Research Monograph

A Novel Multimethod Approach to Investigate Whether Tests Delivered at a Test Centre are Concordant with those Delivered Remotely Online

An Investigation of the Concordance of the CAEL

Bruno D. Zumbo

Paragon Testing Enterprises
Test Research & Development Division

THE UNIVERSITY OF BRITISH COLUMBIA
UBC Paragon Research Initiative
Statistical Science of Measurement

First Published 2021 by Test Research and Development Division of Paragon Testing,
Paragon Testing Enterprises, Vancouver B.C., Canada
and by
First Published 2021 by University of British Columbia, UBC Paragon Research Initiative,
Vancouver B.C., Canada

Suggested citation (APA Style):
Zumbo, B. D. (2021). *A Novel Multimethod Approach to Investigate Whether Tests Delivered at a Test Centre are Concordant with those Delivered Remotely Online: An Investigation of the Concordance of the CAEL [monograph]*. Vancouver, BC: Paragon Testing Enterprises/UBC Paragon Research Initiative, University of British Columbia.

Errata

- July 17, 2021, updated version of the monograph includes a correction of a typographical error in the GLIM model specification on page 115.
- July 21, 2021, correction to CAEL band 60 descriptors, page 49.

This project was coordinated through Paragon Testing Enterprises' 2020/2021 *Strategic Projects Initiative (#3.2)*. The project owner was Dr. Bruno Zumbo, and the divisional lead was Dr. You-Min Lin.

Dr. You-Min Lin provided coordination and oversight of the project and contributed their expertise in test research and development and linguistic theory to classify test takers' reported first languages into two major language families in *Sections 5.2* and *5.4*.

Dr. Michelle Chen provided coordination and oversight of test taker data extraction by the psychometrics and quality assurance team members (Asbury, Chau, Tang) and contributed their test research and psychometrics expertise. She led the work with contributions by Dr. Bruno Zumbo, Taylor Asbury, Lok Heng Chau, and William Tang to apply the statistical matching methods to score comparability and misclassification analysis reported in Chapters 9 and 10.

Dr. Bruno Zumbo is the primary author of this report. The integrated assessment theory and research methods developed herein reflect his thirty-year research program on the statistical science of measurement. He is Professor and Distinguished University Scholar, Paragon UBC Professor of Psychometrics and Measurement. In 2020 he was awarded the Tier I Canada Research Chair in Psychometrics and Measurement. Its central theme of *Equity and Fairness at the Nexus of Data Science, Digital Innovation, and Social Justice* is reflected in this report. He is also Paragon Testing's Chief Scientific Officer.

Paragon Testing Enterprises was a subsidiary of the University of British Columbia incorporated in 2009 to commercialize UBC's English language proficiency tests. In April 2021, it was acquired by Prometric Canada Testing Services®, parent company Prometric LLC.

You-Min Lin, Ph.D., Test Research and Development Specialist | Acting Manager, Test R&D Projects, Planning and Budgets, Test R&D Division, Paragon Testing Enterprises

Michelle Y. Chen, Ph.D., Validation Studies and Test Research Lead, Test R&D Division, Paragon Testing Enterprises

Taylor Asbury, M.Sc., Psychometrics Lead, Test Operations Division, Paragon Testing Enterprises

Lok Heng Chau, M.A., Psychometrics Specialist, Test Operations Division, Paragon Testing Enterprises

William Tang, B.Sc., Quality Assurance Assistant, Test Operations Division, Paragon Testing Enterprises

Summary

A novel multimethod research methodology and accompanying statistical methods for operational and validity research is described in response to the emergence of remote online proctored test administration. The multimethod strategy was designed to allow for a robust comparison of the test centre and online test performance that far exceeds conventional methods to investigate the comparability of tests- i.e., their concordance.

The rigour and logic of our methodology are grounded in test validity and a framework based on four key principles. First, Angoff's (1993) matching principle allows for the definition of optimal statistical psychometric methods that do not confound concordance with true differences in item performance (i.e., impact). Second, the equity principle states it should be a matter of indifference to a test taker or a test user about which test takers choose between two modes of test administration (test centre or online). Third, the test use principle states that the comparison across test administrations should focus on the scale on which scores are reported- for example, band scores on each of the (four) components of a language test rather than an item-by-item comparison. Fourth, there is an overall principle of multiple sources of evidence (multimethod methodology) that calls for more than one source of evidence supporting the concordance investigation to rule out rival plausible alternative interpretations and ferreting out multiple sources of potentially hidden invalidity.

This novel methodology is applied to investigate the concordance of the *CAEL* delivered at a test centre and online. The concordance study used a sample of 1,455 *CAEL* test takers, 765 test takers who completed the *CAEL* at a test centre, and 690 who completed it remotely online between June and October 2020. The findings from the six statistical and psychometric methods are consistent. The sample of the test centre and online test takers were equivalent, and the test performance was found to be consistently concordant. Together, this is strong evidence that the conclusions from the *CAEL* band scores from the test centre and online versions are concordant, fully comparable band score performance of test takers at various levels of *CAEL*'s language domains.

The results of this analysis will serve as evidence to support the interpretation and use of scores from tests administered online using a remote-proctored test-delivery platform or at a test centre. The multimethod approach introduced in this monograph is a general model for other concordance studies that provides a principled rationale for designing such studies to investigate any delivery modes; for example, a concordance study may investigate a test administered simultaneously at a test centre, remotely online, at pop-up administration centres, and in paper-and-pencil format. A rigorous test of the concordance is of importance to test users, test providers, and external stakeholders who rely on valid and comparable test performance and test use across different test administration modalities.

Table of Contents

Summary 4

Table of Contents 5

Part I - A Novel Multimethod Approach for Operational and Validity Research 10

Chapter 1 Does Remote Online Test Administration have a Significant Effect on Test Scores? 11

1.1 The Emergence of Remote Online Test Administration 11

1.1.1 The Kinds of Online Tests 11

1.2 Purpose and Organization of the Report 14

Chapter 2 Our Validity Theory and Validation Methodology 17

2.1 Theoretical Framework for Our Methodology to Investigate Online Testing 17

2.1.1 Blending Zumbo's Model and Kane's Argument-Based Approach and the *Bona Fides* of the Claims from Test Scores 19

2.2 The Theoretical Framework That Forms the Principles and Logic of Our Methodology to Investigate Online Testing 22

2.2.1 Principles, Criteria, and Evidence Emerging from Zumbo's DLD Framework 29

2.2.2 Toward a Psychometrics For Studies of Concordance - A Brief History of Concordance Methods Leading to a Definition of Concordance 34

2.3 Research Design Options and Decisions Implied From the Matching Principle 37

2.3.1 Basic Concepts and Terminology 38

2.3.2 Option #1: Between Subjects Design 38

2.3.3 Option #2: Within Subjects Design 39

2.3.4 Option #3: Matched Subjects Design 40

2.3.5 Choice of Research Design – There is No Perfect Design; Rather, it is a Tradeoff of Advantages and Disadvantages 42

2.4 Transitioning to Part II – A Concordance Study of the *CAEL* Online 44

Part II - Applying the Novel Multimethod Approach for Operational and Validity Research to the Design of the Concordance of the *CAEL* Delivered at Test Centres and Online 46

Chapter 3 Brief Description of the *Canadian Academic English Language (CAEL)* Test 47

3.1 Test Format 47

3.2 Scoring and Reporting of Results 48

3.3 Online Test Delivery and Test Proctoring 49

Chapter 4 Applying the Novel MultiMethod Approach to Plan the Concordance Study of the *CAEL* 50

4.1 The Kinds of Online Tests (*Section 1.1.1*) 50

4.2 Applying the Theoretical Framework That Forms the Principles and Logic of Our Methodology to Investigate Online Testing (*Section 2.2*) 50

4.3 Unit of Analysis – The Component Band Score (*Section 2.2*) 51

4.4 Concordance Study Design (*Section 2.3*) 51

4.5 Evidence for Score Comparability 51

4.6 Transitioning to Parts III to V of the Report 52

Part III - Investigating the Question of *CAEL Test Taker* Comparability Arising from the Between-Subjects Design 53

Chapter 5 Test Taker Comparability 54

5.1 Sample Size and Composition 54

5.1.1 Adequacy of the Sample Size for Planned Analyses 55

5.1.2 Comparability of the Composition of the Mode of Test Delivery Groups by Gender of the Test Taker 55

5.1.3 Comparability of the Mode of Test Delivery Groups by the Age (in years) of Test Takers 56

5.1.4 Comparability of the Composition of the Mode of Test Delivery Groups by Test Takers' Reported First Language – A Language Families Approach 57

5.1.5 Comparability of the Composition of the Mode of Test Delivery Groups by Gender, Age, First Language (Language Family) 58

5.1.6 Comparison of the Age Distribution and also the Band Scores on the Four Components Earned for Test Takers in 2019 Test Centre, 2020 Test Centre, and 2020 Online 64

5.2 A Consideration of Test Takers' Reported First Language (L1): Language Family as a Grouping Variable 69

5.2.1 Procedures 70

5.2.2 Results 70

5.2.3 Recommended Grouping 72

5.3 Overall Conclusions About the Comparability of Test Takers 73

5.4 Appendix Chapter 5- Test-Taker Count in Each Language Family (Detailed Breakdown, Including Individual L1s) 74

Part IV - Five Strands of Evidence to Investigate *Test Score* Comparability 78

Chapter 6 Covariance Analysis- Dispersion Matrices and Comparative Factor Analyses. 79

