colombian	287	11.37	3.40
Ivorian	286	18.26	4.52
Iranian	220	10.83	3.41
Moroccan	165	18.91	5.04
French	163	25.62	2.80
Cameroonian	156	20.09	4.14
Egyptian	153	12.19	4.29
Chinese	96	12.86	4.62
Brazilian	94	14.10	4.30
Senegalese	79	18.97	4.48
Tunisian	66	19.77	4.88
Algerian	45	20.71	4.34
Indian	23	9.57	3.29
Bangladeshi	22	8.95	2.19
Libanese	21	18.33	5.36
Rwandan	20	15.50	5.02
Mexican	16	15.38	4.38
Togolese	15	18.53	5.22
Beninese	15	20.53	3.93
American	12	18.67	6.02
Kazakh	10	13.50	4.72
Guinean	10	19.20	6.32
Syrian	8	14.75	7.21
Congolese	8	17.13	5.62
Peruvian	7	11.29	4.07
Mauritian	7	19.14	5.30
Russian	6	20.83	6.55
Romanian	6	10.33	4.93
Italian	6	20.83	4.45
Koreanm	6	10.83	3.87
Pakistani	5	7.80	1.64
Malian	5	18.80	4.66
Gabonese	5	24.60	2.07
Burundian	5	12.60	5.86
British	5	13.80	5.07
Vietnamese	4	15.00	6.48
Philippine	4	6.00	2.83
Spanish	4	19.25	1.50
BURKINABE	4	19.75	6.60
Palestinian	3	15.67	5.69
Mauritanian	3	22.00	2.65
Jordanian	3	17.33	7.37
Haitian	3	19.67	3.06
Djiboutian	3	24.67	1.53
Venezuelan	2	22.00	5.66
Portuguese	2	22.50	3.54
Madagascan	2	12.50	4.95
Iraqi	2	17.00	4.24
Costa Rican	2	13.00	2.83
Cambodian	2	15.00	8.49
German	2	21.50	0.71
Ukranian	1	11.00	NA
Taiwanese	1	9.00	NA

Table XXXVI continued

Sri-Lankan	1	10.00	NA
Somalian	1	12.00	NA
Uzbek	1	16.00	NA
Nicaraguan	1	7.00	NA
Moldavian	1	9.00	NA
Lithuanian	1	22.00	NA
Libyan	1	12.00	NA
Kenyan	1	14.00	NA
Japanese	1	8.00	NA
Irish	1	10.00	NA
Indonesian	1	8.00	NA
Hungarian	1	23.00	NA
Dominican	1	16.00	NA
Bulgarian	1	21.00	NA
Bissau-Guinean	1	20.00	NA
Azerbaijani	1	8.00	NA
Armenian	1	23.00	NA
Argentinian	1	25.00	NA

Table XXXVII Descriptive Statistics for First Language in Form B

First Language	<i>n</i>	<i>M</i>	<i>SD</i>
French	678	20.77	5.07
Spanish	322	11.80	3.81
Arabic	285	18.94	4.99
Persian	215	10.93	3.39
Egyptian	142	11.90	4.10
Portuguese	97	14.29	4.37
Mandarin	96	12.68	4.40
English	65	12.49	5.80
Wolof	45	19.04	4.19
Bengali	21	8.90	2.23
Russian	17	14.88	5.69
Kabyle	12	21.50	3.26
Serere	10	19.10	4.72
Multiple languages	10	16.00	3.86
Diola	8	17.00	3.34
Vietnamese	6	15.83	6.24
Romanian	6	10.33	4.93
Korean	6	10.83	3.87
Hindi	4	7.25	1.71
Somali	3	21.00	7.81
Malagasy	3	16.00	7.00
Gujarati	3	10.33	3.51
Creole French	3	20.00	4.36
Creole	3	21.33	5.86
Berber	3	20.33	8.96
Soninke	2	19.50	10.61
Fulani	2	19.50	4.95
Khmer	2	15.00	8.49
Italian	2	24.50	3.54
Creole English	2	20.50	0.71
Bambara	2	18.00	5.66
German	2	21.50	0.71
Ukrainian	1	11.00	NA
Tamil	1	8.00	NA
Swahili	1	17.00	NA
Rwandan	1	27.00	NA
Kirundi	1	13.00	NA
Uzbek	1	16.00	NA
Mandingo	1	17.00	NA
Lithuanian	1	22.00	NA
Kannada	1	11.00	NA
Japanese	1	8.00	NA
Iranian	1	8.00	NA
Indonesian	1	8.00	NA
Hungarian	1	23.00	NA
Ewe	1	23.00	NA
Efik	1	16.00	NA
Bulgarian	1	21.00	NA
Armenian	1	23.00	NA

## Appendix B: Correlation of Residuals, DIF Analyses and Graphics

### Table XXXVIII Correlation of Standardized Rasch Residuals in Form A

Item	1	2	3	4	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26		
2	-0.031																										
3	0.030	0.051																									
4	0.024	-0.046	0.025																								
6	-0.013	0.071	-0.021	-0.034																							
7	-0.075	0.130	0.006	-0.062	0.146																						
8	0.038	-0.022	-0.007	0.049	-0.066	-0.079																					
9	-0.028	0.073	0.034	-0.011	0.001	0.033	-0.026																				
10	0.025	-0.001	-0.071	0.014	-0.033	0.010	-0.014	0.010																			
11	-0.091	0.101	0.003	-0.041	0.076	0.100	-0.065	0.057	0.035																		
12	0.046	-0.147	-0.108	0.006	-0.030	-0.107	-0.100	-0.078	-0.048	-0.097																	
13	-0.033	-0.063	-0.060	-0.009	-0.106	-0.117	-0.017	-0.105	-0.031	-0.082	-0.058																
14	-0.029	0.051	-0.074	-0.035	-0.082	-0.042	-0.039	-0.027	-0.064	-0.053	-0.054	-0.032															
15	-0.087	-0.047	-0.100	-0.024	-0.031	-0.032	-0.053	-0.057	-0.033	-0.042	-0.040	-0.020	-0.029														
16	-0.034	-0.044	-0.069	-0.090	-0.030	-0.029	-0.077	-0.032	-0.045	-0.065	-0.039	-0.052	-0.064	-0.004													
17	-0.130	0.075	-0.007	-0.103	0.037	0.072	-0.093	0.040	-0.042	0.081	-0.101	-0.051	0.012	-0.004	-0.038												
18	0.003	-0.093	-0.044	-0.018	-0.046	-0.076	-0.048	-0.045	-0.043	-0.044	0.012	-0.014	-0.032	-0.093	-0.043	-0.063											
19	0.033	-0.054	-0.058	-0.053	-0.038	-0.089	0.002	-0.003	-0.087	-0.057	-0.062	-0.026	-0.044	-0.068	-0.063	-0.041	-0.042										
20	-0.024	-0.050	-0.029	-0.037	-0.039	-0.048	-0.023	-0.088	-0.026	-0.053	-0.008	-0.091	-0.050	-0.046	-0.052	-0.039	-0.065	-0.026									
21	-0.043	-0.086	-0.019	-0.027	-0.107	-0.063	-0.072	-0.076	-0.081	-0.083	0.015	-0.016	-0.066	-0.058	-0.013	-0.072	-0.016	0.014	0.012								
22	-0.010	-0.057	0.019	-0.021	-0.092	-0.036	-0.030	-0.058	-0.063	-0.077	-0.055	-0.023	-0.034	-0.081	-0.078	-0.074	-0.035	-0.027	-0.062	-0.015							
23	-0.009	-0.044	0.005	-0.081	0.005	-0.025	-0.067	-0.052	-0.085	-0.039	-0.012	0.000	-0.070	-0.023	-0.030	-0.068	-0.075	-0.065	-0.047	-0.068	0.037						
24	0.000	-0.123	-0.042	-0.045	-0.098	-0.106	-0.058	-0.097	-0.054	-0.121	-0.010	-0.067	-0.046	-0.028	-0.006	-0.092	-0.019	-0.046	-0.086	-0.029	-0.008	-0.018					
25	-0.052	-0.073	-0.065	-0.131	-0.031	-0.059	-0.112	-0.057	-0.023	-0.066	-0.040	-0.008	-0.069	0.002	-0.028	-0.011	-0.010	-0.021	-0.006	-0.039	-0.072	-0.086	-0.007				
26	-0.011	-0.107	-0.053	-0.006	-0.062	-0.120	-0.020	-0.090	-0.051	-0.046	-0.015	-0.009	-0.026	-0.036	-0.027	-0.063	-0.024	-0.011	-0.077	-0.022	-0.035	-0.055	-0.027	-0.081			
28	-0.003	-0.074	-0.014	-0.055	-0.033	0.013	0.012	-0.012	-0.010	0.001	-0.033	-0.014	-0.096	-0.034	-0.059	-0.062	-0.083	-0.087	-0.034	-0.133	-0.023	0.045	-0.010	-0.082	-0.021		

### Table XXXIX Correlation of Standardized Rasch Residuals in Form B

Item	1	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	23	24	25				
4	-0.052																									
5	0.024	-0.080																								
6	0.106	-0.034	0.054																							
7	-0.001	-0.056	0.017	0.061																						
8	-0.016	-0.103	0.027	-0.060	0.017																					
9	-0.004	0.012	-0.022	0.009	-0.007	0.002																				
10	-0.016	-0.066	0.008	-0.053	-0.042	0.018	-0.056																			
11	-0.018	0.031	-0.044	0.036	-0.015	-0.073	-0.044	-0.074																		
12	-0.012	-0.016	-0.077	0.043	0.010	-0.077	-0.043	-0.094	0.010																	
13	-0.079	0.017	-0.084	0.013	-0.003	-0.100	-0.037	-0.113	-0.008	0.020																
14	-0.041	-0.020	-0.049	-0.048	-0.037	-0.080	-0.012	-0.101	-0.037	-0.007	0.051															
15	-0.059	-0.046	-0.031	-0.049	-0.024	-0.081	-0.057	-0.092	-0.022	0.002	-0.009	0.015														
16	-0.093	-0.045	-0.063	-0.138	-0.128	-0.054	-0.076	0.000	-0.086	-0.121	-0.077	-0.049	-0.034													
17	-0.036	-0.074	-0.085	-0.096	-0.063	0.003	-0.049	0.005	-0.097	-0.071	-0.093	-0.040	-0.081	0.044												
18	-0.046	-0.060	-0.028	-0.056	-0.061	-0.001	-0.050	-0.012	-0.091	-0.072	-0.061	-0.126	-0.058	0.008	-0.055											
19	-0.026	0.020	-0.062	-0.068	0.023	-0.094	-0.056	-0.066	0.031	-0.038	-0.052	-0.049	0.024	-0.104	-0.065	-0.056										
20	-0.038	-0.064	-0.073	-0.033	-0.067	-0.015	-0.080	-0.045	-0.060	-0.068	-0.088	-0.095	-0.090	-0.003	-0.015	0.014	-0.053									
21	-0.062	-0.044	-0.069	-0.078	-0.097	-0.020	-0.106	-0.028	-0.083	-0.061	-0.031	-0.029	-0.063	0.001	-0.009	-0.041	-0.045	-0.060								
23	-0.023	-0.068	0.001	-0.044	-0.058	-0.030	-0.050	-0.034	-0.071	-0.095	-0.076	-0.055	-0.052	-0.065	-0.085	-0.022	-0.008	-0.057	-0.066							
24	-0.065	0.005	-0.069	-0.100	-0.064	-0.107	-0.062	-0.089	-0.036	0.012	0.031	0.013	0.080	-0.042	-0.050	-0.134	-0.021	-0.022	-0.046	-0.080						
25	-0.043	-0.019	-0.081	-0.027	-0.002	-0.105	-0.064	-0.135	-0.042	0.043	0.047	0.011	0.017	-0.104	-0.069	-0.124	-0.036	-0.073	-0.016	-0.048	0.097					
27	-0.051	-0.082	-0.071	-0.121	-0.106	-0.069	-0.059	-0.036	-0.056	-0.064	-0.082	-0.058	-0.108	0.039	-0.023	-0.060	-0.043	-0.022	-0.033	-0.043	-0.041	-0.059				

Table XL DIF Contrasts for Gender Subgroups in Form A

Item	Reference group	Difficulty	SE	Focal group	Difficulty	SE	DIF contrast	Joint SE	$\chi^2$	<i>p</i>
1	Males	-2.91	0.16	Females	-3.08	0.18	0.17	0.24	0.44	0.506
2	Males	-1.82	0.11	Females	-1.54	0.11	-0.28	-0.16	0.33	0.568
3	Males	-1.42	0.10	Females	-1.04	0.10	-0.38	-0.14	3.69	0.055
4	Males	-1.34	0.10	Females	-1.43	0.11	0.09	0.15	0.48	0.489
6	Males	-0.74	0.09	Females	-0.54	0.10	-0.19	-0.13	1.57	0.210
7	Males	-1.71	0.11	Females	-1.33	0.11	-0.38	-0.15	2.49	0.114
8	Males	0.14	0.08	Females	0.02	0.09	0.12	0.13	1.06	0.303
9	Males	-0.44	0.09	Females	-0.23	0.10	-0.21	-0.13	0.26	0.613
10	Males	-0.56	0.09	Females	-0.93	0.10	0.37	0.13	8.83	0.003
11	Males	-0.32	0.08	Females	-0.07	0.09	-0.25	-0.13	1.03	0.310
12	Males	0.19	0.08	Females	-0.04	0.09	0.23	0.13	0.55	0.458
13	Males	0.31	0.08	Females	0.08	0.09	0.23	0.13	2.67	0.102
14	Males	1.02	0.08	Females	1.20	0.10	-0.18	-0.13	1.39	0.238
15	Males	1.25	0.08	Females	1.04	0.10	0.2	0.13	1.76	0.185
16	Males	1.32	0.08	Females	0.86	0.10	0.45	0.13	8.20	0.004
17	Males	-0.08	0.08	Females	0.01	0.09	-0.08	0.13	0.01	0.933
18	Males	0.38	0.08	Females	0.25	0.09	0.13	0.12	0.18	0.675
19	Males	0.65	0.08	Females	0.30	0.09	0.35	0.12	5.88	0.015
20	Males	0.22	0.08	Females	0.15	0.09	0.07	0.13	0.01	0.909
21	Males	0.05	0.08	Females	-0.04	0.09	0.09	0.13	0.00	0.970
22	Males	0.97	0.08	Females	1.23	0.10	-0.26	-0.13	4.39	0.036
23	Males	1.26	0.08	Females	1.74	0.10	-0.48	-0.13	13.89	0.000
24	Males	0.19	0.08	Females	0.33	0.09	-0.14	-0.12	3.76	0.052
25	Males	0.84	0.08	Females	0.61	0.09	0.23	0.13	1.59	0.207
26	Males	1.42	0.08	Females	1.16	0.10	0.26	0.13	0.92	0.338
28	Males	1.00	0.08	Females	1.39	0.10	-0.39	-0.13	6.75	0.009