6.1 Sample Data and Treating Band Scores as Continous vs Ordered Categorical Variables 80

6.2 Equality of Covariance Matrices 80

6.2.1 Results – Statistical Tests of the Hypothesis of Equal Dispersion Matrices 81
6.2.2 Results – Visualizing Tests of Equality for Covariance Matrices..... 81
6.3 Multi-Group Essential Unidimensionality..... 82
6.4 Multigroup Factor Analysis 83
6.4.1 Results of the Multigroup Factor Analysis..... 84
6.5 Conclusions 85
Chapter 7 Visualizing Comparability: Comparing Groups of Test Takers Using a Novel Kernel Smoothing Band Score Function..... 86
7.1 Overview- The main purposes of this chapter and overall conclusions..... 86
7.2 Visualizing Comparability of the *CAEL* Delivered at a Test Centre as Compared to Online 87
7.2.1 Description of the Graphs Used to Investigate the Comparability 87
7.3 Four Noteworthy Strengths of the Novel Kernel Smoothed Band Function Approach . 88
7.4 Statistical Framework: Definitions, Assumptions, and Modeling..... 89
7.4.1 Assumption of Essential Unidimensionality 91
7.5 Using Kernel Smoothed Band Score Function to Compare the Band Scores for Component Functions of the *CAEL* (Online) and *CAEL* (Test Centre) 91
7.6 Minimum Sample Sizes for the Various Graphical Comparisons 92
7.6.1 Study Data..... 94
7.7 Graphs of the Four Component Band Scores - Test Centre Compared to Online (Main Effect of Mode of Test Delivery)..... 94
7.7.1 Comparability of the Band Scores on the Listening Component 95
7.7.2 Comparability of the Band Scores on the Reading Component..... 96
7.7.3 Comparability of the Band Scores on the Speaking Component 97
7.7.4 Comparability of the Band Scores on the Writing Component..... 97
7.7.5 Overall Comparability of the Band Scores..... 98
7.7.6 Comparability (Density) Distributions of the Observed Overall Band Score..... 99
7.8 Graphs of the Four Component Band Scores - Test Centre Compared to Online for Female and Male Test Takers (Mode by Gender) 100
7.8.1 Delivery Mode by Gender of Test Taker Comparability of the Band Scores on the Listening Component..... 100
7.8.2 Delivery Mode by Gender of Test Taker Comparability of the Band Scores on the Reading Component 101
7.8.3 Delivery Mode by Gender of Test Taker Comparability of the Band Scores on the Speaking Component..... 102

7.8.4 Delivery Mode by Gender of Test Taker Comparability of the Band Scores on the Writing Component 103

7.9 Graphs of the Four Component Band Scores – Gender DIF for Female and Male Test Takers (Main Effect of Gender) 104

7.9.1 Comparability of the Band Scores by Gender of the Test Taker for the Listening Component 104

7.9.2 Comparability of the Band Scores by Gender of the Test Taker for the Reading Component 105

7.9.3 Comparability of the Band Scores by Gender of the Test Taker for the Speaking Component 106

7.9.4 Comparability of the Band Scores by Gender of the Test Taker for the Writing Component 107

7.10 Graphs of the Four Component Band Scores – Test Taker's Report First Language (Language Family) DIF (Main Effect of First Language) 108

7.10.1 Comparability of the Band Scores by First Language of the Test Taker for the Listening Component 108

7.10.2 Comparability of the Band Scores by First Language of the Test Taker for the Reading Component 109

7.10.3 Comparability of the Band Scores by First Language of the Test Taker for the Speaking Component 110

7.10.4 Comparability of the Band Scores by First Language of the Test Taker for the Writing Component 111

7.11 Conclusions 111

Chapter 8 Generalized Linear Model Approaches – DIF Analyses of Test Centre vs Online *CAEL* Test Performance 112

8.1 Overview and Conclusions 112

8.2 Statistical Method 113

8.2.1 Minimum Sample Size 113

8.2.2 Specification of the family of GLIM DIF Models 113

8.3 Results – Concordance of the *CAEL* (Online) and *CAEL* (Test Centre) and Its Moderation by the Test Takers Gender and First Language 116

8.3.1 Listening 116

8.3.2 Reading 118

8.3.3 Speaking 119

8.3.4 Writing 120

8.4 Conclusions 121

8.5 Appendix Chapter 8 – Frequency Plots of the Four Band Scores 122

Chapter 9 Concordance Analysis Using Statistical Matching as an Alternative to Within Subjects Designs or Randomized Experiment 126

9.1 Sample 127

9.2 Exact Matching Method 127

9.3 Results 128

9.4 Conclusion 129

Chapter 10 Measurement Error/Misclassification Analysis from a Statistically Matched Sample: Comparative Decision Consistency and Decision Accuracy 131

10.1 Misclassification and Concordance 131

10.1.1 Evidence of Concordance: Comparative Misclassification Take the Matching Principle Into Account 132

10.2 Sample Data 133

10.3 Decision Accuracy and Decision Consistency Statistics 133

10.4 Results and Conclusions of Decision Accuracy and Decision Consistency – Reduced Matched Sample Data 133

Part V - Bringing the Multimethod Strands Together 137

Chapter 11 Closing Remarks and Conclusion 138

11.1 A Rigorous Integrated Method for Concordance Studies 138

11.2 The Concordance of Test Centre and Online Delivery of the *CAEL* 139

11.3 Next Steps 141

References 142

Part I - A Novel Multimethod Approach for Operational and Validity Research

Investigating the Concordance of Test Performance at a Test Centre and Remote Online Test Delivery

Chapter 1 Does Remote Online Test Administration have a Significant Effect on Test Scores?

1.1 The Emergence of Remote Online Test Administration

Remote proctored test taking provides an additional testing option for candidates, enabling them to take examinations at home, or any setting conducive to testing, without needing to travel to a test centre. A remote test setting will typically be a private, quiet, comfortable room without distractions that allow the testing session to be overseen by a secure, online proctoring solution. As such, the most common remote location is at a test taker's home, but for example, a private work office may also be an option. The test candidates are responsible for ensuring the technical equipment's suitability. Of course, should candidates wish to take their exam in a test centre, this option is still available.

As the use of remote testing options steadily increases at testing agencies, the following question must be answered:

> Does remote testing result in lower, higher, or the same scores as those obtained by taking the test at a test centre? This question of test scores' interchangeability would be an important contribution to the test validity for a testing program allowing for both remote and test centre testing opportunities.

There has been little research investigating whether the remote administration of a given standardized test has a significant effect on test scores. Accordingly, there is a need for empirical studies of the test score interchangeability or equivalence between these different test administration modes. Likewise, these empirical studies should be guided by a theoretical framework and research methodology to investigate test score interchangeability and equivalence that informs ongoing test validation. Although there are nuanced differences between these terms, the terms test ‘administration’ and ‘delivery’ will be used interchangeably throughout this monograph.

1.1.1 The Kinds of Online Tests

The ongoing global pandemic's urgency has led to a surge in remote online test options along with immensely accelerated test research and development timelines. As Isbell and Kremmel (2020) remind us, the administration of high-stakes language proficiency tests at test centres has been disrupted worldwide due to the 2019 novel coronavirus pandemic. Institutions that rely on language test scores have been forced to use scores from tests administered remotely (online) resulting from (i) a ***different test administered online that substitutes for the original*** or (ii) an **online *version of an existing test***.
The test providers may design these online tests as a stopgap during the pandemic, an alternative mode of test administration that continues to be offered at a test centre that can also be taken remotely, or an initially designed test stopgap that later becomes a standard

alternative. It is useful to contrast these three instantiations of the two types of online language tests to acquire a sense of the two modes of test administration's common and unique features.