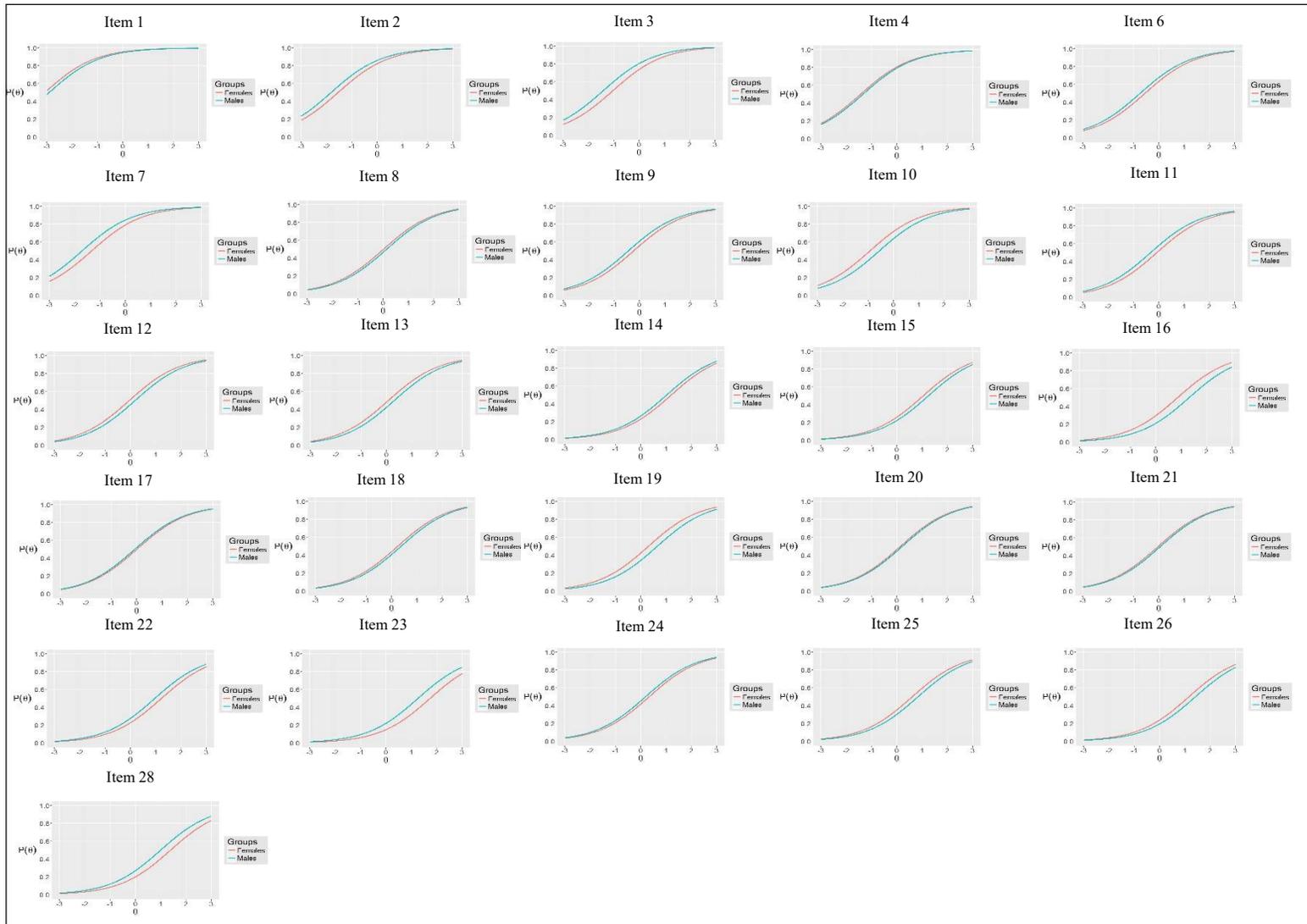


Figure 43 Item Characteristic Curves for Gender DIF Subgroups in Form A

Table XLI DIF Contrasts for L1 Subgroups in Form A

Item	Reference group	Difficulty	SE	Focal group	Difficulty	SE	DIF contrast	Joint SE	$\chi^2$	<i>p</i>
1	French	-2.01	0.21	Arabic	-3.44	0.39	1.42	0.44	11.19	0.001
	French	-2.01	0.21	Spanish	-4.80	0.46	2.78	0.51	21.63	0.000
2	French	-3.98	0.51	Arabic	-3.05	0.33	-0.93	0.60	0.35	0.553
	French	-3.98	0.51	Spanish	-0.70	0.16	-3.28	0.53	19.33	0.000
3	French	-1.36	0.17	Arabic	-1.65	0.20	0.29	0.26	2.01	0.156
	French	-1.36	0.17	Spanish	-0.90	0.16	-0.47	0.23	1.27	0.259
4	French	-0.88	0.14	Arabic	-1.33	0.18	0.45	0.23	4.07	0.044
	French	-0.88	0.14	Spanish	-1.78	0.17	0.90	0.22	6.28	0.012
6	French	-1.56	0.18	Arabic	-1.00	0.17	-0.56	0.25	0.64	0.425
	French	-1.56	0.18	Spanish	0.00	0.17	-1.56	0.25	7.15	0.008
7	French	-2.37	0.24	Arabic	-2.32	0.25	-0.04	0.35	0.37	0.543
	French	-2.37	0.24	Spanish	-0.75	0.16	-1.62	0.29	8.57	0.003
8	French	0.36	0.11	Arabic	-0.15	0.14	0.52	0.18	8.43	0.004
	French	0.36	0.11	Spanish	-0.13	0.17	0.50	0.20	0.30	0.583
9	French	-1.04	0.15	Arabic	-0.60	0.15	-0.44	0.21	0.48	0.489
	French	-1.04	0.15	Spanish	0.37	0.19	-1.41	0.24	2.30	0.129
10	French	-0.60	0.13	Arabic	-0.85	0.16	0.26	0.21	2.02	0.155
	French	-0.60	0.13	Spanish	-0.67	0.16	0.07	0.21	1.50	0.220
11	French	-0.84	0.14	Arabic	-0.33	0.15	-0.51	0.20	1.36	0.244
	French	-0.84	0.14	Spanish	0.65	0.20	-1.48	0.24	4.22	0.040
12	French	0.27	0.11	Arabic	0.24	0.14	0.03	0.18	0.03	0.860
	French	0.27	0.11	Spanish	-0.45	0.16	0.73	0.20	0.37	0.541
13	French	0.55	0.10	Arabic	0.19	0.14	0.36	0.17	5.40	0.020
	French	0.55	0.10	Spanish	-0.19	0.17	0.74	0.20	5.62	0.018
14	French	1.32	0.10	Arabic	0.83	0.14	0.49	0.17	2.33	0.127
	French	1.32	0.10	Spanish	1.25	0.23	0.07	0.25	0.16	0.690
15	French	1.24	0.10	Arabic	1.67	0.14	-0.42	0.17	4.38	0.036
	French	1.24	0.10	Spanish	1.38	0.24	-0.14	0.26	7.27	0.007
16	French	1.00	0.10	Arabic	1.45	0.14	-0.44	0.17	8.19	0.004
	French	1.00	0.10	Spanish	0.98	0.21	0.03	0.23	1.24	0.265
17	French	-0.60	0.13	Arabic	-0.30	0.15	-0.30	0.20	1.20	0.273
	French	-0.60	0.13	Spanish	1.16	0.22	-1.75	0.26	32.91	0.000
18	French	0.59	0.10	Arabic	0.52	0.14	0.07	0.17	2.61	0.106
	French	0.59	0.10	Spanish	-0.07	0.17	0.66	0.20	2.37	0.124
19	French	0.62	0.10	Arabic	0.58	0.14	0.05	0.17	0.47	0.494
	French	0.62	0.10	Spanish	-0.07	0.17	0.70	0.20	5.16	0.023
20	French	0.36	0.11	Arabic	0.26	0.14	0.10	0.18	0.00	0.953
	French	0.36	0.11	Spanish	-0.01	0.17	0.38	0.20	0.82	0.365
21	French	0.21	0.11	Arabic	0.20	0.14	0.01	0.18	0.38	0.535
	French	0.21	0.11	Spanish	-0.22	0.17	0.43	0.20	0.20	0.655
22	French	1.30	0.10	Arabic	1.13	0.14	0.17	0.17	0.95	0.329
	French	1.30	0.10	Spanish	0.85	0.20	0.45	0.23	1.96	0.161
23	French	1.39	0.10	Arabic	1.70	0.14	-0.31	0.18	1.61	0.204
	French	1.39	0.10	Spanish	1.42	0.24	-0.03	0.26	0.92	0.338
24	French	0.42	0.11	Arabic	0.88	0.14	-0.46	0.17	12.67	0.000
	French	0.42	0.11	Spanish	-1.08	0.16	1.50	0.19	5.19	0.023
25	French	0.75	0.10	Arabic	1.08	0.14	-0.33	0.17	8.78	0.003
	French	0.75	0.10	Spanish	0.26	0.18	0.49	0.21	0.00	0.945
26	French	1.80	0.10	Arabic	1.01	0.14	0.80	0.17	16.67	0.000
	French	1.80	0.10	Spanish	0.89	0.21	0.91	0.23	5.25	0.022
28	French	0.94	0.10	Arabic	1.54	0.14	-0.60	0.17	7.72	0.005
	French	0.94	0.10	Spanish	0.92	0.21	0.02	0.23	0.32	0.573

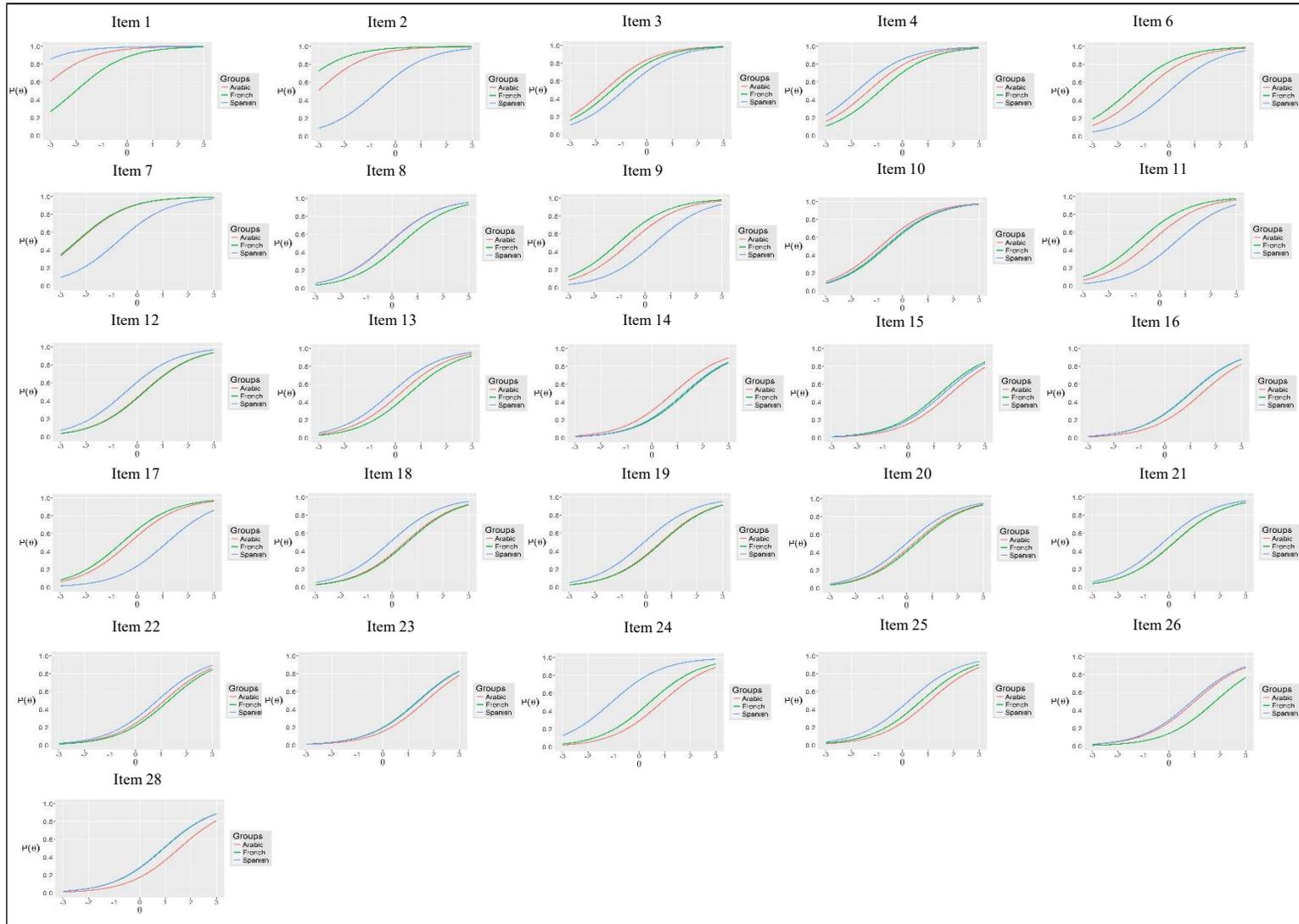


Figure 44 Item Characteristic Curves for L1 DIF Subgroups in Form A

Table XLII DIF Contrasts for Age Subgroups by Immigration Policy in Form A

Item	Reference group	Difficulty	SE	Focal group	Difficulty	SE	DIF contrast	Joint SE	$\chi^2$	<i>p</i>
1	Age group A	-3.40	0.25	Age group B	-2.94	0.18	-0.46	0.31	3.11	0.078
	Age group A	-3.40	0.25	Age group C	-2.72	0.22	-0.68	0.33	6.88	0.009
2	Age group A	-1.76	0.14	Age group B	-1.52	0.12	-0.25	0.18	0.15	0.695
	Age group A	-1.76	0.14	Age group C	-1.84	0.17	0.08	0.22	1.78	0.183
3	Age group A	-1.18	0.12	Age group B	-1.29	0.11	0.10	0.17	3.30	0.069
	Age group A	-1.18	0.12	Age group C	-1.28	0.15	0.10	0.19	1.62	0.203
4	Age group A	-1.38	0.13	Age group B	-1.44	0.12	0.06	0.17	0.01	0.936
	Age group A	-1.38	0.13	Age group C	-1.26	0.15	-0.12	0.19	0.18	0.675
6	Age group A	-0.75	0.11	Age group B	-0.47	0.10	-0.28	0.15	2.10	0.148
	Age group A	-0.75	0.11	Age group C	-0.81	0.14	0.06	0.18	0.77	0.379
7	Age group A	-1.56	0.13	Age group B	-1.66	0.12	0.10	0.18	2.15	0.143
	Age group A	-1.56	0.13	Age group C	-1.28	0.15	-0.28	0.20	0.62	0.430
8	Age group A	0.05	0.10	Age group B	0.09	0.10	-0.04	0.14	0.02	0.897
	Age group A	0.05	0.10	Age group C	0.15	0.13	-0.10	0.16	0.30	0.583
9	Age group A	-0.50	0.11	Age group B	-0.30	0.10	-0.21	0.15	0.60	0.440
	Age group A	-0.50	0.11	Age group C	-0.18	0.13	-0.32	0.17	1.55	0.214
10	Age group A	-0.86	0.11	Age group B	-0.59	0.10	-0.27	0.15	2.25	0.133
	Age group A	-0.86	0.11	Age group C	-0.76	0.14	-0.10	0.18	0.25	0.618
11	Age group A	-0.48	0.11	Age group B	-0.14	0.10	-0.34	0.15	2.57	0.109
	Age group A	-0.48	0.11	Age group C	0.08	0.13	-0.56	0.17	7.99	0.005
12	Age group A	0.22	0.10	Age group B	0.02	0.10	0.20	0.14	0.08	0.776
	Age group A	0.22	0.10	Age group C	-0.01	0.13	0.23	0.16	0.09	0.766
13	Age group A	0.21	0.10	Age group B	0.27	0.10	-0.05	0.14	0.17	0.684
	Age group A	0.21	0.10	Age group C	0.14	0.13	0.07	0.16	0.21	0.644
14	Age group A	1.10	0.10	Age group B	1.21	0.10	-0.11	0.14	0.18	0.676
	Age group A	1.10	0.10	Age group C	0.90	0.13	0.20	0.17	1.27	0.259
15	Age group A	1.20	0.10	Age group B	1.05	0.10	0.15	0.14	0.21	0.647
	Age group A	1.20	0.10	Age group C	1.28	0.13	-0.08	0.17	0.99	0.321
16	Age group A	1.12	0.10	Age group B	1.06	0.10	0.06	0.14	0.26	0.608
	Age group A	1.12	0.10	Age group C	1.22	0.13	-0.10	0.17	0.15	0.695
17	Age group A	-0.19	0.10	Age group B	-0.07	0.10	-0.11	0.14	0.01	0.923
	Age group A	-0.19	0.10	Age group C	0.27	0.13	-0.46	0.17	4.75	0.029
18	Age group A	0.28	0.10	Age group B	0.33	0.10	-0.05	0.14	0.48	0.489
	Age group A	0.28	0.10	Age group C	0.39	0.13	-0.11	0.16	0.16	0.693
19	Age group A	0.55	0.10	Age group B	0.38	0.10	0.17	0.14	0.92	0.338
	Age group A	0.55	0.10	Age group C	0.65	0.13	-0.10	0.16	0.32	0.574
20	Age group A	0.22	0.10	Age group B	0.19	0.10	0.03	0.14	0.19	0.659
	Age group A	0.22	0.10	Age group C	0.19	0.13	0.03	0.16	0.32	0.571
21	Age group A	0.15	0.10	Age group B	-0.06	0.10	0.21	0.14	0.61	0.433
	Age group A	0.15	0.10	Age group C	-0.05	0.13	0.20	0.16	0.63	0.429
22	Age group A	1.12	0.10	Age group B	1.17	0.10	-0.05	0.14	0.01	0.923
	Age group A	1.12	0.10	Age group C	0.82	0.13	0.30	0.17	2.58	0.108
23	Age group A	1.62	0.10	Age group B	1.35	0.11	0.27	0.15	1.29	0.256
	Age group A	1.62	0.10	Age group C	1.35	0.14	0.27	0.17	1.16	0.282
24	Age group A	0.54	0.10	Age group B	0.22	0.10	0.32	0.14	2.73	0.098
	Age group A	0.54	0.10	Age group C	-0.21	0.13	0.76	0.16	15.55	0.000
25	Age group A	0.74	0.10	Age group B	0.72	0.10	0.02	0.14	0.03	0.854
	Age group A	0.74	0.10	Age group C	0.81	0.13	-0.07	0.16	0.47	0.491
26	Age group A	1.18	0.10	Age group B	1.34	0.10	-0.16	0.14	3.13	0.077
	Age group A	1.18	0.10	Age group C	1.48	0.14	-0.30	0.17	4.02	0.045
28	Age group A	1.33	0.10	Age group B	1.13	0.10	0.20	0.14	2.95	0.086
	Age group A	1.33	0.10	Age group C	0.90	0.13	0.43	0.17	5.71	0.017

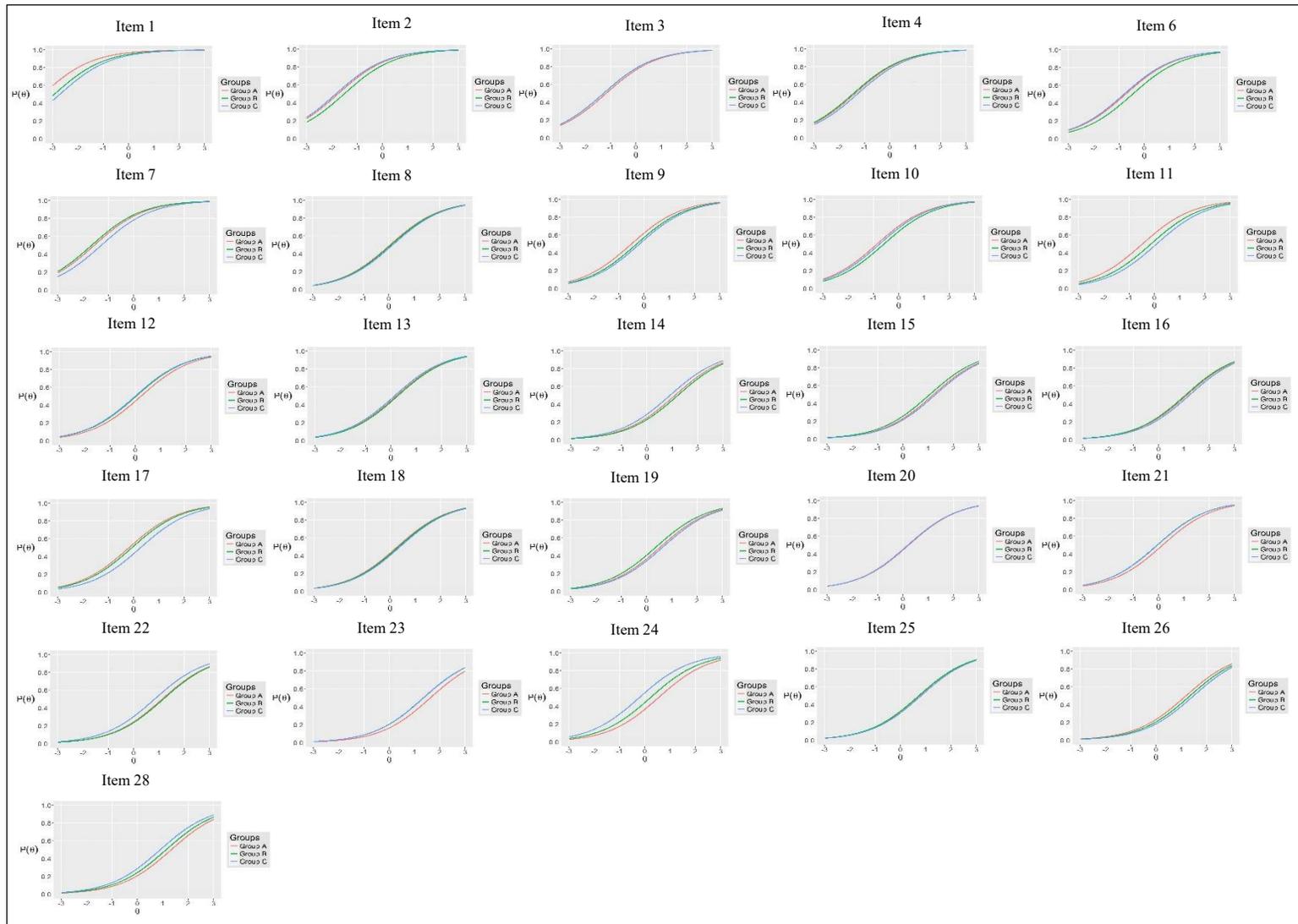


Figure 45 Item Characteristic Curves for Age DIF Subgroups by Immigration Policy in Form A

Table XLIII DIF Contrasts for Age Subgroups by Similar Sample Sizes in Form A

Item	Reference group	Difficulty	SE	Focal group	Difficulty	SE	DIF contrast	Joint SE	$\chi^2$	<i>p</i>
1	Age Group D	-3.36	0.35	Age Group E	-3.25	0.26	-0.11	0.44	0.01	0.919
	Age Group D	-3.36	0.35	Age Group F	-2.88	0.20	-0.48	0.40	2.17	0.141
	Age Group D	-3.36	0.35	Age Group G	-2.72	0.22	-0.64	0.41	3.50	0.061
2	Age Group D	-2.06	0.22	Age Group E	-1.60	0.15	-0.46	0.27	0.86	0.354
	Age Group D	-2.06	0.22	Age Group F	-1.48	0.13	-0.58	0.25	0.95	0.329
	Age Group D	-2.06	0.22	Age Group G	-1.84	0.17	-0.22	0.27	0.01	0.942
3	Age Group D	-1.33	0.18	Age Group E	-1.21	0.14	-0.12	0.23	0.22	0.643
	Age Group D	-1.33	0.18	Age Group F	-1.22	0.13	-0.11	0.22	0.01	0.918
	Age Group D	-1.33	0.18	Age Group G	-1.28	0.15	-0.05	0.23	0.02	0.882
4	Age Group D	-1.50	0.19	Age Group E	-1.38	0.14	-0.13	0.24	0.99	0.321
	Age Group D	-1.50	0.19	Age Group F	-1.42	0.13	-0.08	0.23	0.60	0.440
	Age Group D	-1.50	0.19	Age Group G	-1.26	0.15	-0.24	0.24	0.92	0.338
6	Age Group D	-0.74	0.16	Age Group E	-0.69	0.13	-0.05	0.21	0.00	0.998
	Age Group D	-0.74	0.16	Age Group F	-0.46	0.11	-0.28	0.20	1.11	0.291
	Age Group D	-0.74	0.16	Age Group G	-0.81	0.14	0.07	0.21	0.58	0.446
7	Age Group D	-1.45	0.19	Age Group E	-1.72	0.16	0.27	0.24	0.91	0.340
	Age Group D	-1.45	0.19	Age Group F	-1.62	0.14	0.17	0.23	2.99	0.084
	Age Group D	-1.45	0.19	Age Group G	-1.28	0.15	-0.16	0.24	0.08	0.773
8	Age Group D	0.03	0.15	Age Group E	0.06	0.12	-0.02	0.19	0.00	0.975
	Age Group D	0.03	0.15	Age Group F	0.09	0.11	-0.05	0.18	0.00	0.965
	Age Group D	0.03	0.15	Age Group G	0.15	0.13	-0.12	0.19	0.05	0.828
9	Age Group D	-0.49	0.15	Age Group E	-0.52	0.13	0.03	0.20	0.21	0.649
	Age Group D	-0.49	0.15	Age Group F	-0.25	0.11	-0.24	0.19	1.00	0.317
	Age Group D	-0.49	0.15	Age Group G	-0.18	0.13	-0.31	0.20	1.65	0.199
10	Age Group D	-1.03	0.17	Age Group E	-0.61	0.13	-0.42	0.21	2.69	0.101
	Age Group D	-1.03	0.17	Age Group F	-0.63	0.12	-0.40	0.20	1.87	0.172
	Age Group D	-1.03	0.17	Age Group G	-0.76	0.14	-0.27	0.22	1.14	0.287
11	Age Group D	-0.48	0.15	Age Group E	-0.31	0.12	-0.16	0.20	0.03	0.867
	Age Group D	-0.48	0.15	Age Group F	-0.21	0.11	-0.26	0.19	0.07	0.793
	Age Group D	-0.48	0.15	Age Group G	0.08	0.13	-0.56	0.20	2.38	0.123
12	Age Group D	0.22	0.14	Age Group E	0.24	0.12	-0.02	0.19	0.00	0.982
	Age Group D	0.22	0.14	Age Group F	-0.04	0.11	0.27	0.18	0.17	0.677
	Age Group D	0.22	0.14	Age Group G	-0.01	0.13	0.24	0.19	0.02	0.886
13	Age Group D	0.11	0.14	Age Group E	0.34	0.12	-0.24	0.19	0.98	0.323
	Age Group D	0.11	0.14	Age Group F	0.21	0.11	-0.10	0.18	0.07	0.796
	Age Group D	0.11	0.14	Age Group G	0.14	0.13	-0.03	0.19	0.01	0.942
14	Age Group D	1.03	0.14	Age Group E	1.13	0.12	-0.11	0.19	0.18	0.673
	Age Group D	1.03	0.14	Age Group F	1.26	0.12	-0.24	0.18	0.67	0.412
	Age Group D	1.03	0.14	Age Group G	0.90	0.13	0.12	0.19	0.36	0.549

Table XLIII continued

	Age Group D	1.26	0.14	Age Group E	1.02	0.12	0.25	0.19	0.93	0.336
15	Age Group D	1.26	0.14	Age Group F	1.13	0.11	0.13	0.18	0.03	0.856
	Age Group D	1.26	0.14	Age Group G	1.28	0.13	-0.02	0.20	0.47	0.492
	Age Group D	1.19	0.14	Age Group E	1.01	0.12	0.18	0.19	1.13	0.288
16	Age Group D	1.19	0.14	Age Group F	1.12	0.11	0.07	0.18	0.47	0.494
	Age Group D	1.19	0.14	Age Group G	1.22	0.13	-0.03	0.20	0.01	0.921
	Age Group D	-0.19	0.15	Age Group E	-0.22	0.12	0.03	0.19	0.13	0.721
17	Age Group D	-0.19	0.15	Age Group F	-0.04	0.11	-0.15	0.19	0.10	0.747
	Age Group D	-0.19	0.15	Age Group G	0.27	0.13	-0.46	0.20	4.17	0.041
	Age Group D	0.42	0.14	Age Group E	0.21	0.12	0.21	0.19	1.08	0.300
18	Age Group D	0.42	0.14	Age Group F	0.33	0.11	0.10	0.18	0.00	0.945
	Age Group D	0.42	0.14	Age Group G	0.38	0.13	0.04	0.19	0.20	0.651
	Age Group D	0.59	0.14	Age Group E	0.47	0.12	0.11	0.18	0.06	0.801
19	Age Group D	0.59	0.14	Age Group F	0.37	0.11	0.21	0.18	0.45	0.502
	Age Group D	0.59	0.14	Age Group G	0.65	0.13	-0.06	0.19	0.11	0.742
	Age Group D	0.02	0.15	Age Group E	0.40	0.12	-0.37	0.19	2.30	0.129
20	Age Group D	0.02	0.15	Age Group F	0.11	0.11	-0.09	0.18	0.97	0.325
	Age Group D	0.02	0.15	Age Group G	0.17	0.13	-0.14	0.20	2.08	0.149
	Age Group D	0.20	0.14	Age Group E	0.09	0.12	0.11	0.19	0.01	0.933
21	Age Group D	0.20	0.14	Age Group F	-0.10	0.11	0.30	0.18	0.55	0.457
	Age Group D	0.20	0.14	Age Group G	-0.05	0.13	0.25	0.19	1.05	0.305
	Age Group D	1.26	0.14	Age Group E	1.07	0.12	0.19	0.19	0.31	0.575
22	Age Group D	1.26	0.14	Age Group F	1.15	0.11	0.11	0.18	0.08	0.780
	Age Group D	1.26	0.14	Age Group G	0.82	0.13	0.44	0.19	3.49	0.062
	Age Group D	1.51	0.14	Age Group E	1.63	0.13	-0.12	0.19	0.42	0.518
23	Age Group D	1.51	0.14	Age Group F	1.34	0.12	0.16	0.19	0.05	0.815
	Age Group D	1.51	0.14	Age Group G	1.35	0.14	0.15	0.20	0.08	0.778
	Age Group D	0.69	0.14	Age Group E	0.38	0.12	0.31	0.18	2.84	0.092
24	Age Group D	0.69	0.14	Age Group F	0.22	0.11	0.47	0.18	3.25	0.072
	Age Group D	0.69	0.14	Age Group G	-0.22	0.13	0.90	0.19	17.48	0.000
	Age Group D	0.65	0.14	Age Group E	0.87	0.12	-0.22	0.18	0.78	0.377
25	Age Group D	0.65	0.14	Age Group F	0.64	0.11	0.01	0.18	0.02	0.897
	Age Group D	0.65	0.14	Age Group G	0.81	0.13	-0.16	0.19	0.85	0.356
	Age Group D	1.18	0.14	Age Group E	1.22	0.12	-0.04	0.19	0.20	0.654
26	Age Group D	1.18	0.14	Age Group F	1.33	0.12	-0.15	0.18	1.28	0.258
	Age Group D	1.18	0.14	Age Group G	1.48	0.14	-0.30	0.20	2.86	0.091
	Age Group D	1.56	0.14	Age Group E	1.13	0.12	0.43	0.19	6.05	0.014
28	Age Group D	1.56	0.14	Age Group F	1.13	0.11	0.43	0.18	6.65	0.010
	Age Group D	1.56	0.14	Age Group G	0.90	0.13	0.66	0.20	10.23	0.001

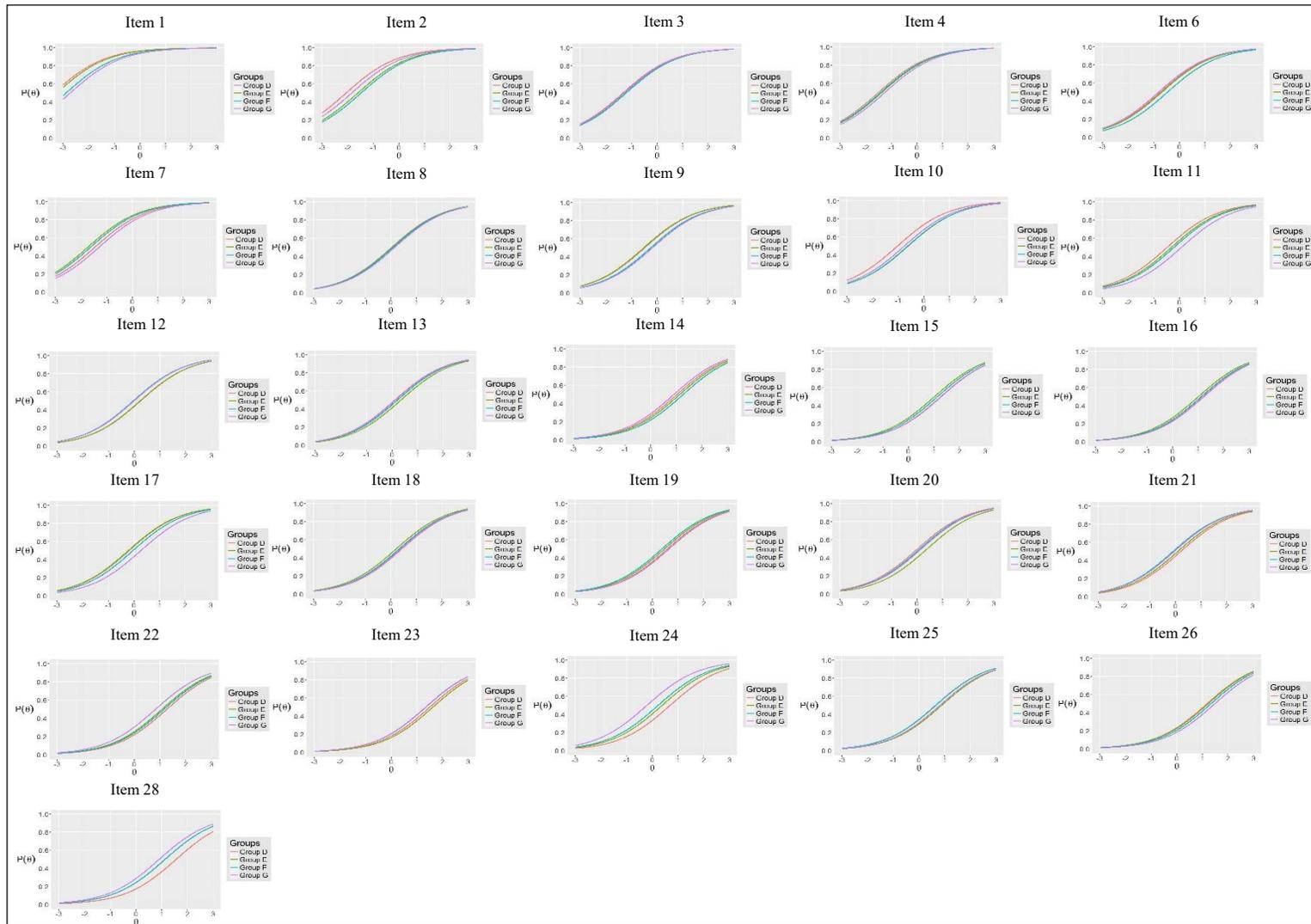


Figure 46 Item Characteristic Curves for Age DIF Subgroups by Similar Sample Sizes in Form A

Table XLIV DIF Contrasts for Geographical Location Subgroups in Form A

Item	Reference group	Difficulty	SE	Focal group	Difficulty	SE	DIF contrast	Joint SE	$\chi^2$	<i>p</i>
1	North Africa	-3.02	0.39	West Africa	-1.93	0.20	-1.09	0.44	3.87	0.049
2	North Africa	-5.03	1.00	West Africa	-3.61	0.42	-1.42	1.09	0.50	0.482
3	North Africa	-1.97	0.25	West Africa	-1.27	0.16	-0.70	0.30	4.77	0.029
4	North Africa	-1.27	0.20	West Africa	-0.64	0.13	-0.63	0.24	7.83	0.005
6	North Africa	-1.31	0.20	West Africa	-1.25	0.16	-0.06	0.26	0.30	0.583
7	North Africa	-2.43	0.30	West Africa	-2.10	0.21	-0.33	0.37	0.98	0.321
8	North Africa	-0.25	0.16	West Africa	0.81	0.11	-1.06	0.19	23.76	0.000
9	North Africa	-0.64	0.17	West Africa	-0.64	0.13	0.00	0.22	0.00	0.958
10	North Africa	-0.92	0.18	West Africa	-0.57	0.13	-0.35	0.23	0.97	0.324
11	North Africa	-0.48	0.16	West Africa	-0.76	0.14	0.28	0.21	0.47	0.494
12	North Africa	0.63	0.14	West Africa	0.43	0.11	0.20	0.18	0.39	0.535
13	North Africa	0.37	0.14	West Africa	0.80	0.11	-0.43	0.18	4.93	0.026
14	North Africa	1.01	0.14	West Africa	1.33	0.11	-0.32	0.18	0.74	0.389
15	North Africa	1.62	0.14	West Africa	1.11	0.11	0.51	0.18	8.31	0.004
16	North Africa	1.59	0.14	West Africa	0.79	0.11	0.80	0.18	15.39	0.000
17	North Africa	-0.46	0.16	West Africa	-0.83	0.14	0.36	0.22	0.62	0.433
18	North Africa	0.66	0.14	West Africa	0.66	0.11	0.00	0.18	0.91	0.341
19	North Africa	0.67	0.14	West Africa	0.83	0.11	-0.16	0.18	0.67	0.412
20	North Africa	0.49	0.14	West Africa	0.21	0.11	0.28	0.18	2.25	0.133
21	North Africa	0.42	0.14	West Africa	0.18	0.11	0.24	0.18	1.62	0.203
22	North Africa	1.12	0.14	West Africa	1.60	0.11	-0.48	0.18	5.44	0.020
23	North Africa	1.61	0.14	West Africa	1.61	0.11	0.00	0.18	0.00	0.959
24	North Africa	0.90	0.14	West Africa	0.37	0.11	0.53	0.18	9.43	0.002
25	North Africa	1.56	0.14	West Africa	0.43	0.11	1.13	0.18	31.73	0.000
26	North Africa	1.28	0.14	West Africa	2.00	0.11	-0.72	0.18	14.57	0.000
28	North Africa	1.59	0.14	West Africa	1.33	0.11	0.25	0.18	2.52	0.113