1) A different test administered online that substitutes for the original: An instantiation of a different test administered online that substitutes for the original is the *IELTS Indicator* (IELTS, 2020). The *IELTS Indicator* is a stopgap to their Academic test during the global pandemic. As Isbell and Kremmel state, IELTS (2020) is very clear on their website that the *IELTS Indicator* provides an "indicative score" only. Furthermore, the test provider is clear that it is a temporary expedient: "The Academic test is available for a limited time while IELTS testing is currently suspended due to COVID-19. Educational providers can use IELTS Indicator to help them gauge the English language."
    - The use of the terms "indicator" and "indicative score" in this setting, as widely used in the social sciences, implies that the indicative scores for each of the four skills -Listening, Reading, Writing and Speaking – can stand in for the less directly quantifiable IELTS test skill scores allowing the test users to carry on as if the four skills were measured with the IELTS test. As IELTS (2020) states: "Educational providers can use *IELTS Indicator* to help them gauge the English language ability of future students while IELTS testing is suspended."
2) An online version of an existing test: In contrast to the *IELTS Indicator*, other test providers offer an online test where nearly everything about the tests themselves is the same except the test administration features that come with remote testing, such as online proctoring and registration. In this family of online tests, the online alternatives are designed so that the test scores arising from the online test administration can be accepted and used in the same way as the test center version.
   Two instantiations of these kinds of online tests are the *CAEL* and the *TOEFL iBT*.
    a) The first example of such a test is the *CAEL*. In June 2020, Paragon Testing launched as an online alternative mode of test delivery with the same test format, content, and reporting scale as the *CAEL* test delivered at test centres. By design, the *CAEL* delivered online is an alternate mode of test delivery alongside the *CAEL* delivered at a test centre. Under current pandemic restrictions, the *CAEL* online has been available to test takers in Canada and the USA (and soon in Mexico). In contrast, the *CAEL* at a test centre is available to test takers internationally.
    b) The second example is the *TOEFL iBT*. As Stacey (2020) notes, in March 2020, Educational Testing Service (ETS) began offering the *TOEFL iBT Special Home Edition* until in-person testing at test centres could resume. This initial positioning and the designator *Special* Home Edition suggest that the *Special Home Edition* may have started as a temporary stopgap (akin to the *IELTS Indicator*) for the test offered at a test centre. However, the recently re-branded *TOEFL iBT Home Edition* is now similar to the *CAEL* test online. Stacey (2020) described that in December 2020, ETS announced that the *TOEFL iBT Special*

Home Edition would be re-branded the *TOEFL iBT Home Edition* and added to its permanent product portfolio. As described in ETS (2021): "The *TOEFL iBT Home Edition* is now a standard option for test takers and will be available for the foreseeable future, along with the option of testing at a test center." Likewise, as a TOEFL iBT product family member, the scoring criteria, scoring process, and score scale of the *Home Edition* are the same as a test taken at a test center. ETS (2020) further stated that "test takers can expect the same valid and reliable tests that are administered in test centers from the comfort of home." As such, the two administrative models' score comparability must be established.

There are other alternative online language tests, see Isbell and Kremmel (2020), but the two classes of alternatives described above capture the essential features of a wide range of options, all of whom appear to fall into one of the two broad categories (i) different test administered online that substitute for the original or (ii) an online version of an existing test.

Treating a test as a stopgap or as an indicator does not cancel the concern about test validity. Whether one has a stopgap test during the pandemic, a test designated an 'indicator,' or an online version of an existing test, high-stakes decisions are made based on the resulting test scores. Likewise, whether the test is available concurrently at a test centre or not, there is a need to determine if administration mode (tests administered at a test centre or remote online testing) affects these language tests' high-stakes decisions. Although it is not the focus of this report, the same reasoning applies to test administration at temporary “pop-up” facilities.

Remote online testing is a relatively new phenomenon born out of the need created by the disruption due to the global pandemic or the timeline was sped-up due to the pandemic for those who were already considering online options.
In either case, ***there is a need for a theoretical framework and research methodology*** to determine if administration mode (test centre and online) affects these tests' high-stakes decisions.

1.2 Purpose and Organization of the Report

As Langenfeld (2020) reminds us, it was in the mid-1990s that saw internet-based testing was introduced in test centers. Shortly after, schools and certification and licensure programs began exploring different formats to allow for more convenient test taking opportunities. The vision of any time, anywhere testing and on-demand testing is the primary advantage of technology-based and internet-based testing that emerged in the late-1990s (Addicott & Foster, 2017; Bartram, 2009; Zumbo, 2002). In *Speaking Personally—With David Foster* (2010), it is clear that from 1990 to 2005, there was a great deal of demand, technological innovation, as well as progress in test design and delivery that lead to a global network of testing centers and test centre providers to deliver high-stakes online tests in secure settings.

A cursory glance at some of the high watermarks in the history of the administration of standardized testing

- ***One of the first uses of computerization in large-scale testing:*** In 1958 the *Iowa Assessments* were computerized. Iowa also introduced computerization to the scoring of tests and production of reports to schools.
- In 1970 ***changes in computer technology and accessibility of fully programmable desktop (personal) computers*** revolutionized psychometric theory and test administration practices.
- The ***transition to computer-based testing*** (CBT): By 1990 many large-scale testing programs moved to CBT. This transition created a need for test centres where these CBTs could be administered in a standardized manner safely and securely.
- Introduction of ***Internet-based testing at test centres***: During 1995 to 2005, there was a great deal of demand, technological innovation, as well as progress in test design and delivery that lead to a global network of testing centers and test centre providers to deliver high-stakes online tests in secure settings
- ***Remote, online at-home, testing emerged in 2020*** in reponse to the fact that administration of high-stakes language proficiency tests at test centres had been disrupted worldwide due to the 2019 novel coronavirus pandemic.

Most certainly, if the history of testing tells us anything, several testing organizations were already considering expanding the alternatives beyond test centres; however, the timeline was sped up due to the pandemic. Several technological advances in online testing that emerged during the rush to move online for remote proctored administration at the start of the

pandemic will likely continue post-pandemic. As seen starting in the early 1990s, test takers and test users like the convenience of any time, anywhere, testing.

Useful Points to Keep In Mind

Even if all guidelines for creating an alternative mode of test delivery have been followed carefully and the smooth delivery have been checked meticulously, there is no guarantee that tests administered in multiple modes behave identically to their original versions.

However, a helpful resource is statistical analyses of the distributions of the test data in populations from the different modes of test delivery, with appropriate care for matching of the test data distributions on the intended-to-be-measured attribute and or other relevant covariates depending on the study design.

The term '***concordance***' is used throughout this monograph to denote the agreement between the interpretation and use of scores arising from a test simultaneously administered in more than one mode of delivery (e.g., at test centre, online, and paper-and-pencil). The term concordance also allows for it to be considered as a matter of degree as described in *Chapter 2*, *Section 2.2*. Other similar terms include interchangeability or exchangeability. However, comparability, although a possible alternative term, is typically used in the context where test scores are used in, for example, ranking or comparing demographic groups of test takers, nations, states, or provinces.

As such, there is a need for a theoretical framework and research methodology to determine if administration mode (test centre and online) affects these tests' high-stakes decisions. To fill this need for a theoretical framework and accompanying research methodology, we have three objectives.

1. ***Describe a theoretical framework*** for unpacking the assumptions that support the validity of claims made from alternate online tests, whether they are an indicator or otherwise. The framework aids in establishing the degree of interchangeability implied when using scores resulting from these two modes of test administration and determining the key empirical evidence needed to support the validity of the claims made from these tests.
2. ***Describe and demonstrate the novel multimethod research methodology*** emerging from this framework while also meeting our third objective.

3. ***Reporting the findings* of a study** investigating whether the *CAEL* test scores delivered at a test centre are concordant with those delivered remotely (online). This study exemplifies how one can adapt and apply the novel framework and research methodology in objectives one and two.

To achieve these three objectives, *Chapter 2* introduces our novel validity theory and validation methodology that highlights the need to investigate test taker and test score comparability across the two modes of test administration- test centre and online. *Part II* of the report signals the transition to this report's second and third objectives by briefly describing the *CAEL* to help readers interpret the findings and applying the methodology in *Chapter 2* to the design of the concordance study of the *CAEL CE*. Throughout this monograph, we use *CAEL* and *CAEL CE* interchangeably. *Part III* of the report contains the findings from the comparability of the test takers, whereas *Part IV* reports on the comparability of the test scores. *Part V* includes the closing chapter of the report that brings the multimethod strands together to make a coherent claim about the concordance of online and test centre administration of tests. We often refer to remote online test delivery of a test as "online" throughout this work.

The intent of this report is to provide the much-needed theoretically-grounded and empirically-based **industry standard** to support the adoption of online testing.

Chapter 2 Our Validity Theory and Validation Methodology

2.1 Theoretical Framework for Our Methodology to Investigate Online Testing

As Shear and Zumbo (2014) show, over the past 50 years, the concepts and theories of test validity have grown increasingly expansive, and the methods for test validation have become increasingly complex and multi-faceted. Shear and Zumbo state that validity theorists have highlighted the important distinction between validity and validation (Borsboom et al., 2004; Zumbo 2007a, 2009). Whereas validity is the property or relationship we are trying to judge, validation is an activity geared towards understanding and making that judgment. Zumbo (2009) reminds us of the importance that a guiding rationale (i.e., validity theory) must play in selecting and applying appropriate analyses (i.e., validation), while Borsboom et al. note that failing to distinguish between validity and validation can lead to conceptual and methodological confusion. These authors are highlighting the importance of having a clear concept of validity, which can then be used to guide the use of validation methods.

Since the early 1990s, Michael Kane has been one of the main proponents of an argument-based approach to validation, focusing on score interpretation and use driving the argument's structure. This approach helps focus validation efforts and clarify intended interpretations and uses. As Kane notes, the main advantage of the argument-based approach to validation is its guidance in allocating research effort and gauging progress in the validation effort (Kane, 2006). He argues that treating validity as a property of test score interpretations and uses allows for flexibility in the sources or kinds of evidence used to support inferences and uses with different kinds of tests or particular situations.