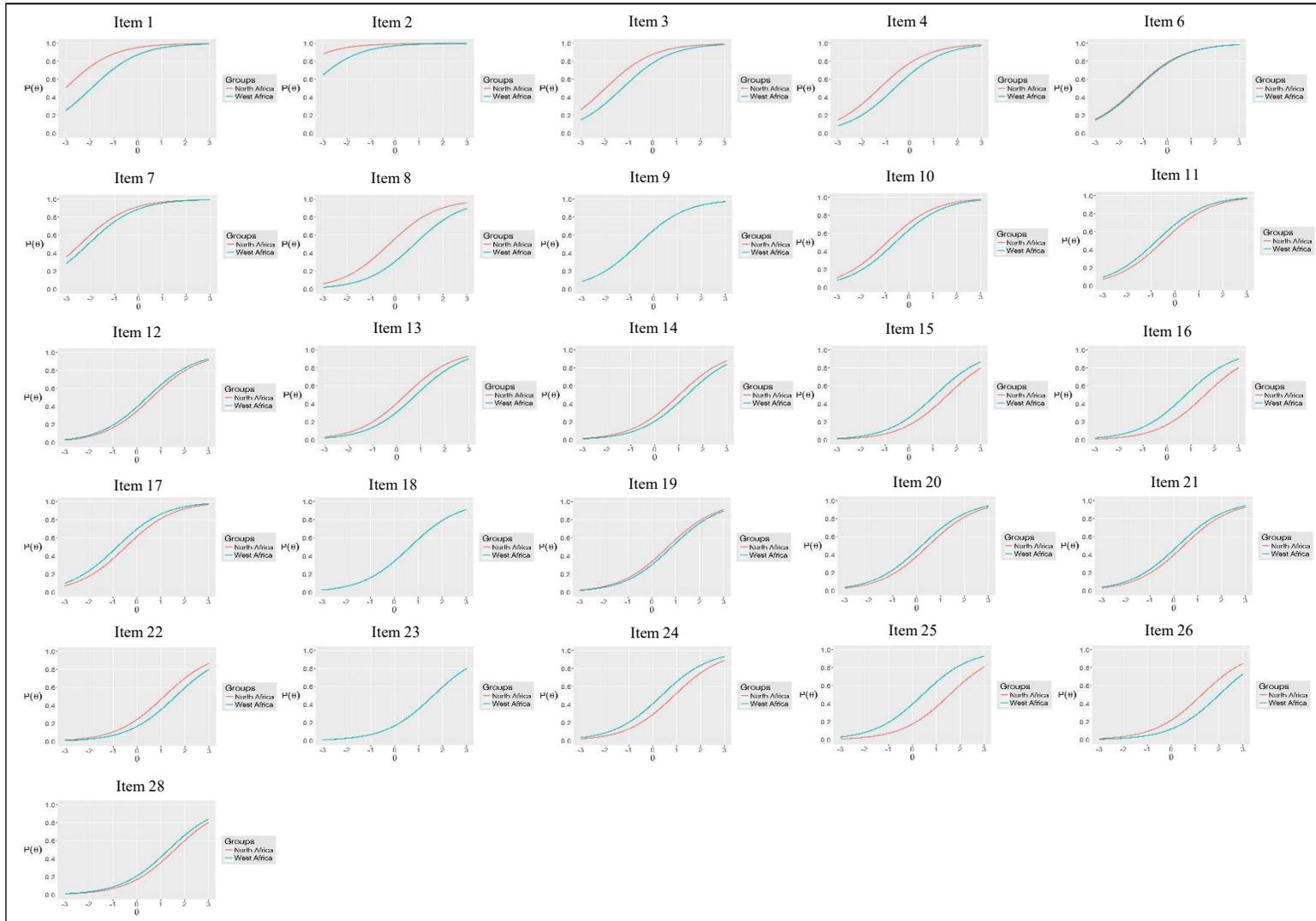


Figure 47 Item Characteristic Curves for Geographical Location DIF Subgroups in Form A

Table XLV DIF Contrasts for Gender Subgroups in Form B

Item	Reference group	Difficulty	SE	Focal group	Difficulty	SE	DIF contrast	Joint SE	$\chi^2$	<i>p</i>
1	Males	-1.85	0.10	Females	-1.85	0.10	0.00	0.14	0.06	0.803
4	Males	-0.56	0.07	Females	-0.49	0.08	-0.06	0.11	0.01	0.923
5	Males	-0.80	0.08	Females	-0.72	0.08	-0.08	0.11	0.25	0.618
6	Males	-1.40	0.09	Females	-1.25	0.09	-0.15	-0.12	0.65	0.420
7	Males	-1.43	0.09	Females	-1.53	0.09	0.10	0.13	3.58	0.059
8	Males	-0.73	0.08	Females	-0.73	0.08	0.00	0.11	0.05	0.825
9	Males	-0.63	0.07	Females	-1.01	0.08	0.38	0.11	12.37	0.000
10	Males	0.31	0.07	Females	-0.04	0.08	0.35	0.10	4.13	0.042
11	Males	-0.35	0.07	Females	-0.30	0.08	-0.04	0.11	0.01	0.916
12	Males	-0.72	0.08	Females	-0.87	0.08	0.14	0.11	3.03	0.082
13	Males	-0.32	0.07	Females	-0.06	0.08	-0.26	-0.11	2.53	0.112
14	Males	0.32	0.07	Females	0.24	0.08	0.08	0.10	0.99	0.321
15	Males	0.36	0.07	Females	0.31	0.08	0.05	0.10	1.22	0.270
16	Males	1.12	0.07	Females	1.06	0.08	0.06	0.11	0.26	0.607
17	Males	1.87	0.08	Females	1.48	0.08	0.39	0.11	5.29	0.022
18	Males	0.59	0.07	Females	0.83	0.08	-0.24	-0.11	8.39	0.004
19	Males	-0.07	0.07	Females	0.16	0.08	-0.24	-0.10	2.65	0.103
20	Males	0.89	0.07	Females	1.14	0.08	-0.25	-0.11	7.69	0.006
21	Males	0.55	0.07	Females	0.55	0.08	0.00	0.10	0.68	0.409
23	Males	0.71	0.07	Females	0.71	0.08	0.00	0.11	0.16	0.693
24	Males	0.60	0.07	Females	0.72	0.08	-0.11	-0.10	0.47	0.495
25	Males	0.00	0.07	Females	0.11	0.08	-0.11	-0.10	0.03	0.857
27	Males	1.57	0.07	Females	1.52	0.08	0.05	0.11	0.00	0.998

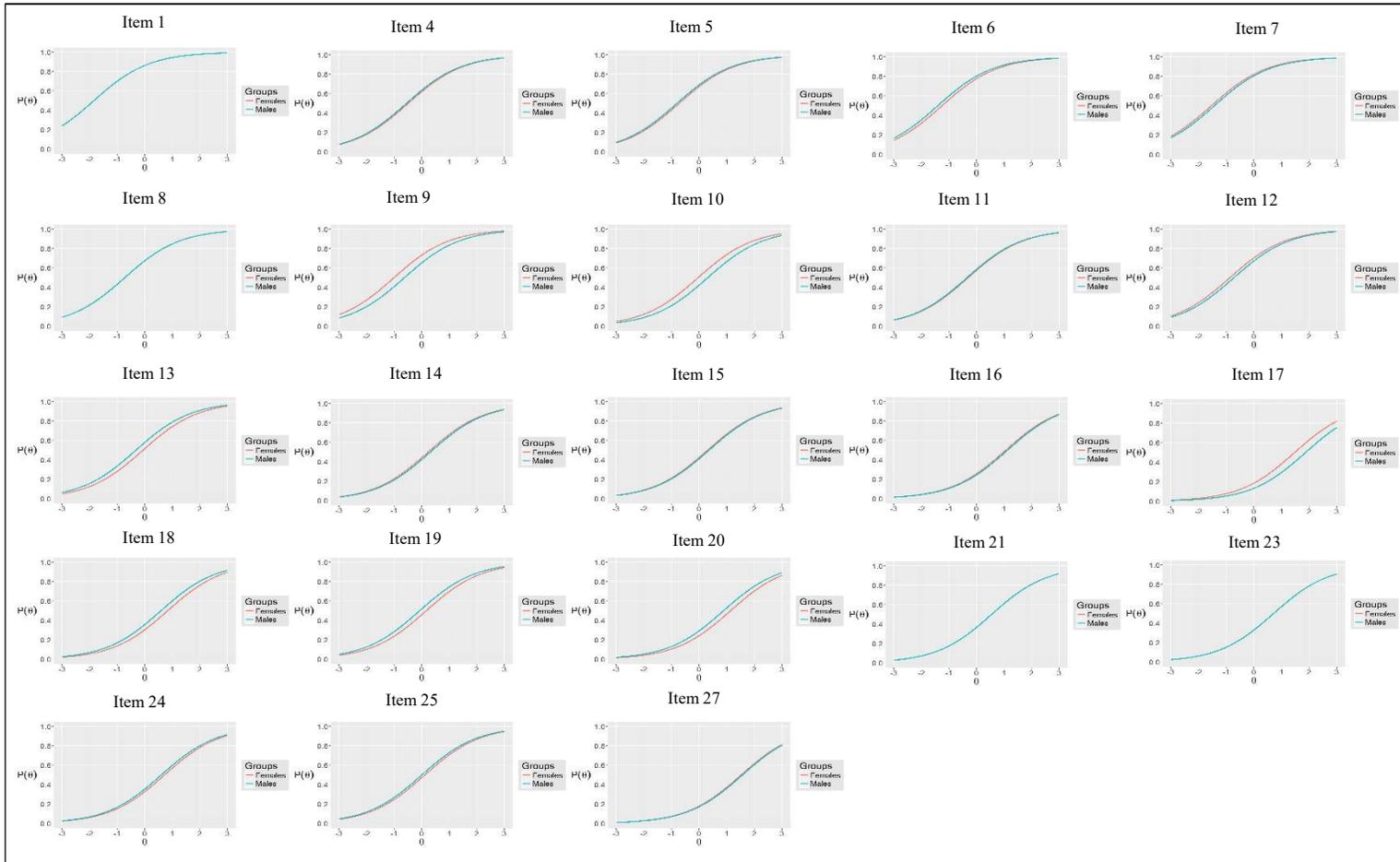


Figure 48 Item Characteristic Curves for Gender DIF in Form B

Table XLVI DIF Contrasts for L1 Subgroups in Form B

Item	Reference group	Difficulty	SE	Focal group	Difficulty	SE	DIF contrast	Joint SE	$\chi^2$	<i>p</i>
1	French	-1.65	0.18	Arabic	-2.45	0.19	0.80	0.26	7.98	0.005
	French	-1.65	0.18	Persian	-1.93	0.17	0.28	0.24	1.99	0.158
	French	-1.65	0.18	Spanish	-1.62	0.13	-0.03	0.22	0.00	0.961
4	French	-1.30	0.16	Arabic	0.20	0.11	-1.51	-0.19	43.33	0.000
	French	-1.30	0.16	Persian	-0.72	0.14	-0.58	-0.21	0.02	0.878
	French	-1.30	0.16	Spanish	-0.30	0.12	-1.00	-0.20	4.85	0.028
5	French	-0.18	0.11	Arabic	-1.10	0.13	0.92	0.17	23.06	0.000
	French	-0.18	0.11	Persian	-0.61	0.14	0.43	0.18	3.17	0.075
	French	-0.18	0.11	Spanish	-1.11	0.12	0.93	0.17	15.54	0.000
6	French	-1.68	0.18	Arabic	-1.54	0.15	-0.14	0.23	1.26	0.261
	French	-1.68	0.18	Persian	-0.97	0.15	-0.71	-0.23	0.02	0.879
	French	-1.68	0.18	Spanish	-1.21	0.12	-0.48	-0.22	2.95	0.086
7	French	-1.82	0.19	Arabic	-2.08	0.17	0.26	0.26	3.86	0.049
	French	-1.82	0.19	Persian	-1.12	0.15	-0.70	-0.24	0.02	0.877
	French	-1.82	0.19	Spanish	-1.59	0.13	-0.23	-0.23	1.52	0.218
8	French	0.21	0.10	Arabic	-0.82	0.13	1.03	0.16	18.43	0.000
	French	0.21	0.10	Persian	-0.77	0.14	0.98	0.18	3.99	0.046
	French	0.21	0.10	Spanish	-1.42	0.13	1.62	0.17	22.07	0.000
9	French	-0.83	0.13	Arabic	-0.92	0.13	0.09	0.18	0.01	0.931
	French	-0.83	0.13	Persian	-1.01	0.15	0.18	0.20	2.27	0.132
	French	-0.83	0.13	Spanish	-0.76	0.12	-0.07	0.18	1.06	0.303
10	French	0.94	0.09	Arabic	0.16	0.11	0.78	0.15	18.55	0.000
	French	0.94	0.09	Persian	-0.06	0.15	1.00	0.18	6.60	0.010
	French	0.94	0.09	Spanish	-0.86	0.12	1.81	.15 1	47.59	0.000
11	French	-0.81	0.13	Arabic	0.17	0.11	-0.98	-0.17	21.79	0.000
	French	-0.81	0.13	Persian	-0.11	0.15	-0.70	-0.20	3.83	0.050
	French	-0.81	0.13	Spanish	-0.81	0.12	0.00	0.18	2.39	0.122
12	French	-1.88	0.19	Arabic	-0.97	0.13	-0.91	-0.23	8.40	0.004
	French	-1.88	0.19	Persian	-1.07	0.15	-0.81	-0.24	1.10	0.295
	French	-1.88	0.19	Spanish	0.17	0.12	-2.05	-0.23	56.25	0.000
13	French	-0.78	0.13	Arabic	-0.56	0.12	-0.23	-0.18	0.19	0.664
	French	-0.78	0.13	Persian	0.31	0.16	-1.10	-0.20	6.70	0.010
	French	-0.78	0.13	Spanish	0.50	0.13	-1.28	-0.18	12.43	0.000
14	French	-0.21	0.11	Arabic	0.10	0.11	-0.30	-0.16	0.00	0.996
	French	-0.21	0.11	Persian	0.56	0.16	-0.77	-0.20	5.68	0.017
	French	-0.21	0.11	Spanish	0.77	0.13	-0.98	-0.18	14.24	0.000
15	French	-0.21	0.11	Arabic	0.23	0.11	-0.44	-0.16	0.00	0.964
	French	-0.21	0.11	Persian	0.43	0.16	-0.63	-0.20	0.18	0.672
	French	-0.21	0.11	Spanish	0.96	0.14	-1.16	-0.18	9.09	0.003
16	French	1.42	0.09	Arabic	1.23	0.12	0.20	0.15	0.01	0.918
	French	1.42	0.09	Persian	0.51	0.16	0.91	0.19	2.62	0.106
	French	1.42	0.09	Spanish	0.98	0.14	0.44	0.17	0.69	0.407

Table XLVI continued

	French	1.95	0.09	Arabic	1.94	0.13	0.00	0.16	1.94	0.164
17	French	1.95	0.09	Persian	1.04	0.19	0.91	0.21	0.01	0.931
	French	1.95	0.09	Spanish	1.40	0.16	0.55	0.18	0.05	0.825
	French	1.30	0.09	Arabic	0.85	0.11	0.45	0.15	1.30	0.255
18	French	1.30	0.09	Persian	-0.08	0.15	1.38	0.17	15.01	0.000
	French	1.30	0.09	Spanish	0.57	0.13	0.73	0.16	7.18	0.007
	French	-0.35	0.12	Arabic	0.22	0.11	-0.57	-0.16	2.17	0.141
19	French	-0.35	0.12	Persian	0.41	0.16	-0.76	-0.20	0.07	0.787
	French	-0.35	0.12	Spanish	-0.13	0.12	-0.22	-0.17	0.65	0.420
	French	1.32	0.09	Arabic	1.14	0.12	0.18	0.15	0.12	0.730
20	French	1.32	0.09	Persian	0.62	0.17	0.70	0.19	0.41	0.525
	French	1.32	0.09	Spanish	0.73	0.13	0.59	0.16	0.43	0.513
	French	0.82	0.09	Arabic	0.35	0.11	0.47	0.15	8.17	0.004
21	French	0.82	0.09	Persian	0.30	0.16	0.52	0.18	0.60	0.437
	French	0.82	0.09	Spanish	0.25	0.12	0.57	0.15	2.16	0.142
	French	0.85	0.09	Arabic	0.94	0.12	-0.09	0.15	0.99	0.319
23	French	0.85	0.09	Persian	1.32	0.21	-0.47	-0.23	1.09	0.296
	French	0.85	0.09	Spanish	0.09	0.12	0.76	0.15	8.72	0.003
	French	0.08	0.11	Arabic	0.52	0.11	-0.44	-0.15	0.50	0.478
24	French	0.08	0.11	Persian	0.47	0.16	-0.39	-0.19	0.01	0.929
	French	0.08	0.11	Spanish	1.86	0.18	-1.78	-0.21	22.64	0.000
	French	-1.03	0.14	Arabic	0.01	0.11	-1.03	-0.18	23.69	0.000
25	French	-1.03	0.14	Persian	0.60	0.17	-1.63	-0.22	28.58	0.000
	French	-1.03	0.14	Spanish	0.53	0.13	-1.56	-0.19	37.07	0.000
	French	1.79	0.09	Arabic	1.69	0.12	0.10	0.15	0.95	0.329
27	French	1.79	0.09	Persian	1.30	0.20	0.49	0.22	6.60	0.010
	French	1.79	0.09	Spanish	1.06	0.14	0.73	0.17	2.22	0.136