Kane's more recent views use the terms ‘interpretation or use argument’ (IUA) and ‘validity argument’ (VA) (Kane, 2012). Kane writes:

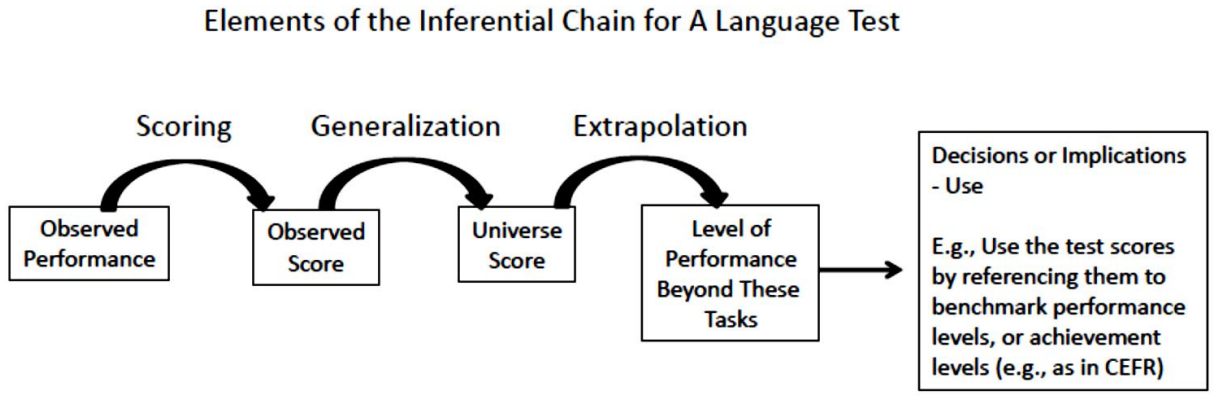
> *The argument-based framework is quite simple and involves two steps. First, specify the proposed interpretations and uses of the scores in some detail. Second, evaluate the overall plausibility of the proposed interpretations and uses.*
>
> *The argument-based framework is quite flexible in the sense that it does not specify any particular kind of interpretation or use for assessment scores, and invites assessment developers and users to specify their proposed interpretations and uses. Any kind of interpretation or use can be proposed, but the claims being made should be justified, and more ambitious interpretations and uses impose more demands for justification.* (Kane, 2012, p. 4)

Additionally, Sireci (2013) called for a simpler approach to developing validity arguments that focuses on explicating the testing purposes, as suggested by the *Standards for Educational and Psychological Testing* (AERA et al., 2014). At that point, developing an interpretive argument

becomes unnecessary, and validation can directly address intended interpretations and uses. Especially relevant is the statement in the *Standards* that "... documentation of the purpose and intended uses of the test, as well as detailed decisions about test content, format, test length, psychometric characteristics of the items and test, delivery mode, administration, scoring and score reporting" (AERA et al., 2014, p. 76).

Figure 2.1 depicts the scoring, generalization, and extrapolation inferential chain linking, for example, observed test performance on a component of a language test (e.g., the listening component) to the test use and decisions by a test user.

Figure 2.1 A Schematic of Kane's Argument-Based Approach to Validation



Kane's model unpacks the hidden assumptions when making a claim from a test score. It is the investigation of the backing for these claims that is the central focus of Kane's model.

In unpacking the meaning of the IUA and VA, he goes on to explain that (i) the central point of the interpretive argument is to make the assumptions and inferences in the interpretation [of test scores/outcomes] as clear as possible, and (ii) the validity argument, on the other hand, is meant to provide a coherent analysis of the evidence in support of the proposed interpretation while allowing to rule-out rival plausible alternate interpretations.

In the validity framework presented herein, a logico-mathematical approach (the DLD framework) provides a mode of analysis to rigorously investigate both the IUA and VA and inform the resultant multimethod statistical methodology to support the validity argument.

We agree with the sentiment expressed in Kane's tenant; however, in line with Zumbo and Hubley (2016), we would caution that Kane's approach to validation may lead some researchers to take a precarious minimalist approach to validation leading to hidden sources of invalidity. Our 'non-minimalist' theoretical orientation will be evidenced in our advocacy for a multimethod research methodology. As Zumbo and Hubley state:

> *We do not think that the number of inferences or the plausibility of the assumptions made ('if the proposed interpretation is simple,' see Kane,*

> *2016) necessarily means less validity evidence is required. Provided evidence is appropriate to the test and situation, we strongly argue that presenting more, rather than fewer, sources or kinds of validity evidence is better. Rather like consulting more than one timepiece to determine the current time, this may lead to conflicting results, but we would argue that ultimately one would hope to come to a more well-rounded and informed conclusion.* (2016, p. 300)

Kane (2016) reminds the reader that there is debate about whether the evaluation of score-based uses should be included under the heading of 'validity,' per se. Our position on this matter is clearly stated in Hubley and Zumbo's integrated test validity framework (2011, 2013). A key focus is the idea of what inference and the strength of that information one is validating are important alongside the consequences of that decision.

2.1.1 Blending Zumbo's Model and Kane's Argument-Based Approach and the *Bona Fides* of the Claims from Test Scores

Since the publication of Messick's groundbreaking review of validity (Messick, 1989), the field of measurement, assessment, and testing has been calling out for a new and expanded evidential basis for test validation. We respond to Messick's call by blending key ideas in Kane's and in Zumbo's theories. The approach we are advocating builds on Kane's (2016) explication of (argument based) validity with the foci and refinements described by Zumbo and his colleagues that emphasize an explanation-focused view, transparency, and trending away from routine validation practices to shine a light on often hidden forms of test invalidity (Addey, Maddox, & Zumbo, 2020; Hubley & Zumbo, 2011, 2013; Stone & Zumbo, 2016; Zumbo & Hubley, 2016; Zumbo, 2007a, 2009, 2017).

As an explanatory model of test score variation, Zumbo's *explanation focused view of validity* is embedded within an ecological model of item responding that is situated within a pragmatic view of abductive explanation wherein one develops validity evidence for tests through abductive reasoning (Stone & Zumbo, 2016; Zumbo, 2007a, 2009). In contrast to inductive reasoning or deductive reasoning, abductive reasoning neither construes the meaning of the scores purely from empirical evidence nor presumes the meaning of the test to explain the score. Rather, abductive reasoning seeks the enabling conditions under which the score makes sense. In merging key ideas in Kane's and Zumbo and his colleagues' work, we make a case for a validity argument developed and supported empirically from various sources and kinds of validity evidence, focusing on the validation of test score use.

The central concept is establishing and recognizing the ***bona fides*** of the interpretation, decisions or implications (i.e., the claims) from test scores. As depicted in *Figure 2.2*, when one follows the line of evidence established by following the dark heavy arrows, there is a clear explanatory focus to the validation objectives, as per Zumbo's explanatory focused view of validity, all the while marshalling and organizing the evidence in the green boxes in the oval. This model drives our validation program of research stressing (i) transparency and the

unpacking of typically hidden sources of invalidity in validation practices (see Zumbo & Chan, 2014), and (ii) the evidence of the qualifications of the test for the proposed use and interpretation.

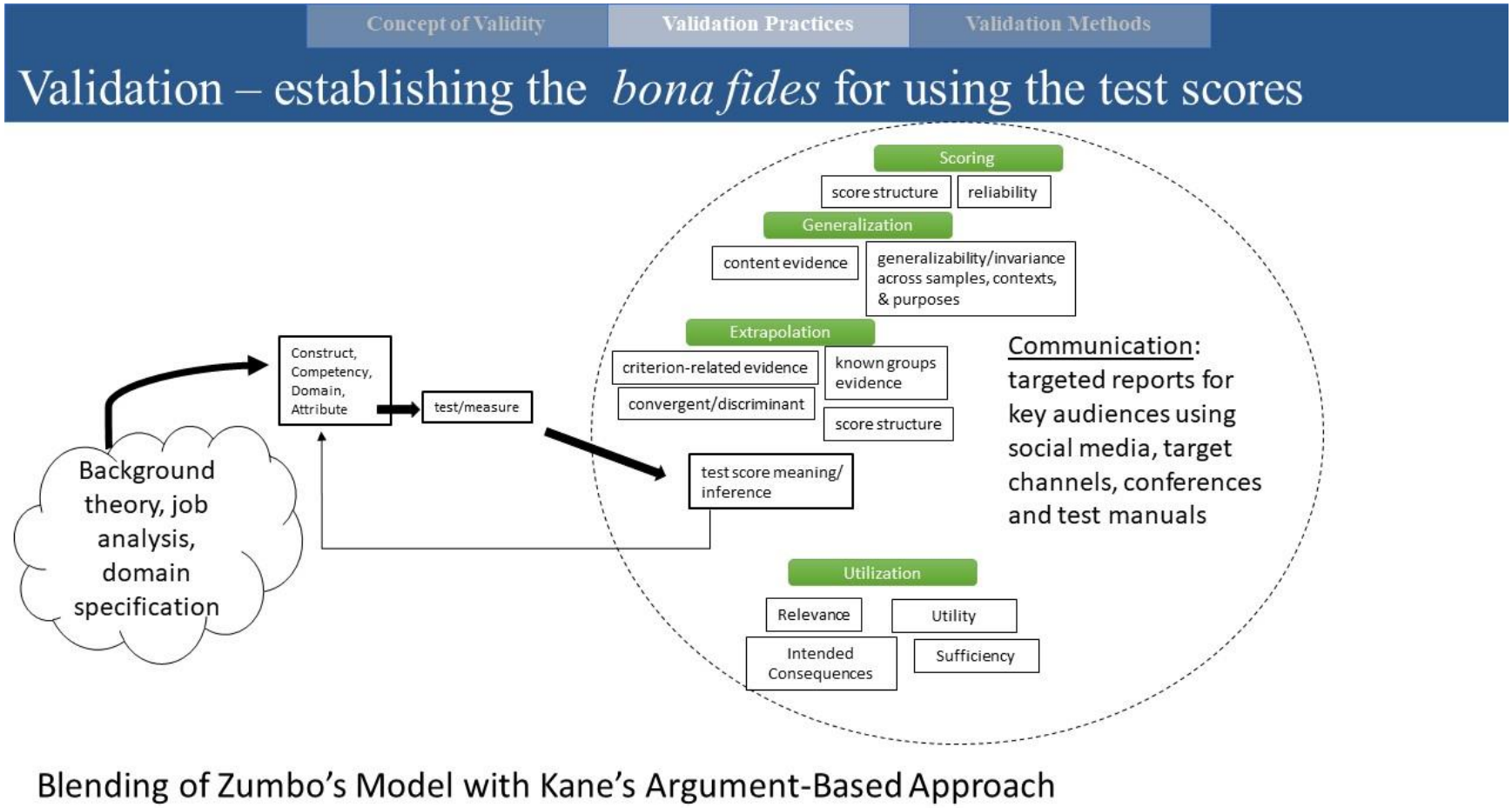
In terms of validation and test validity in business practices, in the testing industry, the ***bona fides*** approach we are introducing highlights that there is an implicit transaction, or social contract if you wish, among the test developer, test user, and test takers. We wish to highlight that test validation practices are meant to provide empirical evidence (and an argument/rationale) that the transaction is in good faith—hence transparent evidentiary trail and rationale for test use and interpretation as well as entering into this social contract with honesty, authenticity, and acting without the intention of ignoring potential hidden sources of invalidity in certain test use embodied in construct validity.

It should be noted that many authors refer to construct validity as the most important characteristic of a test but is seldom defined. A clear statement of what a construct is and the logic of construct validation was presented by Cronbach and Meehl (1955). These authors wrote:

> "*A construct is some postulated attribute of people, assumed to be reflected in test performance. In test validation the attribute about which we make statements in interpreting a test is a construct. We expect a person at any time to possess or not possess a qualitative attribute (amnesia) or structure, or to possess some degree of a quantitative attribute (cheerfulness). … Persons who possess this attribute will, in situation X, act in manner Y (with a stated probability). The logic of construct validation is invoked whether the construct is highly systematized or loose, used in ramified theory or in a few simple propositions, used in absolute propositions or probability statements. We seek to specify how one is to defend a proposed interpretation of a test . . . ."* (p. 247)

In short, a test is valid for a construct when it produces results that can be interpreted in terms of the construct definition under consideration. What has caused some confusion in testing is that tests that are construct valid provide information about (i) the test taker in terms of the construct (e.g., they are highly proficient in spoken English) and (ii) how the *construct definition* itself can be strengthened or extended. Distinguishing these two types of information and recognizing the importance of the second type is highlighted in the *Blended Zumbo and Kane Argument Based Approach* to validation introduced and adopted in this report, depicted in *Figure 2.2*, wherein there is reciprocal feedback information from the "test score meaning/inference" to "Construct, Competency, Domain, Attribute." Consistent with Zumbo's (2007, 2009) description of validity and validation as an integrative cognitive judgment involving a form of contextualized and pragmatic best explanation, our re-envisioning of contemporary argument based approaches to validation draws attention to the point that the practice of validation will (should) inform the construct, competency or attribute we posit to be measuring. As van Fraasen (2008) highlighted in his study of the history and philosophy of measurement, **the theory of the phenomenon and its measurement cannot be answered independently of each other, and that they co-evolve**.

Figure 2.2 A Schematic Depicting Establishing the Bona Fides for Using the Test Scores – Blending Zumbo's Model and Kane's Argument-Based Approach



Blending of Zumbo's Model with Kane's Argument-Based Approach

Our final observation on our validity theory is that Kane's (2016) notions of generalisability and invariance are reminiscent of Zumbo's (2007a) Draper-Lindley-De Finetti (DLD) framework and the bounds on the inferences therefrom. Zumbo's DLD framework's advantage is that it introduces a formalism based on Bayesian reasoning that focuses our research methodology on the interchangeability (or, more formally, exchangeability[1]) of the scores from two test versions of test administration—online versus test centre. Moreover, unlike other widely used psychometric frameworks, Zumbo's DLD brings to the forefront a similar exchangeability question about test takers. Based on these two forms of exchangeability, the kinds of claims made from alternate test scores arising from test centres or online test administration will shape the bounds on the inferences made from those test scores.

In the next section, a version of Zumbo's DLD will be described to aid in assembling the validity arguments and research methodology to investigate the validity of inferences from online testing.

2.2 The Theoretical Framework That Forms the Principles and Logic of Our Methodology to Investigate Online Testing

In a series of invited addresses, papers, and book chapters over the last 20 years, Zumbo has developed the Draper-Lindley-De Finetti (DLD) framework. (see, for example, Zumbo, 2001, 2002, 2007a, 2013, 2016; Kroc & Zumbo, 2020; Shear & Zumbo, 2013; Zimmerman & Zumbo, 2001). Building on Draper's (1995), Lindley's (1972), and de Finetti's (1974-1975) Bayesian predictive approach to inference, Zumbo's DLD framework highlight the necessity to be explicit about the sorts of inferences one makes, and that one can make from a test design and implementation.

At the heart of Zumbo's DLD framework is de Finetti's notion of 'exchangeability' to describe a certain sense in which, for example, test scores treated as random variables in a probability specification are thought to be similar. Zumbo (2007a) points to a definition of exchangeability in the setting of testing.

> Definition: A set of *n* units is termed *exchangeable* in the universe of test scores denoted *Y* if the joint probability distribution $p(Y_1,\ldots,Y_n)$ is invariant under the units' permutations denoted by the subscript indices. Additional *k*

[1] We use the term 'exchangeable' rather than 'parallel' throughout this line of research following developments by Louis Guttman (1945, 1953a, 1953b) who rejected the notion of parallel tests because of practical and statistical theoretical difficulties associated with the concept. Later, Novick (1966), Lord and Novick (1968), Kroc and Zumbo (2020), and others showed that parallel measurements can be abstractly defined and integrated with the rest of psychometric theory in an unambiguous way using advances in statistical theory. Zimmerman and Zumbo (2001), Zumbo (2007), showed mathematically that in fact, the concept is essentially the same as exchangeable random variables in mathematical formulations of probability theory. See, for example, De Finetti's (1975) introduction of this notion and other texts in probability theory (Loève, 1963; Rényi, 1970).

units are exchangeable in *Y* with the set if all (*n* + *k*) test scores are so exchangeable.

The definition of exchangeability can be interpreted that test score(s) as a scalar quantity or vector are similar and hence, in a mechanical sense, exchangeable or interchangeable. There are other definitions of exchangeability in probability and statistics, but the form given here is adequate for the present report's psychometric applications. Our objective is to unpack and analyze the logic of judgments of exchangeability (or similarity, or homogeneity) to clarify the roles of context and psychometric analysis in determining if administration mode (test centre and online) affects these tests' high-stakes decisions.

This definition invokes the concept of a 'unit' similar to De Finetti's more general term 'event' to denote any outcome from testing where either a quantity or vector that has been or can be observed (Zumbo, 2007a). As Zumbo and Kroc (2019) and Kroc and Zumbo (2020) describe in detail, a measurement is a choice, and the judgements involved in defining units and measurements rely on test use and context logically precede judgments of exchangeability.

Using the definition of exchangeability in our setting, we define the 'units' according to the following conditions.

- Zimmerman and Zumbo (2001) introduce a measure-theoretic (Hilbert space) approach to test data. Their approach can be extended for our purposes such that test data can be characterized as the realization of a stochastic event defined on a product space $\Omega = \Omega_I \times \Omega_J \times \Omega_K$ where the orthogonal components, Ω_I, Ω_J, and Ω_K, are the probability spaces for items, examinees, and test settings (test centre or online), respectively. We will limit our discussion to the three components, but it should be noted that the joint product space can be expanded to include other spaces induced by raters or measurement occasions for repeat testers.
- A set of *n* language test scores is termed *exchangeable* in the universe of test scores, *Y* if the joint probability distribution $p(Y_1,\ldots,Y_n)$ is invariant under the units' permutations denoted by the subscript indices. An additional *k* unit(s) is exchangeable in *Y* with the set if all (*n* + *k*) units are so exchangeable.
- In language assessment, we often deal with a profile of test scores reflecting listening, reading, writing, and speaking, so we will use a vector to denote a multidimensional observation on a single test taker unit.
- Keeping the three-component product space in mind, the interpretation of this test data involving online testing ***minimally*** requires a judgement of exchangeability (similarity or homogeneity) of a vector of language testing component scores, examinees, and test settings, as well as the specification of a stochastic process that is supposed to have generated the data (Zimmerman & Zumbo, 2001).
- Next, one needs to decide which of the online testing scenarios described in *Section 1.1* involving either (i) different tests administered online that substitute

for the original or (ii) an online version of an existing test. This decision necessitates a judgement jointly by the test user and psychometric researcher. How the test is used in high-stakes decision-making will lead to the appropriate descriptions of the necessary exchangeability for each of the three orthogonal components characterizing the stochastic event.