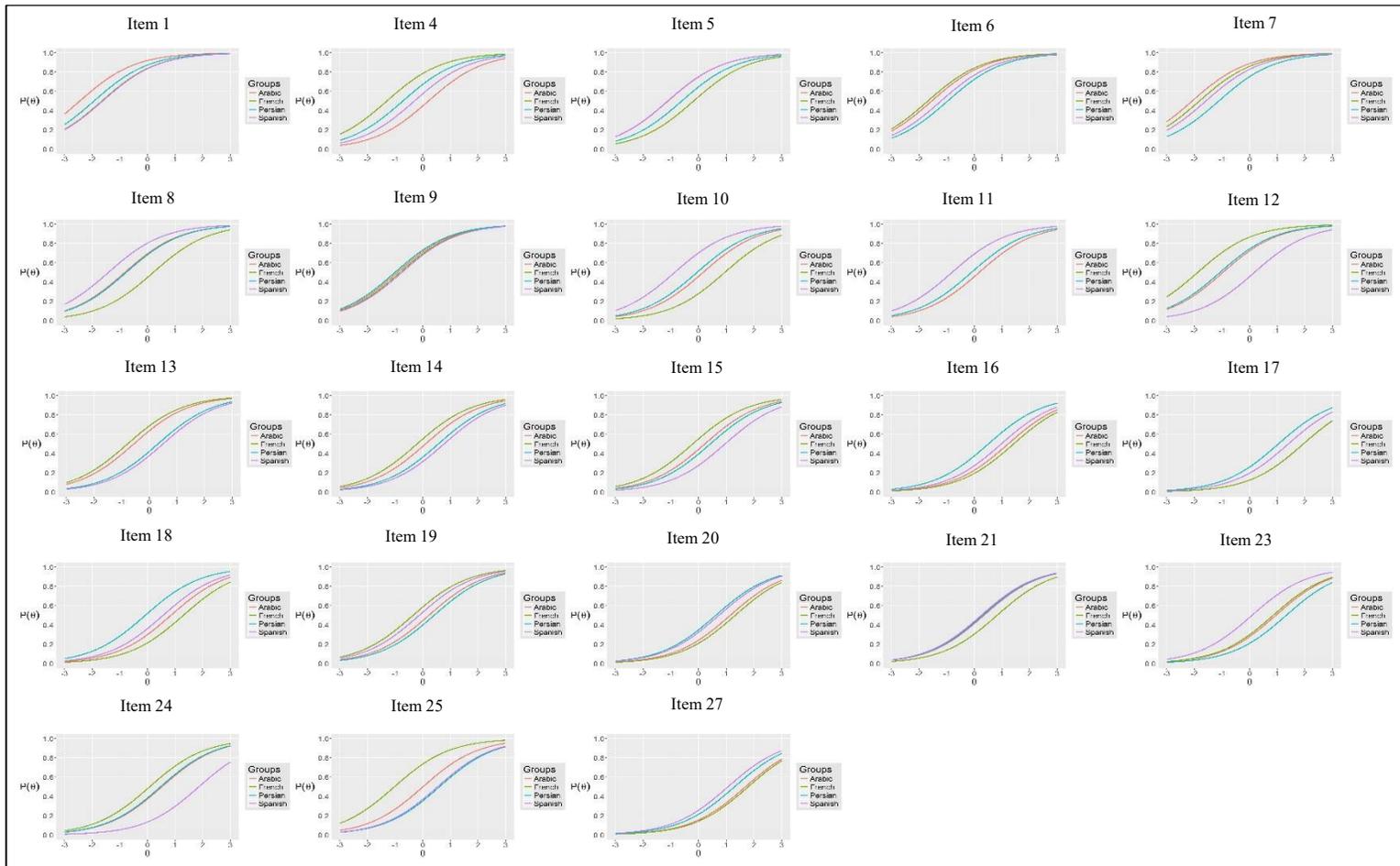


Figure 49 Item Characteristic Curves for L1 DIF Subgroups in Form B

Table XLVII DIF Contrasts for Age Subgroups by Immigration Policy in Form B

Item	Reference group	Difficulty	SE	Focal group	Difficulty	SE	DIF contrast	Joint SE	$\chi^2$	<i>p</i>
1	Age group A	-2.03	0.12	Age group B	-1.77	0.11	-0.25	0.16	2.47	0.116
	Age group A	-2.03	0.12	Age group C	-1.69	0.15	-0.33	0.19	3.17	0.075
4	Age group A	-0.50	0.08	Age group B	-0.55	0.09	0.05	0.12	0.18	0.674
	Age group A	-0.50	0.08	Age group C	-0.53	0.12	0.03	0.15	0.00	0.988
5	Age group A	-0.79	0.09	Age group B	-0.67	0.09	-0.11	0.12	0.77	0.380
	Age group A	-0.79	0.09	Age group C	-0.91	0.13	0.12	0.15	0.45	0.503
6	Age group A	-1.58	0.10	Age group B	-1.17	0.09	-0.41	0.14	9.78	0.002
	Age group A	-1.58	0.10	Age group C	-1.19	0.13	-0.39	0.17	6.35	0.012
7	Age group A	-1.57	0.10	Age group B	-1.42	0.10	-0.15	0.14	1.03	0.311
	Age group A	-1.57	0.10	Age group C	-1.41	0.14	-0.17	0.17	0.67	0.414
8	Age group A	-0.89	0.09	Age group B	-0.73	0.09	-0.16	0.12	0.58	0.445
	Age group A	-0.89	0.09	Age group C	-0.43	0.12	-0.46	0.15	5.70	0.017
9	Age group A	-0.83	0.09	Age group B	-0.91	0.09	0.09	0.13	0.69	0.406
	Age group A	-0.83	0.09	Age group C	-0.56	0.12	-0.27	0.15	3.36	0.067
10	Age group A	0.22	0.08	Age group B	0.15	0.08	0.06	0.12	0.01	0.912
	Age group A	0.22	0.08	Age group C	0.06	0.12	0.16	0.14	0.95	0.329
11	Age group A	-0.33	0.08	Age group B	-0.26	0.08	-0.07	0.12	0.00	0.994
	Age group A	-0.33	0.08	Age group C	-0.45	0.12	0.12	0.15	1.13	0.288
12	Age group A	-0.86	0.09	Age group B	-0.75	0.09	-0.11	0.12	0.36	0.547
	Age group A	-0.86	0.09	Age group C	-0.72	0.12	-0.14	0.15	0.55	0.458
13	Age group A	-0.26	0.08	Age group B	-0.12	0.08	-0.14	0.12	0.40	0.528
	Age group A	-0.26	0.08	Age group C	-0.25	0.12	-0.01	0.14	0.01	0.938
14	Age group A	0.34	0.08	Age group B	0.28	0.08	0.07	0.12	0.47	0.494
	Age group A	0.34	0.08	Age group C	0.17	0.12	0.18	0.14	2.48	0.115
15	Age group A	0.47	0.08	Age group B	0.21	0.08	0.26	0.12	6.84	0.009
	Age group A	0.47	0.08	Age group C	0.34	0.12	0.13	0.14	1.00	0.318
16	Age group A	1.13	0.08	Age group B	1.04	0.09	0.09	0.12	0.03	0.859
	Age group A	1.13	0.08	Age group C	1.15	0.12	-0.02	0.15	0.05	0.828
17	Age group A	1.65	0.09	Age group B	1.62	0.09	0.03	0.13	0.13	0.718
	Age group A	1.65	0.09	Age group C	1.94	0.13	-0.28	0.16	2.25	0.134
18	Age group A	0.70	0.08	Age group B	0.64	0.08	0.05	0.12	0.00	0.997
	Age group A	0.70	0.08	Age group C	0.80	0.12	-0.10	0.14	0.71	0.399
19	Age group A	0.13	0.08	Age group B	-0.01	0.08	0.14	0.12	1.41	0.235
	Age group A	0.13	0.08	Age group C	-0.07	0.12	0.20	0.14	1.05	0.306
20	Age group A	1.04	0.08	Age group B	0.94	0.09	0.10	0.12	0.32	0.571
	Age group A	1.04	0.08	Age group C	1.03	0.12	0.01	0.14	0.00	0.994
21	Age group A	0.58	0.08	Age group B	0.57	0.08	0.01	0.12	0.02	0.896
	Age group A	0.58	0.08	Age group C	0.44	0.12	0.14	0.14	0.57	0.449
23	Age group A	0.73	0.08	Age group B	0.71	0.08	0.02	0.12	0.15	0.696
	Age group A	0.73	0.08	Age group C	0.71	0.12	0.02	0.14	0.04	0.852
24	Age group A	0.70	0.08	Age group B	0.63	0.08	0.07	0.12	3.03	0.082
	Age group A	0.70	0.08	Age group C	0.61	0.12	0.09	0.14	0.45	0.503
25	Age group A	0.10	0.08	Age group B	0.05	0.08	0.04	0.12	0.62	0.431
	Age group A	0.10	0.08	Age group C	-0.02	0.12	0.12	0.14	1.12	0.289
27	Age group A	1.63	0.09	Age group B	1.66	0.09	-0.03	0.13	0.16	0.685
	Age group A	1.63	0.09	Age group C	1.21	0.12	0.41	0.15	3.11	0.078

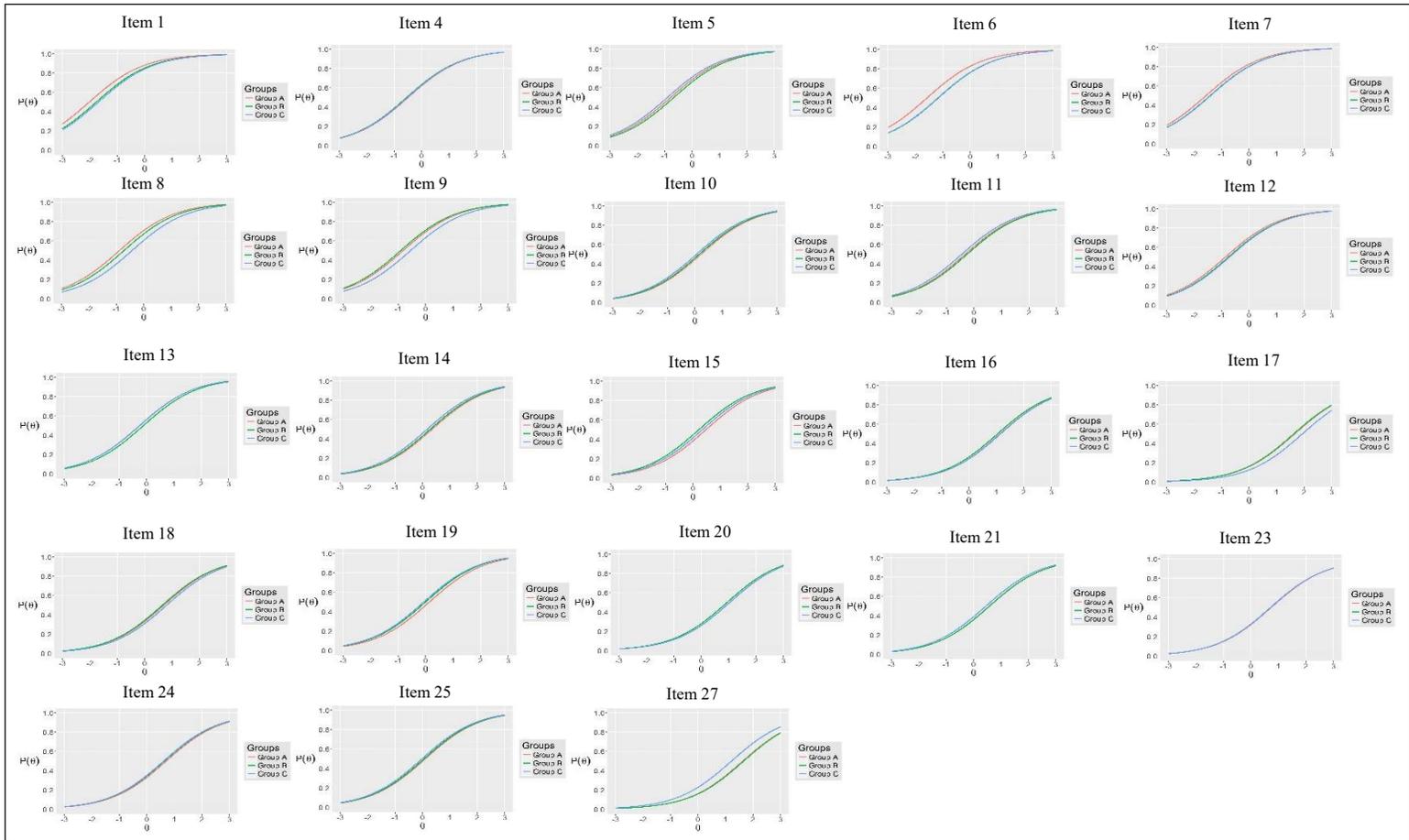


Figure 50 Item Characteristic Curves for Age DIF Subgroups by Immigration Policy in Form B

Table XLVIII DIF Contrasts for Age Subgroups by Similar Sample Sizes in Form B

Item	Reference group	Difficulty	SE	Focal group	Difficulty	SE	DIF contrast	Joint SE	$\chi^2$	<i>p</i>
1	Age group D	-2.100	0.160	Age group E	-1.940	0.140	-0.160	0.210	0.043	0.835
	Age group D	-2.100	0.160	Age group F	-1.660	0.130	-0.440	0.200	3.933	0.047
	Age group D	-2.100	0.160	Age group G	-1.750	0.130	-0.350	0.210	2.455	0.117
4	Age group D	-0.470	0.110	Age group E	-0.660	0.110	0.190	0.150	1.035	0.309
	Age group D	-0.470	0.110	Age group F	-0.490	0.110	0.030	0.150	0.002	0.962
	Age group D	-0.470	0.110	Age group G	-0.500	0.110	0.030	0.150	0.034	0.854
5	Age group D	-0.920	0.120	Age group E	-0.630	0.110	-0.290	0.160	2.575	0.109
	Age group D	-0.920	0.120	Age group F	-0.690	0.110	-0.240	0.160	2.567	0.109
	Age group D	-0.920	0.120	Age group G	-0.840	0.110	-0.080	0.160	0.064	0.801
6	Age group D	-1.790	0.140	Age group E	-1.240	0.120	-0.560	0.190	5.062	0.025
	Age group D	-1.790	0.140	Age group F	-1.190	0.120	-0.610	0.180	10.511	0.001
	Age group D	-1.790	0.140	Age group G	-1.200	0.120	-0.590	0.190	7.751	0.005
7	Age group D	-1.610	0.140	Age group E	-1.480	0.130	-0.130	0.180	0.014	0.906
	Age group D	-1.610	0.140	Age group F	-1.460	0.120	-0.150	0.180	0.386	0.534
	Age group D	-1.610	0.140	Age group G	-1.410	0.120	-0.210	0.180	0.526	0.468
8	Age group D	-0.920	0.120	Age group E	-1.000	0.110	0.080	0.160	0.236	0.627
	Age group D	-0.920	0.120	Age group F	-0.510	0.110	-0.410	0.160	4.477	0.034
	Age group D	-0.920	0.120	Age group G	-0.530	0.110	-0.390	0.160	4.429	0.035
9	Age group D	-0.890	0.120	Age group E	-0.810	0.110	-0.090	0.160	0.074	0.786
	Age group D	-0.890	0.120	Age group F	-0.880	0.110	-0.010	0.160	0.077	0.782
	Age group D	-0.890	0.120	Age group G	-0.660	0.110	-0.240	0.160	1.439	0.230
10	Age group D	0.210	0.100	Age group E	0.210	0.100	0.000	0.150	0.023	0.879
	Age group D	0.210	0.100	Age group F	0.150	0.100	0.060	0.150	0.231	0.631
	Age group D	0.210	0.100	Age group G	0.040	0.110	0.170	0.150	0.176	0.675
11	Age group D	-0.360	0.110	Age group E	-0.300	0.110	-0.060	0.150	0.152	0.696
	Age group D	-0.360	0.110	Age group F	-0.140	0.100	-0.220	0.150	0.399	0.528
	Age group D	-0.360	0.110	Age group G	-0.520	0.110	0.160	0.150	1.324	0.250
12	Age group D	-0.790	0.110	Age group E	-0.790	0.110	0.000	0.160	0.002	0.962
	Age group D	-0.790	0.110	Age group F	-0.880	0.110	0.080	0.160	0.511	0.475
	Age group D	-0.790	0.110	Age group G	-0.720	0.110	-0.070	0.160	0.011	0.916
13	Age group D	-0.430	0.110	Age group E	0.030	0.100	-0.450	0.150	5.423	0.020
	Age group D	-0.430	0.110	Age group F	-0.200	0.100	-0.230	0.150	0.593	0.441
	Age group D	-0.430	0.110	Age group G	-0.200	0.110	-0.230	0.150	1.112	0.292
14	Age group D	0.360	0.100	Age group E	0.320	0.100	0.040	0.150	0.179	0.673
	Age group D	0.360	0.100	Age group F	0.170	0.100	0.190	0.150	2.203	0.138
	Age group D	0.360	0.100	Age group G	0.280	0.100	0.080	0.150	0.614	0.433
15	Age group D	0.410	0.100	Age group E	0.460	0.100	-0.050	0.150	0.205	0.651
	Age group D	0.410	0.100	Age group F	0.130	0.100	0.280	0.150	5.041	0.025
	Age group D	0.410	0.100	Age group G	0.340	0.100	0.080	0.150	0.146	0.703