Using the mathematical notion of exchangeability, Zumbo's original focus was on sources and problems of measurement error and sampling problems regarding a domain of items or a target universe of test takers. As Zumbo (2007a) notes, his DLD framework is a natural extension of ideas in the 1940s and 1950s by Louis Guttman, who wrote about generalizing from measures (or tasks) that we have created to more tasks of the same kind with a kind of Bayesian thinking about inferences. Together there is also a tidy connection to the methodology of measurement invariance discussed in more detail below.

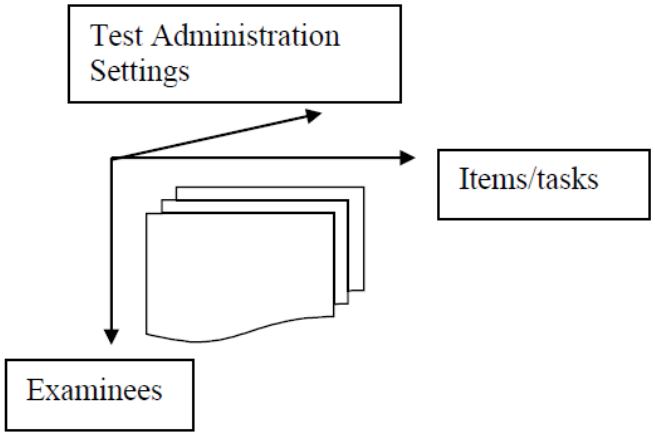
In summary, the concordance of the test centre and online test scores for either case described in *Section 1.1* (either different test administered online that substitute for the original or an online version of an existing test) is concerned with combining information from different observational units and making inferences from the resulting test data to prospective measurements on the same or other units. These psychometric operations will be useful only when combined units are judged to be concordant (comparable or homogeneous).

Using Zumbo's DLD framework to investigate the test validity of online and test centre testing will result in a description of the:

a) kind and level of exchangeability (similarity or homogeneity) that each invokes,
b) validity, and more importantly, the sources of hidden invalidity, of score use of language tests, administered remotely for high-stakes decisions, and
c) kinds of evidence and subsequent research methodology that supports valid test use.

Keeping the three-component product space in mind, interpreting test data, and investigating the validity using online (remote) test administration, one can depict the space of exchangeability as a three-dimensional variant of Zumbo's DLD. A third dimension denoted 'test administration settings' reflects the two test modes administration as seen in *Figure 2.3* below. Therefore, one can have any degree of exchangeability and resultant evidential strength supporting the validity of the claims as a point in the three-space depicted in *Figure 2.3*.

Figure 2.3 A Three-Dimensional Variant of Zumbo's DLD Framework



The three-dimensional representation can be greatly simplified by collapsing the items and settings dimensions, see *Figure 2.4*. Therefore, in our setting wherein we investigate if remote online testing is concordant with test scores from test centres, it can be addressed with a question of the *exchangeability* of items or tasks and the *exchangeability* of a target universe of test takers online and test takers at test centres.

In the simplified revised DLD framework depicted in *Figure 2.4*, the horizontal axis reflects the degree of exchangeability of the test centre and online test scores. In contrast, the vertical axis reflects the test takers' degree of exchangeability of those test takers who choose the test centre or remote online test administration in the target population of test takers. *Figure 2.4* depicts the various types of inference graphically in a two-dimensional space using Zumbo's (2007a) terms for the various forms of inference (i.e., calibrative, specific sampling, specific measurement, and general measurement).

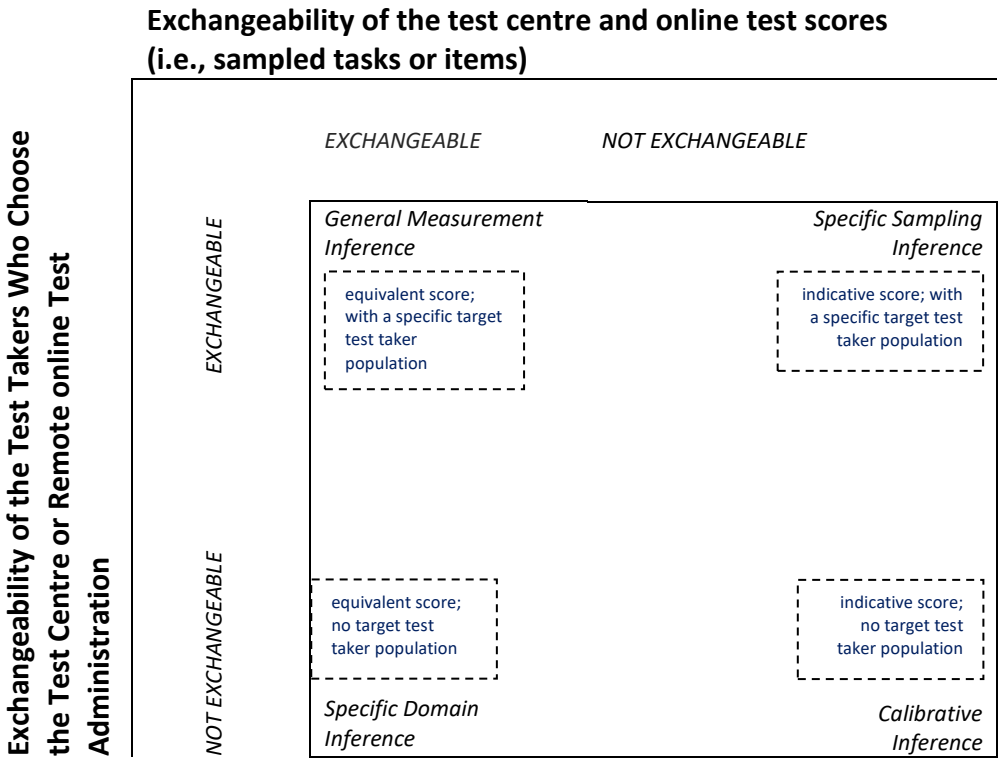
Please note that the kinds of tests and inferences in *Figure 2.4* could be placed anywhere in the rectangle, reflecting the degree of exchangeability on either dimension. Here one will see that at the four corners of the quadrants, starting at the bottom right corner and going counterclockwise, one has what Zumbo (2007a) describes as either calibrative, specific sampling, general measurement, or specific domain inference.

As further clarification, as Zumbo (2007a) notes, there is an implied continuum of inferential strength depicted at the four quadrants' extremes in *Figure 2.4* wherein in terms of inferential strength, both initial calibrative and specific sampling are less strong than specific domain, which in turn is less strong than general measurement inference. General measurement inference is the strongest. Please note that in this setting, 'strength of inference' is being used in a logical-mathematical sense rather than a common language usage. As such, it is important to note that it is not that some inference is necessarily better than others (because this sort of value judgment needs to take the purpose of the testing into account), but rather that credible and defensible testing practices require one to be explicit about the sorts of inferences that are made

and that can be made in a given context. Doing so helps lay bare the hidden sources of invalidity in certain test use embodied in construct validity.

Two points are noteworthy. First, as the DLD framework highlights, construct validity with remote testing needs to be considered in light of the test takers' exchangeability who choose the test centre or remote online test administration, and the test score results from the test centre and remotely administered tests. Second, it is important to keep in mind that all four of these degrees of validity (based on inferential) claims require evidence to support their use of test scores resulting from remote test administration.

Figure 2.4 The various forms of measurement inference of remote online testing alternatives - Zumbo's DLD Framework.



Note: Adapted from Zumbo, B. D. (2007a). Validity: Foundational issues and statistical methodology. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics* (Vol. 26: Psychometrics, pp. 45-79). Amsterdam: Elsevier Science.

It is instructive to contrast four alternatives depicted in dashed-boxes in *Figure 2.4*. Let us imagine an English language proficiency test for students planning to study in a Canadian post-secondary institution. Imagine that an ongoing language test is offered at a test centre with a well-specified test blueprint and design as well as the validity and reliability evidence. Finally, imagine that the test provider offers the prospective test taker a choice

of a test administered online using a remote-proctored test-delivery platform or at a test centre.

The focus throughout the description that follows is on the test administered online.

1) The bottom right corner of *Figure 2.4* represents low degrees of exchangeability of test scores and test takers.
    - An online test that is designed as an indicator test with no intention (or evidence) of exchangeability with its counterpart administered in a test centre.
    - The test provider has no reason to believe that the online test takers are exchangeable with those taking the test at a test centre.
    - Suppose this indicator test is well designed but not exchangeable with the online version. In that case, the test is limited to a type of calibrative inference because there is no evidence of the exchangeability of either the online tests or test takers with their counterparts at a test centre. This results in a different test with a different testing population than the well-established test centre version and is of limited to no value for test users as a substitute for the well-establish test at a test centre.
2) The top right corner of *Figure 2.4* represents a low degree exchangeability of test scores but a high degree of exchangeability of test takers.
    - An online test that is designed as an indicator test with no intention (or evidence) of exchangeability with its counterpart administered in a test centre.
    - However, the test provider has good reason to believe that the online test takers are exchangeable with those taking the test at a test centre.
    - In this case, the test is limited to specific sampling inference because there is evidence to support that the online test takers are exchangeable with those taking the test at a test centre. Suppose the indicator test is well designed but not exchangeable with the online version; test users may find that this is a worthwhile stopgap to get them by until test centres reopen on a wide scale. The online and test centre test takers' exchangeability may bootstrap the inferences made for a well-designed indicator test. However, this test would be treated as an independent alternative unrelated to the test centre test.
3) The bottom left corner of *Figure 2.4* represents a high degree of exchangeability of test scores but a low degree of exchangeability of test takers.
    - An online test designed as an alternative mode of test delivery with the same test format, content, and reporting scale as the test delivered at test centres and intended to be used as an alternate mode of test delivery alongside the test delivered at a test centre. The test provider has good reason to believe that the online test is exchangeable with its counterpart administered in a test centre.
    - However, the test provider has no reason to believe that the online test takers are exchangeable with those taking the test at a test centre.

- In this case, the test is limited to specific domain inference because there is no evidence to support that the online test takers are exchangeable with those taking the test at a test centre. The interchangeability of the online and test centre test versions of the test may bootstrap the inferences made for a well-designed test. However, much like the case of specific sampling inference, this test would be treated as an independent alternative to the test centre test because of the lack of test takers' exchangeability that may threaten the validity of the standard-setting and cut-scores or item response theory equating or calibration- may need to be calibrated separately with a different cut-score.

4) The top left corner of *Figure 2.4* represents a high degree of exchangeability of test scores and a high degree of exchangeability of test takers.
   - An online test designed as an alternative mode of test delivery with the same test format, content, and reporting scale as the test delivered at test centres and intended to be used as an alternate mode of test delivery alongside the test delivered at a test centre. The test provider has good reason to believe that the online test is exchangeable with its counterpart administered in a test centre.
   - The test provider has good reason to believe that the online test takers are exchangeable with those taking the test at the test centre.
   - In this case, the test allows for general measurement inference because the evidence supports the test scores' exchangeability and the online test takers' exchangeability with those taking the test at a test centre. In this case, the test scores from the online and test centre modes of administration are fully interchangeable.

The objective of this classification and ordering of inferences in the DLD framework is to encourage test researchers, providers, and users to be explicit about the types of inferences they can make. The additional focus is on the range of possible conditions under which concordance (invariance) is expected to hold. It depends, then, on the type (or strength) of inferences one wants to draw. Psychometric researchers need to be explicit about the information they have about the level of exchangeability of individuals and tasks or items used in validation studies of tests administered online using a remote-proctored test-delivery platform. This explicitness will go a long way toward creating credible scientific measurement evidence leading to *bona fide* claims of test validity.

For example, it is sometimes implied with indicator tests that given that the indicator test is derived from the full test offered at a test centre, one does not have to worry about validating the exchangeability (or concordance) of an indicator test and the full test used at a test centre; relying on a kind of face validity. Clearly, from the above framework, one is still making inferences from "indicative scores" to the full test centre test (of which the indicative score is a gauge or indicator) because one still does not have the individual completing the full test, per se. The indicative score is, by design, easier to attain than the full test centre test during a pandemic or time when access to test centres is limited,

hence shortening the inferential distance from the indicative score to the full target test. Importantly, an inference is still being made from those indicative scores to other full tests like them. ***One should be cautious not to let the DLD framework be interpreted in a weak form, hence allowing the same slippage as is noted above with weak forms of construct validity***.

2.2.1 Principles, Criteria, and Evidence Emerging from Zumbo's DLD Framework

Four methodological principles emerge from Zumbo's DLD theoretical framework that forms the basis and logic to investigate the concordance of the test centre and online testing. *Table 2.1* below describes the four principles and their corresponding implications for criteria and evidence of concordance.

A few points are noteworthy. First, the equity principle is defined differently for the cases of an online version of an existing test than an indicator test that substitutes for the test centre version. In line with Zumbo's DLD framework, the former implies a stricter principle of indifference. The latter is a weaker form of no hindrance or disadvantage for the test taker having to take the indicator test. Of course, the former implies the latter.

Table 2.1 A Description of the Four Principles and their Implications for an Online Version of an Existing Test or an Indicator Test that Substitutes for the Original. (continued on next pages)

Principle	**An online version of an existing test**	**Online version is a different test that substitutes for the original**
The Equity Principle	It should be a matter of indifference to a test taker or a test user as to which of two modes of test administration (test centre or online) the test taker chooses to take the test.	A test taker should not be hindered or disadvantaged by having to take the indicator test.
Implications of the Equity Principle	This principle implies the interchangeability of the test scores across the two modes of administration. Evidence and Criteria: i. The online and test centre versions should be built into a common blueprint and be designed to	As a substitute for the original full test (which is completed at a test centre), this principle implies the following evidence and criteria: i. In practice, an indicator test is an analogue of the full test.

Principle	An online version of an existing test	Online version is a different test that substitutes for the original
	measure the same constructs. Nearly everything about the tests themselves is the same. ii. The online and test centre tests should both be measures of the same language skills and for the same use.	ii. There is a correspondence in kind or quality between the members of pairs or sets of forms of test administration (test centre and remote) that serve as a basis for creating the indicator test. That is, to serve as a suitable substitute, the indicator test is like the full test.
The Test Use Principle	As Zumbo and Rupp (2004) note, scoring test data eventually brings about consequences for test-takers, and these consequences are mathematically dependent upon accurately estimating the uncertainty associated with test takers' scores. This is most crucial for test-takers with an observed test score somewhere around the cut-score but less so for any test taker. Therefore comparison across test administrations to establish the level of concordance should be established on the scale on which scores are reported- for example, band scores on each of the (four) components of a language test.	
Implications of The Test Use Principle	There should be a **high degree of interchangeability** (exchangeability) between the reported score from the test centre and online tests. This could also be described more formally as a bijection or bijective function between the sets representing the reported scores from the test centre and online tests. Evidence and Criteria: The two tests should have (i) similar levels of classification consistency or other relevant	There should be an **order-preserving correspondence** between the reported scores from the indicator and the full tests. This could also be described more formally as a monotonic function (or monotone function) that preserves the given order between the reported test scores from the indicator and the full tests (these are ordered sets). Evidence and Criteria: If relying on specific sampling inference to justify using the

Principle	An online version of an existing test	Online version is a different test that substitutes for the original
	forms of measurement uncertainty, (ii) similar joint distribution (e.g., covariance matrix) of the scores on the test components, (iii) meaningful sub-groups of test takers defined by gender or first language should have similar test scores. In addition, as implied by the DLD, there should be a homogeneity or equivalence for (sub) populations of online test takers and those at a test centre.	indicator test, the correspondence of the reported test scores of online and test centres should be the same regardless of the choice of (sub) population from which it is derived. As implied by the DLD, a homogeneity of online test takers and those at a test centre may bootstrap the claims made from the online test.
<u>The Matching Principle</u>	The equity and test use principles do not, as stated above, require a formal definition of concordance, but they hint at the matching principle. The formal definition of concordance and statistical psychometric methods emerged in the early 1980s based on the matching principle (Angoff, 1993). Concordance is synonymous with test bias, with a particular focus on comparing different modes of test administration. Matching Principle: Working from Millsap's (2011) description of the more general concept of item bias, let us imagine any two test takers from different groups reflecting two modes of test administration who are identical on the attribute(s) being measured by the test item or component. • The item or component is concordant in relation to the different test administration groups if the probability of attaining any particular score on the item is the same for the two individuals. • Therefore, the score on an item or component that lacks concordance will depend not only on the attribute being measured but also on the individual's mode of administration under consideration. The matching idea enters into this definition in the requirement that the two individuals be identical or matched on the attribute(s) being measured.	

Principle	An online version of an existing test	Online version is a different test that substitutes for the original
Implications of The Matching Principle	Angoff's (1993) matching principle allows for the definition of optimal statistical psychometric methods that do not confound concordance with true differences in item performance (i.e., impact). The matching principle disentangles the lack of concordance from true differences in item performance. • As Millsap, Angoff, Mellenbergh (1989, 1994), and others state, it is essential that we compare individuals from different groups who are matched rather than randomly chosen pairs of individuals. • Comparing mean test scores, item or component scores, or pass rates for two groups is considered uninterpretable and flawed psychometric evidence because they confound concordance with true differences in item performance. The same goes for simply comparing group psychometric properties such as test reliability for each group. • These unconditional approaches were replaced over three decades ago by statistical methods to investigate concordance based on the matching principle. • As Zumbo and Hubley (2003) note, there are at least three statistical frameworks for investigating concordance that has evolved in the research literature: (1) modeling item responses via contingency tables or regression models, (2) item response theory, and (3) multidimensional models. Each framework provides useful logic and organizing principles for describing the lack of concordance and developing methods for detecting it. It is useful to discuss concordance in relation to the concept of measurement invariance of item response theory. Measurement invariance indicates that the expected item or component scores, denoted *Y*, given the true score or latent variable, *T* computed in subpopulation $G = g$ are equal to those computed in the total group, $E(Y \mid T, G = g) = E(Y \mid T)$, and $E(Y \mid X, G = g) = E(Y \mid X)$, where X denotes a proxy for the true score or latent variable such as the total observed test score, item-corrected observed test score, or a vector of multiple covariates, and $g = 1, 2$ denoting groups of test takers who complete the test either at a test centre or online. Another way to think about the mathematical expression above is that item or component test scores can be predicted from a latent variable (or observed score proxy), the invariance of which can be evaluated as	