Table XLII continued

	Age group D	1.300	0.110	Age group E	0.890	0.110	0.410	0.150	4.319	0.038
16	Age group D	1.300	0.110	Age group F	0.940	0.110	0.360	0.150	3.797	0.051
	Age group D	1.300	0.110	Age group G	1.260	0.110	0.040	0.150	0.022	0.883
	Age group D	1.660	0.110	Age group E	1.630	0.110	0.020	0.160	0.002	0.965
17	Age group D	1.660	0.110	Age group F	1.640	0.110	0.020	0.160	0.144	0.704
	Age group D	1.660	0.110	Age group G	1.860	0.120	-0.200	0.160	1.297	0.255
	Age group D	0.790	0.100	Age group E	0.540	0.100	0.250	0.150	0.563	0.453
18	Age group D	0.790	0.100	Age group F	0.700	0.100	0.100	0.150	0.020	0.887
	Age group D	0.790	0.100	Age group G	0.730	0.110	0.060	0.150	0.003	0.954
	Age group D	0.140	0.100	Age group E	0.130	0.100	0.010	0.150	0.259	0.611
19	Age group D	0.140	0.100	Age group F	0.040	0.100	0.110	0.150	0.538	0.463
	Age group D	0.140	0.100	Age group G	-0.170	0.110	0.310	0.150	4.129	0.042
	Age group D	1.060	0.100	Age group E	1.000	0.110	0.070	0.150	0.317	0.573
20	Age group D	1.060	0.100	Age group F	0.860	0.100	0.210	0.150	1.673	0.196
	Age group D	1.060	0.100	Age group G	1.070	0.110	0.000	0.150	0.026	0.872
	Age group D	0.630	0.100	Age group E	0.550	0.100	0.080	0.150	0.199	0.656
21	Age group D	0.630	0.100	Age group F	0.550	0.100	0.080	0.150	0.379	0.538
	Age group D	0.630	0.100	Age group G	0.480	0.100	0.150	0.150	1.009	0.315
	Age group D	0.710	0.100	Age group E	0.710	0.110	0.000	0.150	0.430	0.512
23	Age group D	0.710	0.100	Age group F	0.740	0.100	-0.030	0.150	1.133	0.287
	Age group D	0.710	0.100	Age group G	0.690	0.110	0.020	0.150	0.000	0.995
	Age group D	0.760	0.100	Age group E	0.560	0.100	0.200	0.150	3.038	0.081
24	Age group D	0.760	0.100	Age group F	0.650	0.100	0.110	0.150	2.778	0.096
	Age group D	0.760	0.100	Age group G	0.620	0.100	0.130	0.150	1.427	0.232
	Age group D	0.120	0.100	Age group E	0.100	0.100	0.020	0.150	0.000	0.988
25	Age group D	0.120	0.100	Age group F	0.090	0.100	0.020	0.150	0.232	0.630
	Age group D	0.120	0.100	Age group G	-0.090	0.110	0.210	0.150	3.834	0.050
	Age group D	1.710	0.110	Age group E	1.620	0.110	0.090	0.160	0.358	0.550
27	Age group D	1.710	0.110	Age group F	1.510	0.110	0.200	0.160	0.154	0.695
	Age group D	1.710	0.110	Age group G	1.370	0.110	0.340	0.160	0.614	0.434

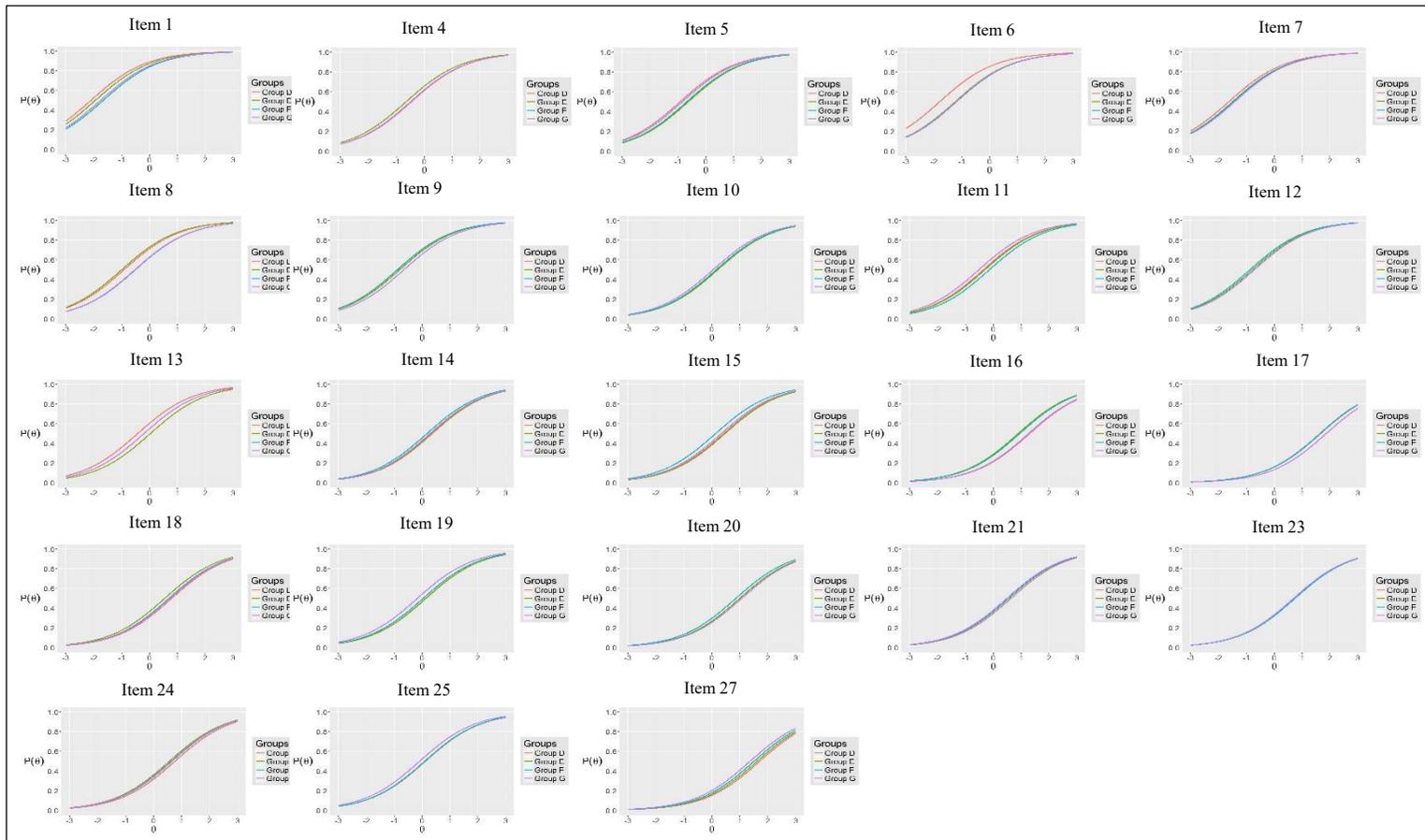


Figure 51 Item Characteristic Curves for Age DIF Subgroups by Sample Sizes in Form B

Table XLIX DIF Contrasts for Geographical Location Subgroups in Form B

Item	Reference group	Difficulty	SE	Focal group	Difficulty	SE	DIF contrast	Joint SE	$\chi^2$	<i>p</i>
1	North Africa	-2.88	0.39	West Africa	-1.53	0.18	-1.35	0.43	7.63	0.006
4	North Africa	-0.27	0.16	West Africa	-1.23	0.16	0.96	0.23	12.42	0.000
5	North Africa	-0.99	0.20	West Africa	-0.05	0.12	-0.95	0.23	10.89	0.001
6	North Africa	-2.11	0.28	West Africa	-1.53	0.18	-0.58	0.33	3.42	0.065
7	North Africa	-2.85	0.39	West Africa	-1.70	0.19	-1.15	0.43	6.26	0.012
8	North Africa	-0.72	0.18	West Africa	0.23	0.11	-0.95	0.21	14.98	0.000
9	North Africa	-0.48	0.17	West Africa	-0.72	0.13	0.24	0.22	2.01	0.156
10	North Africa	0.50	0.15	West Africa	1.35	0.10	-0.85	0.18	19.39	0.000
11	North Africa	0.17	0.15	West Africa	-0.63	0.13	0.80	0.20	19.75	0.000
12	North Africa	-1.54	0.23	West Africa	-1.71	0.19	0.16	0.30	0.89	0.346
13	North Africa	-1.14	0.20	West Africa	-0.84	0.14	-0.30	0.25	2.71	0.099
14	North Africa	-0.13	0.16	West Africa	-0.09	0.11	-0.04	0.20	0.37	0.545
15	North Africa	-0.08	0.16	West Africa	0.00	0.11	-0.08	0.19	0.64	0.424
16	North Africa	1.74	0.14	West Africa	1.50	0.10	0.24	0.17	1.80	0.180
17	North Africa	2.30	0.15	West Africa	1.96	0.10	0.34	0.18	2.59	0.107
18	North Africa	1.15	0.14	West Africa	1.45	0.10	-0.30	0.17	2.10	0.148
19	North Africa	0.12	0.15	West Africa	-0.27	0.12	0.39	0.19	3.39	0.066
20	North Africa	1.35	0.14	West Africa	1.37	0.10	-0.03	0.17	0.00	0.997
21	North Africa	0.46	0.15	West Africa	1.00	0.10	-0.55	0.18	7.48	0.006
23	North Africa	1.37	0.14	West Africa	0.82	0.10	0.55	0.17	7.28	0.007
24	North Africa	0.30	0.15	West Africa	0.17	0.11	0.13	0.18	0.00	0.996
25	North Africa	0.18	0.15	West Africa	-0.96	0.14	1.14	0.21	31.44	0.000
27	North Africa	1.83	0.14	West Africa	1.49	0.10	0.34	0.17	1.85	0.174

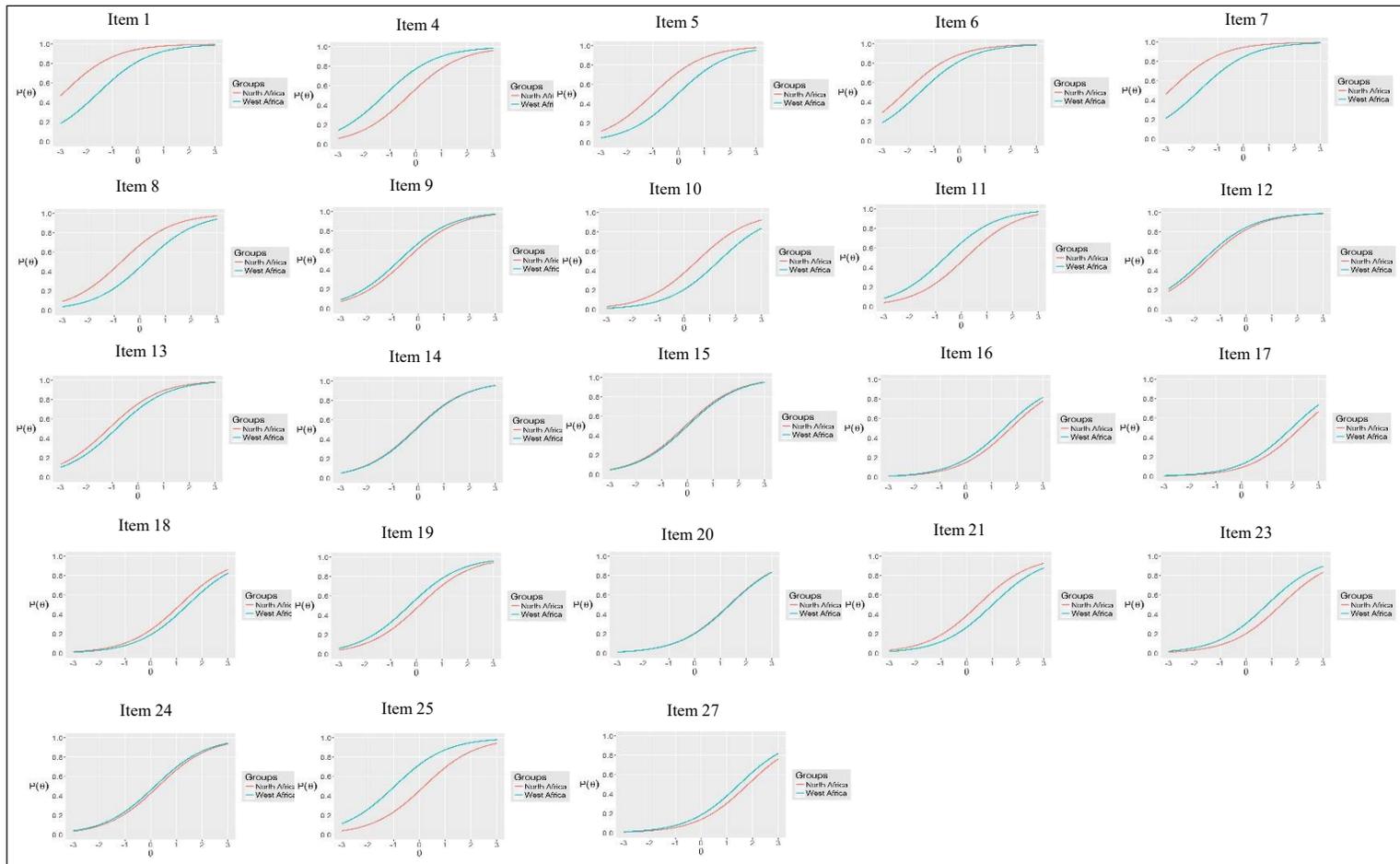


Figure 52 Item Characteristic Curves for Geographical Location DIF Subgroups in Form B

### Appendix C: Nominal Response Model Parameters and Graphics

Table L Intercept and Slope NRM Parameters for Males and Females in Form A

Item	Reference group (males)								Focal group (females)								Likelihood ratio test		
	$\zeta_1$	$\zeta_2$	$\zeta_3$	$\zeta_4$	$\lambda_1$	$\lambda_2$	$\lambda_3$	$\lambda_4$	$\zeta_1$	$\zeta_2$	$\zeta_3$	$\zeta_4$	$\lambda_1$	$\lambda_2$	$\lambda_3$	$\lambda_4$	LogLik difference	df	p
1	-0.245	-1.384	3.412	-1.783	-0.164	-0.443	0.768	-0.161	-0.673	-0.957	3.394	-1.764	-0.377	-0.488	0.636	0.230	2.886	6	0.823
2	-1.430	-1.516	4.045	-1.099	-0.694	-0.244	2.437	-1.499	-1.605	-1.642	3.779	-0.532	-0.893	-0.465	2.281	-0.923	3.004	6	0.808
3	-0.526	-0.670	-1.555	2.750	0.080	-0.973	-0.532	1.425	-0.576	-0.121	-2.082	2.778	-0.043	-0.742	-0.877	1.662	7.090	6	0.313
4	-0.788	2.366	-0.776	-0.802	-0.061	1.102	-0.691	-0.351	-0.940	2.322	-0.274	-1.109	-0.279	0.875	-0.311	-0.284	8.930	6	0.178
5	-0.863	1.100	-0.476	0.239	-0.139	0.351	-0.489	0.277	-0.752	1.109	-0.510	0.153	-0.263	0.420	-0.614	0.457	4.606	6	0.595
6	-0.757	2.295	-1.226	-0.312	-0.482	1.770	-0.792	-0.496	-1.009	2.310	-1.324	0.022	-0.644	1.828	-0.866	-0.319	2.858	6	0.826
7	-1.281	-0.731	-1.098	3.109	-0.344	-0.399	-0.940	1.682	-1.096	-0.983	-0.918	2.997	-0.236	-0.684	-0.895	1.816	3.423	6	0.754
8	1.612	-0.054	0.559	-2.117	1.339	0.008	0.323	-1.670	1.934	-0.165	0.828	-2.598	1.541	-0.093	0.507	-1.954	6.478	6	0.372
9	-0.383	-1.358	-0.539	2.281	0.212	-1.401	-1.065	2.253	-0.516	-0.954	-0.439	1.908	0.067	-0.786	-0.978	1.697	11.511	6	0.074
10	-0.329	0.517	2.283	-2.470	-0.273	0.176	1.599	-1.503	-0.454	0.642	2.661	-2.849	-0.303	0.373	1.518	-1.588	13.937	6	0.030
11	2.053	0.168	-0.809	-1.412	1.906	-0.160	-1.089	-0.657	2.346	0.387	-0.994	-1.740	2.664	0.046	-1.363	-1.348	9.771	6	0.135
12	-0.286	-0.937	1.188	0.035	-0.281	-0.411	0.661	0.030	-0.214	-1.255	1.405	0.064	-0.454	-0.399	0.789	0.065	8.614	6	0.196
13	0.116	-0.700	1.125	-0.541	-0.421	-0.254	0.790	-0.115	0.002	-0.644	1.357	-0.714	-0.512	-0.469	0.997	-0.016	9.764	6	0.135
14	-0.179	-0.833	0.706	0.306	0.211	-0.920	0.925	-0.216	-0.224	-0.775	0.603	0.397	0.171	-0.848	0.968	-0.291	3.379	6	0.760
15	0.865	0.875	0.363	-2.103	0.141	1.261	0.024	-1.426	0.688	0.993	0.329	-2.009	-0.002	1.155	0.137	-1.290	5.730	6	0.454
16	-0.564	-0.048	0.176	0.437	-0.303	-0.298	-0.183	0.784	-0.698	-0.134	0.070	0.762	-0.363	-0.155	-0.320	0.838	14.936	6	0.021
17	-0.936	1.551	-0.880	0.266	-0.716	1.231	-0.437	-0.078	-1.848	1.853	-0.386	0.380	-1.591	1.783	-0.100	-0.092	17.754	6	0.007
18	0.292	1.179	-1.017	-0.454	-0.031	1.091	-1.003	-0.058	-0.039	1.186	-0.740	-0.407	-0.111	0.957	-0.632	-0.215	10.291	6	0.113
19	-0.127	0.095	0.924	-0.892	0.072	-0.478	0.927	-0.521	-0.085	-0.440	1.236	-0.711	-0.093	-0.816	1.178	-0.269	24.090	6	<b>0.001</b>
20	-0.249	1.147	-0.513	-0.385	-0.694	0.631	0.082	-0.019	-0.041	1.264	-0.548	-0.675	-0.672	0.832	0.157	-0.317	8.661	6	0.194
21	-0.953	-0.414	1.297	0.071	-0.494	-0.199	0.788	-0.095	-1.020	-0.234	1.287	-0.033	-0.053	-0.225	0.557	-0.279	12.751	6	0.047
22	0.318	-0.554	-0.459	0.694	-0.262	-0.206	-0.405	0.873	0.263	-0.264	-0.528	0.529	-0.379	0.070	-0.660	0.969	10.823	6	0.094
23	0.405	-0.350	-0.050	-0.005	1.189	-0.568	-0.440	-0.181	0.126	-0.172	-0.050	0.096	0.884	-0.341	-0.373	-0.171	22.194	6	<b>0.001</b>
24	-0.733	0.316	-0.858	1.275	-0.395	0.284	-0.678	0.788	-0.606	0.218	-0.717	1.104	-0.256	0.113	-0.557	0.700	4.455	6	0.615
25	0.039	-0.639	-0.154	0.754	-0.046	-0.390	-0.301	0.738	0.074	-0.711	-0.270	0.908	0.117	-0.461	-0.338	0.682	4.968	6	0.548
26	-0.894	0.708	1.022	-0.837	-0.949	1.175	0.370	-0.596	-1.268	0.889	0.875	-0.497	-1.116	1.134	0.418	-0.436	11.175	6	0.083
27	0.385	0.425	-0.805	-0.006	-0.250	0.342	0.057	-0.149	0.422	0.372	-0.632	-0.163	-0.383	0.369	0.200	-0.186	6.081	6	0.414
28	0.870	0.170	0.517	-1.557	1.531	0.042	0.100	-1.673	0.494	0.116	0.496	-1.106	1.396	0.037	-0.209	-1.224	16.842	6	0.010
29	0.178	0.411	1.484	-2.073	0.222	0.008	0.702	-0.932	0.062	0.376	1.254	-1.692	0.137	0.001	0.667	-0.804	4.886	6	0.559
30	-0.451	0.003	-0.150	0.598	0.519	-0.066	-0.385	-0.068	-0.559	0.103	-0.135	0.591	0.569	-0.025	-0.424	-0.120	1.866	6	0.932

Table LI Intercept and Slope NRM Parameters for Males and Females in Form B

Item	Reference group (males)								Focal group (females)								Likelihood ratio test		
	$\zeta_1$	$\zeta_2$	$\zeta_3$	$\zeta_4$	$\lambda_1$	$\lambda_2$	$\lambda_3$	$\lambda_4$	$\zeta_1$	$\zeta_2$	$\zeta_3$	$\zeta_4$	$\lambda_1$	$\lambda_2$	$\lambda_3$	$\lambda_4$	LogLik difference	df	p
1	-0.393	2.997	-0.591	-2.013	-0.169	1.330	-0.222	-0.938	0.283	3.563	-0.691	-3.154	0.373	1.777	-0.572	-1.577	9.639	6	0.141
2	-0.263	-0.470	-0.413	1.147	-0.146	-0.121	-0.182	0.449	-0.406	-0.465	-0.297	1.168	-0.207	-0.057	-0.214	0.477	2.774	6	0.837
3	-0.480	-0.455	-0.592	1.527	0.087	-0.223	-0.258	0.394	-0.480	-0.455	-0.592	1.527	0.087	-0.223	-0.258	0.394	7.927	6	0.243
4	-0.881	-1.246	2.081	0.046	-0.874	-0.894	1.538	0.230	-0.603	-0.826	1.815	-0.386	-0.704	-0.589	1.170	0.123	14.905	6	0.021
5	-0.616	2.185	-1.695	0.127	-0.035	1.248	-1.051	-0.162	-0.326	2.237	-2.127	0.217	0.261	1.312	-1.402	-0.171	3.710	6	0.716
6	-1.884	3.283	-0.258	-1.141	-1.220	2.148	-0.358	-0.571	-1.736	3.138	-0.448	-0.953	-1.024	2.051	-0.666	-0.361	2.407	6	0.879
7	-0.355	-1.380	-1.218	2.953	-0.295	-0.793	-0.671	1.759	-0.673	-1.336	-0.844	2.852	-0.571	-0.669	-0.232	1.472	5.150	6	0.525
8	-0.824	1.784	-0.047	-0.913	0.204	0.595	-0.393	-0.406	-0.979	1.791	-0.007	-0.806	0.105	0.633	-0.378	-0.360	1.007	6	0.985
9	0.258	-1.364	2.009	-0.903	0.240	-1.168	1.166	-0.237	-0.007	-1.118	2.306	-1.180	0.023	-1.034	1.204	-0.193	16.862	6	0.010
10	-0.480	-0.809	1.110	0.179	-0.208	-0.201	0.557	-0.149	-0.420	-0.914	1.309	0.026	-0.197	-0.122	0.648	-0.329	11.218	6	0.082
11	1.990	-1.645	0.370	-0.715	1.479	-0.955	-0.227	-0.296	1.925	-1.533	0.224	-0.616	1.351	-0.803	-0.497	-0.051	3.976	6	0.680
12	-1.879	0.135	2.485	-0.742	-0.745	-0.263	1.664	-0.656	1.925	-1.533	0.224	-0.616	1.351	-0.803	-0.497	-0.051	15.295	6	0.018
13	-0.215	-1.152	1.942	-0.575	-0.107	-1.028	1.753	-0.619	-0.744	-0.571	1.679	-0.364	-0.444	-0.655	1.604	-0.505	17.129	6	<b>0.009</b>
14	1.328	-0.277	-1.253	0.201	1.450	-0.261	-1.214	0.025	1.477	-0.183	-1.146	-0.149	1.615	-0.116	-1.190	-0.308	10.219	6	0.116
15	-0.071	-0.122	-1.133	1.326	0.090	-0.457	-1.349	1.717	0.022	-0.066	-1.455	1.498	-0.014	-0.218	-1.470	1.702	10.438	6	0.107
16	-0.867	0.749	0.682	-0.564	-0.120	0.614	0.001	-0.495	-0.650	0.715	0.533	-0.597	-0.239	0.784	-0.008	-0.538	11.176	6	0.083
17	-0.924	0.245	0.256	0.424	-0.728	0.679	0.065	-0.017	-1.059	0.471	0.285	0.304	-0.663	0.877	-0.032	-0.182	18.847	6	<b>0.004</b>
18	0.844	-0.218	1.316	-1.943	0.333	-0.185	1.010	-1.158	0.644	-0.085	0.980	-1.539	0.098	-0.060	0.835	-0.873	14.080	6	0.029
19	-0.370	-0.252	-0.930	1.552	-0.422	-0.131	-0.822	1.376	-0.142	-0.272	-1.005	1.419	-0.292	-0.301	-0.802	1.396	6.400	6	0.380
20	-0.684	0.799	-0.544	0.429	-0.164	0.651	-0.155	-0.332	-0.695	0.674	-0.556	0.577	-0.197	0.874	-0.299	-0.378	12.565	6	0.050
21	-0.368	0.993	-0.055	-0.570	0.202	0.920	-0.415	-0.707	-0.457	0.967	-0.027	-0.483	0.268	0.831	-0.492	-0.606	3.265	6	0.775
22	-0.157	0.391	0.120	-0.354	0.556	-0.197	-0.030	-0.329	-0.022	0.643	0.027	-0.648	0.558	-0.140	0.043	-0.462	18.703	6	0.005
23	0.123	-0.335	-0.655	0.867	0.090	-0.225	-0.679	0.814	0.112	-0.348	-0.663	0.899	0.083	-0.288	-0.738	0.943	1.642	6	0.950
24	-0.483	1.029	0.141	-0.687	-0.832	1.457	-0.423	-0.202	-0.466	0.997	0.195	-0.726	-0.662	1.373	-0.520	-0.192	4.971	6	0.548
25	-0.190	1.512	-0.500	-0.822	-0.405	1.454	-0.573	-0.475	-0.215	1.421	-0.461	-0.745	-0.237	1.336	-0.632	-0.467	5.012	6	0.542
26	1.088	0.921	-0.654	-1.355	0.546	0.888	-0.933	-0.501	1.088	0.921	-0.654	-1.355	0.546	0.888	-0.933	-0.501	11.193	6	0.083
27	-0.641	0.668	1.003	-1.030	-0.329	0.849	0.415	-0.935	-0.614	0.582	0.865	-0.833	-0.281	0.794	0.281	-0.794	3.894	6	0.691
28	-0.070	-0.450	0.448	0.072	-0.238	0.023	-0.246	0.461	-0.006	-0.711	0.612	0.104	-0.166	-0.173	-0.061	0.399	15.262	6	0.018
29	0.287	0.320	0.304	-0.911	0.545	-0.117	-0.033	-0.395	0.240	0.518	0.363	-1.122	0.493	-0.023	-0.134	-0.336	12.688	6	0.048
30	-0.471	-0.049	-0.022	0.543	0.262	-0.014	-0.310	0.063	-0.496	0.014	0.005	0.477	0.205	0.074	-0.201	-0.079	6.554	6	0.364

Table LII Intercept and Slope NRM Parameters with Standard Errors for the Parsimonious Model of Form A

Item	$\zeta_1$	$SE_1$	$\zeta_2$	$SE_2$	$\zeta_3$	$SE_3$	$\zeta_4$	$SE_4$	$\lambda_1$	$SE_5$	$\lambda_2$	$SE_6$	$\lambda_3$	$SE_7$	$\lambda_4$	$SE_8$
1	-0.407	0.192	-1.176	0.251	3.357	0.132	-1.774	0.295	-0.234	0.187	-0.494	0.235	0.704	0.131	0.024	0.298
2	-1.492	0.331	-1.603	0.374	3.822	0.192	-0.726	0.248	-0.811	0.277	-0.402	0.326	2.386	0.186	-1.173	0.207
3	-0.560	0.159	-0.388	0.152	-1.731	0.245	2.679	0.106	0.014	0.158	-0.889	0.139	-0.662	0.223	1.537	0.112
4	-0.861	0.132	2.284	0.072	-0.511	0.120	-0.913	0.136	-0.176	0.132	0.979	0.078	-0.496	0.116	-0.307	0.134
5	-0.801	0.076	1.082	0.044	-0.462	0.071	0.182	0.054	-0.207	0.081	0.388	0.046	-0.547	0.074	0.366	0.055
6	-0.829	0.143	2.214	0.085	-1.232	0.165	-0.153	0.115	-0.541	0.137	1.799	0.103	-0.827	0.152	-0.431	0.111
7	-1.187	0.223	-0.816	0.192	-0.979	0.197	2.983	0.119	-0.299	0.204	-0.533	0.171	-0.932	0.170	1.764	0.122
8	1.657	0.079	-0.095	0.099	0.634	0.086	-2.196	0.204	1.408	0.083	-0.031	0.097	0.387	0.084	-1.765	0.176
9	-0.467	0.117	-1.077	0.145	-0.462	0.117	2.006	0.078	0.129	0.132	-1.058	0.142	-1.036	0.118	1.965	0.107
10	-0.369	0.142	0.529	0.119	2.337	0.107	-2.497	0.280	-0.277	0.128	0.249	0.108	1.513	0.102	-1.484	0.226
11	2.057	0.089	0.263	0.104	-0.818	0.138	-1.502	0.179	2.231	0.119	-0.051	0.109	-1.188	0.132	-0.993	0.171
12	-0.238	0.069	-1.034	0.093	1.234	0.047	0.038	0.061	-0.376	0.073	-0.380	0.097	0.711	0.054	0.045	0.065
13	0.096	0.062	-0.653	0.080	1.167	0.047	-0.609	0.075	-0.451	0.068	-0.367	0.086	0.875	0.057	-0.057	0.084
14	-0.209	0.059	-0.765	0.081	0.617	0.049	0.357	0.051	0.195	0.066	-0.881	0.086	0.945	0.057	-0.258	0.057
15	0.786	0.069	0.865	0.069	0.339	0.074	-1.990	0.170	0.079	0.067	1.203	0.073	0.078	0.073	-1.360	0.153
16	-0.602	0.065	-0.079	0.054	0.148	0.050	0.533	0.045	-0.323	0.072	-0.229	0.059	-0.242	0.055	0.794	0.052
17	-1.172	0.129	1.561	0.063	-0.685	0.104	0.295	0.076	-1.025	0.124	1.436	0.079	-0.311	0.108	-0.100	0.080
18	0.155	0.061	1.126	0.050	-0.848	0.093	-0.434	0.074	-0.054	0.066	1.024	0.061	-0.816	0.095	-0.153	0.080
19	-0.114	0.062	-0.049	0.064	0.977	0.047	-0.814	0.083	-0.020	0.069	-0.570	0.070	1.008	0.060	-0.418	0.090
20	-0.136	0.067	1.150	0.045	-0.541	0.070	-0.473	0.069	-0.708	0.072	0.716	0.052	0.115	0.077	-0.123	0.076
21	-0.953	0.088	-0.330	0.069	1.255	0.046	0.028	0.061	-0.279	0.093	-0.216	0.074	0.675	0.053	-0.179	0.065
22	0.306	0.050	-0.418	0.062	-0.465	0.066	0.577	0.047	-0.316	0.056	-0.067	0.069	-0.531	0.073	0.914	0.056
23	0.246	0.050	-0.248	0.057	-0.040	0.053	0.042	0.051	1.031	0.059	-0.446	0.065	-0.406	0.060	-0.179	0.057
24	-0.659	0.079	0.262	0.056	-0.765	0.086	1.162	0.048	-0.325	0.083	0.199	0.060	-0.621	0.088	0.747	0.054
25	0.048	0.052	-0.648	0.070	-0.187	0.058	0.786	0.044	0.030	0.057	-0.421	0.075	-0.314	0.063	0.705	0.050
26	-0.990	0.103	0.724	0.057	0.932	0.054	-0.666	0.086	-1.003	0.102	1.149	0.066	0.394	0.058	-0.539	0.091
27	0.418	0.044	0.382	0.045	-0.735	0.064	-0.065	0.051	-0.320	0.047	0.353	0.045	0.128	0.065	-0.162	0.053
28	0.637	0.058	0.131	0.064	0.496	0.059	-1.264	0.115	1.453	0.074	0.040	0.071	-0.067	0.064	-1.426	0.114
29	0.114	0.073	0.390	0.070	1.347	0.062	-1.851	0.157	0.185	0.073	0.007	0.069	0.692	0.063	-0.884	0.146
30	-0.524	0.061	0.048	0.048	-0.124	0.052	0.600	0.041	0.542	0.056	-0.046	0.049	-0.404	0.055	-0.092	0.042

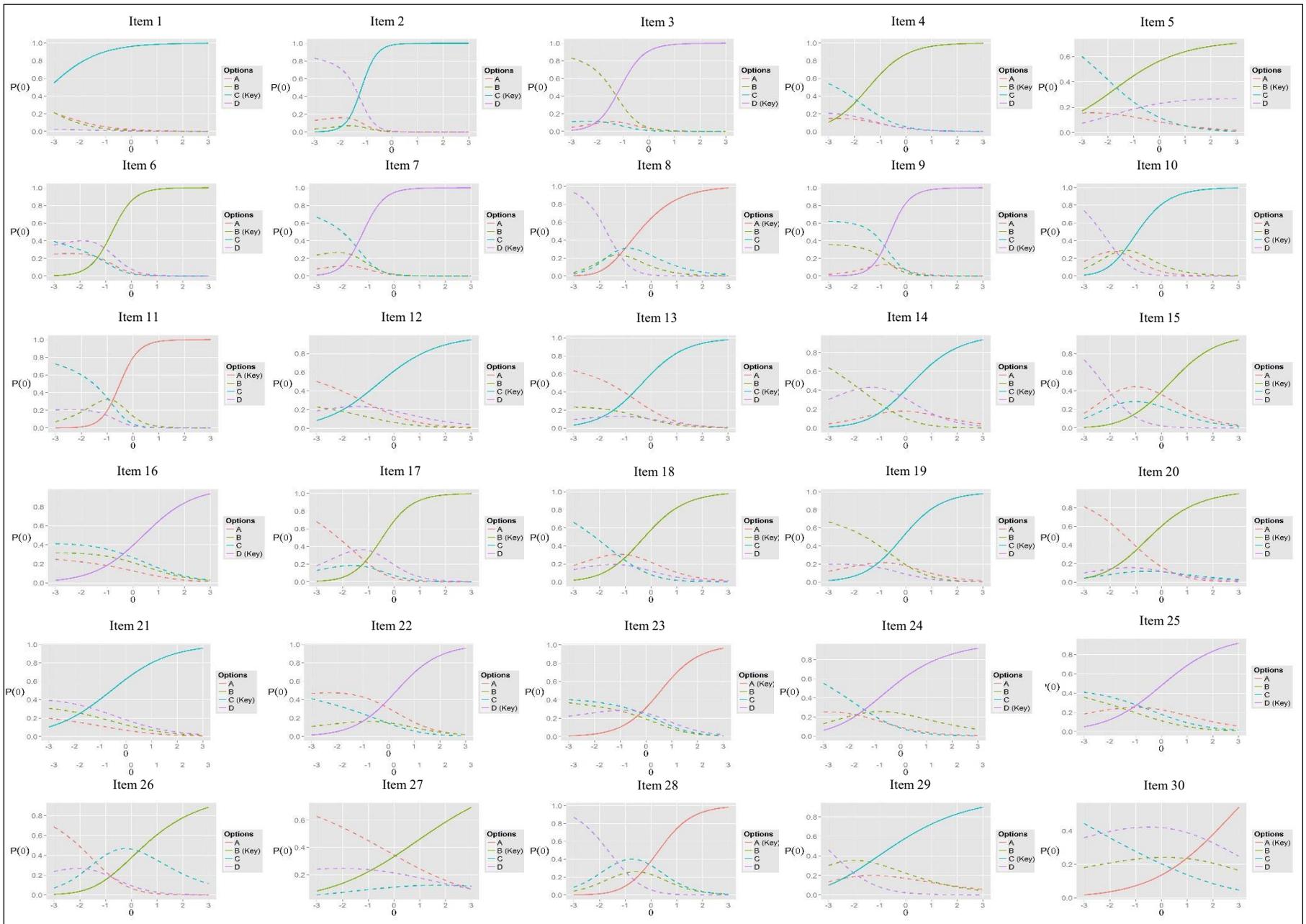


Figure 53 NRM Item Characteristic Response Curves for Form A

Table LIII Intercept and Slope NRM Parameters with Standard Errors for the Parsimonious Model of Form B

Item	$\zeta_1$	$SE_1$	$\zeta_2$	$SE_2$	$\zeta_3$	$SE_3$	$\zeta_4$	$SE_4$	$\lambda_1$	$SE_5$	$\lambda_2$	$SE_6$	$\lambda_3$	$SE_7$	$\lambda_4$	$SE_8$
1	-0.156	0.151	3.092	0.118	-0.629	0.17	-2.307	0.292	0.034	0.143	1.482	0.112	-0.406	0.155	-1.11	0.247
2	-0.314	0.052	-0.465	0.054	-0.347	0.053	1.126	0.035	-0.165	0.057	-0.091	0.059	-0.205	0.058	0.461	0.038
3	-0.485	0.061	-0.441	0.062	-0.577	0.065	1.503	0.038	0.087	0.064	-0.223	0.067	-0.259	0.07	0.395	0.041
4	-0.704	0.098	-0.986	0.109	1.865	0.057	-0.175	0.078	-0.793	0.101	-0.732	0.112	1.35	0.07	0.174	0.088
5	-0.494	0.101	2.126	0.069	-1.806	0.168	0.174	0.086	0.100	0.107	1.274	0.076	-1.202	0.156	-0.172	0.088
6	-1.733	0.218	3.085	0.112	-0.329	0.146	-1.023	0.184	-1.111	0.193	2.101	0.121	-0.523	0.136	-0.467	0.173
7	-0.475	0.135	-1.314	0.178	-1.009	0.162	2.798	0.092	-0.422	0.13	-0.729	0.166	-0.457	0.155	1.608	0.099
8	-0.897	0.081	1.748	0.045	-0.006	0.064	-0.845	0.085	0.168	0.088	0.613	0.051	-0.387	0.07	-0.393	0.092
9	0.148	0.077	-1.2	0.127	2.051	0.061	-0.999	0.112	0.165	0.083	-1.122	0.123	1.156	0.069	-0.199	0.12
10	-0.444	0.058	-0.849	0.067	1.156	0.038	0.137	0.048	-0.207	0.064	-0.163	0.074	0.587	0.043	-0.217	0.053
11	1.872	0.063	-1.537	0.142	0.325	0.076	-0.66	0.100	1.415	0.077	-0.878	0.143	-0.356	0.081	-0.18	0.109
12	-1.842	0.196	0.152	0.103	2.39	0.085	-0.699	0.129	-0.755	0.189	-0.263	0.104	1.671	0.096	-0.653	0.127
13	-0.447	0.085	-0.803	0.098	1.706	0.056	-0.455	0.086	-0.26	0.095	-0.834	0.104	1.672	0.078	-0.577	0.093
14	1.287	0.049	-0.234	0.068	-1.136	0.100	0.082	0.062	1.506	0.068	-0.204	0.078	-1.213	0.105	-0.09	0.071
15	-0.034	0.066	-0.083	0.068	-1.174	0.103	1.291	0.051	0.027	0.078	-0.343	0.078	-1.378	0.109	1.694	0.074
16	-0.747	0.061	0.683	0.04	0.613	0.04	-0.55	0.06	-0.193	0.069	0.697	0.044	0.008	0.044	-0.512	0.067
17	-0.941	0.068	0.289	0.042	0.268	0.042	0.384	0.041	-0.693	0.075	0.757	0.045	0.018	0.045	-0.082	0.044
18	0.729	0.054	-0.149	0.065	1.099	0.051	-1.678	0.122	0.213	0.056	-0.122	0.069	0.922	0.056	-1.012	0.122
19	-0.24	0.069	-0.248	0.069	-0.922	0.092	1.41	0.048	-0.361	0.078	-0.216	0.078	-0.813	0.098	1.39	0.065
20	-0.681	0.059	0.696	0.039	-0.532	0.056	0.516	0.041	-0.176	0.067	0.751	0.045	-0.215	0.064	-0.36	0.047
21	-0.423	0.055	0.931	0.039	-0.015	0.051	-0.492	0.062	0.236	0.062	0.88	0.047	-0.456	0.058	-0.66	0.07
22	-0.135	0.044	0.512	0.035	0.072	0.040	-0.449	0.049	0.549	0.042	-0.184	0.037	0.003	0.041	-0.368	0.053
23	0.114	0.046	-0.323	0.054	-0.615	0.065	0.825	0.039	0.089	0.052	-0.253	0.061	-0.708	0.072	0.872	0.046
24	-0.431	0.067	0.931	0.044	0.193	0.053	-0.693	0.07	-0.748	0.078	1.419	0.064	-0.477	0.064	-0.194	0.085
25	-0.183	0.067	1.39	0.047	-0.448	0.075	-0.759	0.084	-0.322	0.076	1.403	0.066	-0.607	0.083	-0.474	0.093
26	1.057	0.049	0.869	0.05	-0.601	0.079	-1.325	0.099	0.548	0.054	0.889	0.057	-0.935	0.085	-0.501	0.108
27	-0.613	0.064	0.578	0.045	0.920	0.042	-0.885	0.078	-0.307	0.071	0.823	0.049	0.354	0.045	-0.87	0.083
28	-0.036	0.041	-0.549	0.049	0.526	0.035	0.06	0.04	-0.211	0.044	-0.049	0.051	-0.164	0.037	0.424	0.039
29	0.234	0.041	0.412	0.039	0.336	0.039	-0.981	0.063	0.52	0.042	-0.075	0.041	-0.084	0.042	-0.361	0.069
30	-0.496	0.048	-0.025	0.04	0.003	0.04	0.518	0.034	0.235	0.045	0.026	0.04	-0.259	0.042	-0.001	0.034

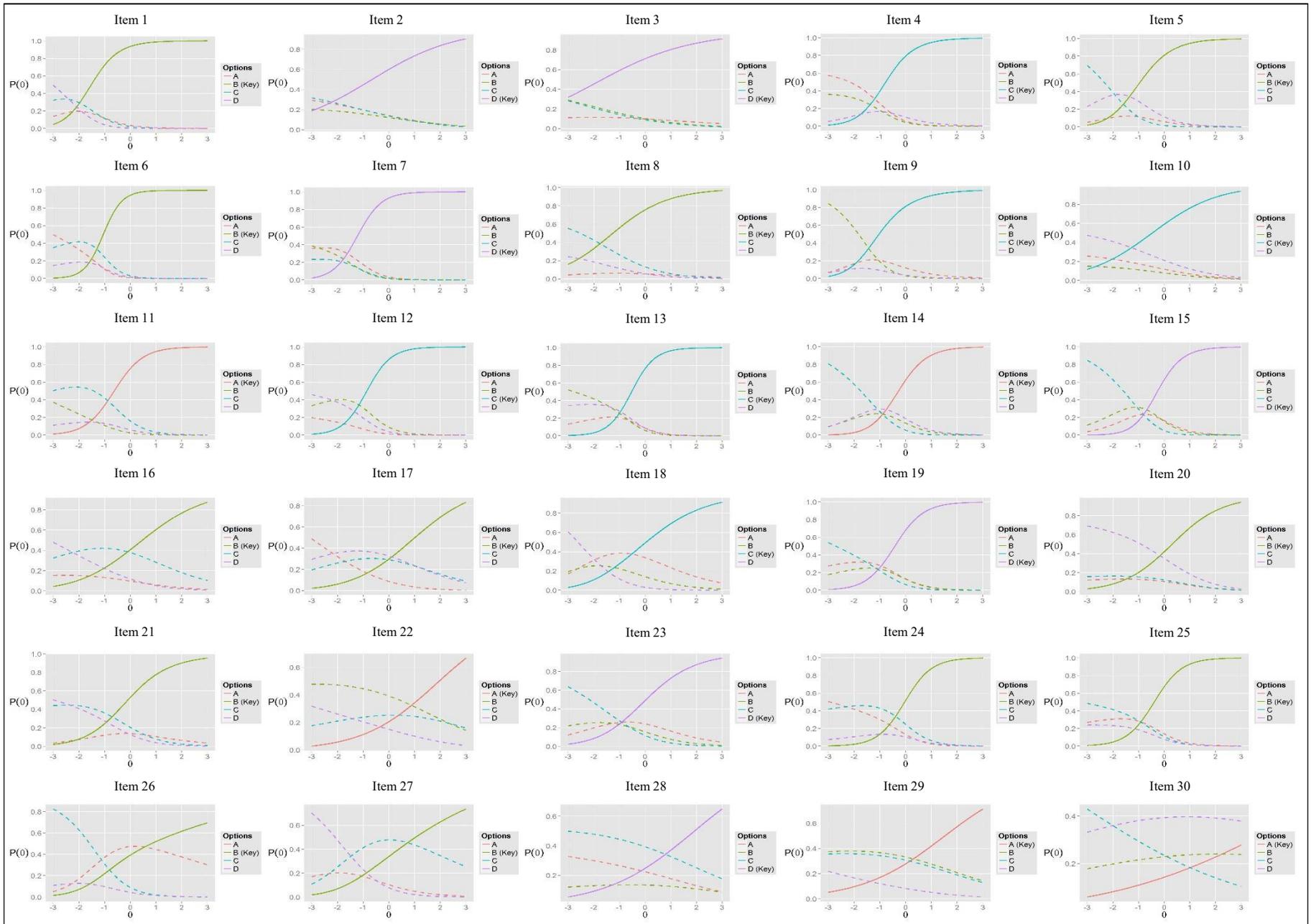


Figure 54 NRM Item Characteristic Response Curves for Form B

## Appendix D: R codes

The data were fitted to the unidimensional dichotomous Rasch model in Winsteps, but the graphics were generated in R using the ggplot2 and reshape2 packages as well as customized R functions. The R code used to create the customized functions and the uniform DIF graphics was specified as follows:

```
#####  
### Packages  
#####  
library(ggplot2)  
library(reshape2)  
  
#####  
### Functions  
#####  
  
# Unidimensional dichotomous Rasch model  
prob_rasch_dicho = function(b, theta){return(1 / (1 + exp(-(theta - b))))}  
  
# Function to create a uniform DIF graphic  
plot_rasch_dicho_dif = function(b_params, group_names, line_types='solid'){  
  n_b = length(b_params)  
  if (n_b != length(group_names)){  
    stop("The number of parameters must be equal to the number of groups")  
  }  
  
# Generating the vector of theta  
  theta = seq(-3, 3, 0.01)  
  N = length(theta)  
  
# Formatting the data matrices  
  probs = lapply(b_params, prob_rasch_dicho, theta)  
  probs = as.data.frame(matrix(unlist(probs), ncol=n_b))  
  colnames(probs) = group_names  
  X = cbind(theta, probs)  
  X = data.frame(X)  
  X = melt(X, id.vars = 'theta')  
  
# Defining the line types  
  len_type_error_message = paste("The argument 'line_types' can be a string 'solid' or a vector of  
of size", n_b, " (i.e. number of groups)")  
  if (length(line_types) == 1){ if (line_types == 'solid'){line_types = rep('solid', n_b)}  
else{stop(len_type_error_message)}}else  
  if (length(line_types) != n_b){stop(len_type_error_message)}
```

### **# Plotting**

```
g = ggplot(data = X, aes(x = theta, y = value, color = variable, linetype=variable)) +  
geom_line(lwd = 1) + # lwd = width of line  
ggtitle("") + # Modifying the title. By default, it is empty "
```

### **# Formatting the x axis**

```
scale_x_continuous(name = expression(theta), # setting the label for the x axis  
breaks = seq(-3, 3, 1)) + # setting the values of the x axis: from -3 to 3 by increment of 1
```

### **# Formatting the y axis**

```
scale_y_continuous(name = expression(paste(P, (theta))), breaks = seq(0, 1, 0.2), limits=c(0, 1))  
+
```

### **# Formatting the linetype**

```
scale_linetype_manual(values = line_types, name="Gender", labels=group_names) +
```

### **# Formatting the labels of the legend**

```
scale_color_discrete(name="Gender", labels=group_names) +
```

### **# General settings for the graph**

```
theme(text=element_text(size=18), # font size  
axis.title.y=element_text(angle=0, vjust=0.5), # orientation (angle) and position of the y label  
legend.background = element_rect(fill="gray90", size=.5), # setting for the legend  
panel.grid.minor = element_blank()) # eliminate the inner grid of the graphic  
return(g)}
```

### **# Function to create a uniform DIF graphic from a table**

```
plot_rasch_dicho_dif_from_table = function(dif_table, entry, line_types='solid'){  
i = which(dif_table$Entry == entry)  
params = dif_table[i, 2:ncol(dif_table)]  
x = colnames(dif_table)[2:ncol(dif_table)]  
group_names = sub('!',',', x, fixed=T)  
plot_rasch_dicho_dif(params, group_names, line_types)}
```

### **# Function to save all graphics**

```
save_multi_plot_rasch <- function(dif_table, my_directory, line_types=c('solid', 'dotted')) {  
entries = dif_table$Entry  
for (entry in entries){
```

### **# Function to create each graphic**

```
plot_rasch_dicho_dif_from_table(dif_table, entry, line_types)
```

### **# Function to save each graphic**

```
ggsave(filename = paste0('Item', entry, '.jpg'),  
path=my_directory, width = 7, height = 4)}
```

```
#####
### Import the data (i.e., item difficulty parameters by group membership)
#####

infile = 'file path'
dif_table = read.csv(infile)
dif_table

# Create one graphic at a time
plot_rasch_dicho_dif_from_table(dif_table, entry='Item 1', line_types=c('solid', 'dotted'))

# Save all graphics in a directory
save_multi_plot_rasch(dif_table, 'file path', line_types = c('solid', 'dotted'))
```

The nominal response model ([NRM], Bock, 1972) was fitted to the data using the R package mcIRT and the graphics were generated using the ggplot2, reshape2 and RColorBrewer packages as well as customized R functions. The R code used to create the customized functions and to fit the NRM model was specified as follows:

```
#####
### Packages
#####
library(ggplot2)
library(reshape2)
library(RColorBrewer)
library(mcIRT)

#####
### Functions
#####

# Function to create the item category response curves

# Zv : vector of zeta parameters
# Lv: vector of lambda parameters
# theta: vector of theta parameters
prob_nrm = function(Zv, Lv, theta){
  n = length(theta)
  v1 = rep(1, n)
  catexp = exp(t(t(v1)) %*% t(Zv) + t(t(theta)) %*% t(Lv)) # numerator
  sumcat = apply(catexp, 1, sum) #denominator
  probs = catexp / sumcat
  return(probs)}
}
```

```

# Function to generate a graphic for an item
# fit: output after fitting the NRM
# item: selected item
# key: position of the correct response (from 1 to 4)

plot_nrm = function(fit, item = 1, key = 1, filepath_to_save = NULL, line_type_key='solid',
line_types = 'dashed'){

# Extracting Zeta and Lambda parameters for the given item
ZLpar = fit$ZLpar[[1]][[item]]
npar = length(ZLpar)
Z = as.vector(ZLpar[1:(npar/2)])
L = as.vector(ZLpar[(1 + npar/2):npar])

# Generating the vector of theta
theta = seq(-3, 3, 0.01)
N = length(theta)

# Computing the probabilities for the distractors and formatting the data matrices
probs = prob_nrm(Z, L, theta)
colnames(probs) = LETTERS[1:(npar/2)]
X = cbind(theta, probs)
X = data.frame(X)
X = melt(X, id.vars = 'theta')

# Defining the line types
len_type_error_message = "The argument 'line_types' can be a string 'dashed' or a vector of of
size 4" if (length(line_types) == 1){if (line_types == 'dashed'){line_types = rep('dashed', 4)}
else{stop(len_type_error_message)}}else
if (length(line_types) != 4){stop(len_type_error_message)}
line_types[key] = 'solid'

# Defining the correct option and formatting the labels for the legend
opts = LETTERS[1:4]
opts[key] = paste(opts[key], '(Key)')

# Plotting
g = ggplot(data = X, aes(x = theta, y = value, color = variable, linetype=variable)) +
geom_line(lwd = 1) + # lwd = width of line
ggtitle("") + # Modifying the title. By default it is empty "

# Formatting the x axis
scale_x_continuous(name = expression(theta), # setting the label for the x axis
breaks = seq(-3, 3, 1)) + # setting the values of the x axis: from -3 to 3 by increment of 1

```

```

# Formatting the y axis
scale_y_continuous(name = expression(paste(P, (theta))), breaks = seq(0, 1, 0.2)) +

# Formatting the linetype
scale_linetype_manual(values = line_types, name="Options", labels = opts) +

# Formatting the labels of the legend
scale_color_discrete(name="Options", labels = opts) +

# General settings for the graph
theme(text=element_text(size=18), # font size
axis.title.y=element_text(angle=0, vjust=0.5), # orientation (angle) and position of the y label
legend.background = element_rect(fill="gray90", size=.5), # setting for the legend
panel.grid.minor = element_blank()) # eliminate the minor grid
return(g)}

# Function to save graphics
save_plot <- function(fit, key, items, my_directory, line_type_key, line_types) {
  for (i in items){

# Function to generate each graphic
plot_nrm(fit, item = i, key = key[i], line_type_key=line_type_key, line_types=line_types)

# Saving the graphics
  ggsave(filename = paste0('Item', i, '.jpg'),
    path=my_directory,
    width = 7,
    height = 4)
  }
}

#####
### NRM data modelling
#####

library(mcIRT)

# Setting reshOBJ to model the data. This is required by mcIRT
dp = "direct path to import the data"
da = read.csv(dp,header=TRUE)
items = 3:32 # items = seq(2, 31, 1)
da = da - 1

# Items
items = 2:31 # items = seq(2, 31, 1)

```

```

# Correct answer or key
correct = c(a vector with the key)
correct = correct - 1

# reshOBJ # this is to reshape raw data prior to modelling
reshOBJ = reshMG(da = da, items = items, correct = correct)

# Fitting NRM
fit1 = nrm(reshOBJ)

#####
### Generating the graphics
#####

key = correct + 1 # for the graphics the values must start at 1

# Plotting one graphic at a time
plot_nrm(fit1, item = 1, key = key[1], line_type_key = 'solid', line_types=c('twodash', 'dotted',
'dashed', 'dotdash'))
plot_nrm(fit1, item = 2, key = key[2])
plot_nrm(fit1, item = 3, key = key[3])
plot_nrm(fit1, item = 4, key = key[4])
plot_nrm(fit1, item = 5, key = key[5])
# ...

# Saving the graphics in a directory
my_directory = "file path"
save_plot(fit1, key, items = 1:30, my_directory, line_type_key = 'solid', line_types=c('twodash',
'dotted', 'dashed', 'dotdash'))

```