Principle	An online version of an existing test	Online version is a different test that substitutes for the original
	questions about the differential item and test functioning (Shealy & Stout, 1993), in other words, concordance. The lack of concordance implies that for individuals from this population of test takers, the conditional distribution of observed scores on *Y* at some T (or X) value will be different from *Y*'s conditional distribution for individuals from the other populations. With this statistical description in hand, **Zumbo and Hubley's three statistical frameworks for investigating concordance, the three frameworks operationalize the matching principle differently**. 1. The modeling item responses via contingency tables or regression models framework, in essence, consists of conditional methods that study the effect of the grouping variable(s) and the interaction term(s) over-and-above (i.e., while conditioning on) the total score. 2. The item response theory framework considers two item or component trace lines of the same item or component but computed from the two groups. Lack of concordance is exhibited when these trace functions are identifiably different for the groups. In its essence, the IRT approach is focused on determining the area between the trace lines of the two groups. Unlike the contingency table or regression modeling methods, the IRT approach does not match the groups by conditioning on the total score wherein one computes the difference function between the groups conditionally. Comparing the trace lines operationalizes the matching principle by 'integrating out' the matching continuum of variation (the latent variable) in the sense that one computes the area between the trace lines across the distribution of the continuum of variation, the latent variable. 3. As implemented in Stout's SIBTEST, the multidimensional model framework similarly operationalizes the matching principle as the first framework, modeling item responses via contingency tables or regression models by conditioning on matching variable(s).	
<u>Multimethod Principle</u>	There is an overall principle of multiple sources of evidence (multimethod methodology) that calls for more than one source of evidence.	

Principle	An online version of an existing test	Online version is a different test that substitutes for the original
Implications of the Multimethod Principle	The multiple sources of evidence are intended to support the concordance investigation and allow a researcher to rule out rival plausible alternative interpretations and ferret out multiple sources of evidence potentially hidden invalidity.	

In the next section of this report, we describe a brief history of concordance methodology that sets the stage for a formal definition of concordance that implies a statistical framework of psychometrics for concordance studies.

2.2.2 Toward a Psychometrics For Studies of Concordance - A Brief History of Concordance Methods Leading to a Definition of Concordance

A Brief History of Concordance Studies That Sets the Stage for a Formal Definition

As Zumbo (2007b) notes, concerns about concordance emerged under the rubric of "item bias" within the context of test bias and high-stakes decision-making involving achievement, aptitude, certification, and licensure tests in which matters of fairness and equity were paramount. Historically, concerns about test bias have centred around differential performance by groups based on gender or race. If the average test scores for such groups (e.g. men vs. women, Blacks vs. Whites) were found to be different, then the question arose as to whether the difference reflected bias in the test. Given that a test is comprised of items, questions soon emerged about which specific items might be the source of such bias.

Chronologically, concordance research developed from item analysis methods without an explicit definition of the concept of concordance. There is value in letting "a thousand flowers bloom" because, as we learn from contemporary philosophies of science, the idea that we should begin all scientific inquiry with an adequate definition of all key terms is often inappropriate. However, as the use of concordance studies expanded and began to take a role in test validity, the lack of explicit definitions began to impede the development and implementation of more rigorous testing for concordance, where and when appropriate. More specifically, vague conceptualization and theorizing of this nature stood in the way of developing logical and principled methods, resulting in continual calls for better psychometric statistical methods without any real progress or frame of reference to judge their suitability and performance.

The statistical methods to investigate concordance can be divided into unconditional and conditional ones (Mellenbergh, 1982, 1983). Examples of unconditional methods to investigate concordance relied on statistical analyses that focused on the percentage of

examinees responding correctly, or the average score, to each item. The implicit definition of concordance in unconditional methods is based on group-by-item interaction. Items may differ in difficulty, and groups may differ in ability to solve the item correctly, but, importantly, that does not indicate item bias. As Mellengbergh notes, item bias is conceived as an interaction of sorts; the difference in item difficulty or passing rates between groups is not constant for all items and items that deviate from the general trend are considered to be biased. This unconditional approach has been criticized since the late 1970s (as simply an omnibus test of item score differences, which confound concordance with true differences in item performance (i.e., impact). These flawed early unconditional approaches compared average performance for the two groups to determine if outcomes from two groups of test takers (e.g., in our setting, those taking the test with live remote proctoring and test center) were comparable.

Zumbo and Hubley (2003) noted that the conditional methods the concept of concordance are developed based on a formal definition are based on differences in item difficulty between groups given the level of ability. An item is considered not concordant when it differs in difficulty between subjects of identical ability from different groups. The methods are based on item response theory or its approximations. From conceptual and statistical mathematics points of view, conditional methods must be preferred above unconditional methods because they are founded on a formal definition of concordance.

As Zumbo (2007b) noted, due to the highly politicized environment in which concordance was initially examined, two inter-related changes occurred. First, the expression 'item bias' was replaced by the more palatable term 'differential item functioning' or DIF in many descriptions. DIF was the statistical term used to describe the situation in which persons from one group answered an item correctly more often than equally knowledgeable persons from another group. Second, introducing the term 'differential item functioning' allowed one to distinguish item impact from item bias. Item impact described the situation in which DIF exists because there were true differences between the groups in the underlying ability of interest being measured by the item. Zumbo (2007b) described item bias as situations in which there is DIF because of some characteristic of the test item or testing context that is not relevant to the underlying ability of interest (hence the test purpose).

Since the early 1980s, comparing average test scores, average item scores, or average pass rates for the two groups are considered uninterpretable and flawed psychometric evidence because they confound concordance with true differences in item performance (i.e., impact).
The same goes for simply comparing group psychometric properties such as test reliability for each group.
These unconditional approaches have been replaced over three decades ago by a formal definition of, and statistical methods to investigate concordance based on the matching principle (see, for example, Angoff, 1993).

Formal Psychometric Definition of Concordance

It may be helpful to think about the worst-case scenario, a lack of concordance, to get a sense of the concept of concordance. Let us take the *CAEL* as an example to help motivate the problem. First, let us imagine we have two groups of test takers, one having completed the *CAEL* at a test centre and the other online.

Keeping with *CAEL* as an example, lack of concordance for a language component (i.e., reading, listening, writing, or speaking) is defined to occur when the probability of a language component score, for test takers with the same intended-to-be-measured language ability, differs because of the setting in which they took the test. It is noteworthy that, as is appropriate, this definition is silent about whether the score distributions of the two groups of test takers are identical or are stochastically ordered (Stout, 2002). This means whether one takes the *CAEL* at a test centre or online impacts a test taker's eventual test score.

To draw these sorts of conclusions, one needs to balance the two groups of test takers regarding their intended-to-be-measured language ability and important related variables that affect test performance, like motivation. There are a few research designs or statistical techniques in our data science toolkit to balance the two groups so that they are interchangeable, if you wish, except for which mode of test administration. However, setting aside how we balance the two test taker groups, the key thing is that they need to be balanced to make sense of the results.

The absence of concordance, also known as differential item functioning, has been studied extensively in latent variable models such as confirmatory factor analysis and item response theory and observed score models involving contingency tables or generalized linear models.

In the psychometric research literature, concordance is defined with respect to a grouping or selection variable, S, such as mode of test administration (online or test

centre), and concerns the measurement model relating observed scores to underlying latent variables.

The measurement model has been treated as the same for all groups in the sense that the probability of observing a given item score is equal for members of different groups who have the same score on the underlying latent variable.

It is widely seen in the psychometric research literature (e.g., Lubke, Dolan, & Neale, 2004) that, more formally, concordance (characterized as measurement invariance) has been defined as

$$f\left(Y|\eta,s\right)=f\left(Y|\eta\right),$$

where observed variables are denoted as Y, latent variables as η, and the grouping variable as S. A situation where measurement invariance is absent, that is,

$$f\left(Y|\eta,s\right)\neq f\left(Y|\eta\right),$$

an observed variable *Y* is non-invariant with respect to a grouping variable *S* if the observed score depends not only on the latent variables η but also on *S*, or variable(s) related to *S*.

Following the seminal work of Mellenbergh (1989) and Meredith (1993), there are three different types of effects of S or variable(s) related to S that may or may not occur simultaneously:

- Constant for all possible scores on η, which results in a group difference in the intercept of the regression of Y on η.
- The effect can increase or decrease as a function of η, resulting in a group difference with respect to the steepness of the regression.
- The regression curves (or non-linear regression) on η are equal across groups, but the regression residuals differ.

2.3 Research Design Options and Decisions Implied From the Matching Principle

There is an agreement in the psychometric research literature that comparing unmatched (unconditional) mean test scores, item or component scores, or pass rates for two groups is flawed psychometric evidence because they confound concordance with true differences in test performance. There is less agreement in the research literature about the alternative research designs to investigate concordance that operationalizes the matching principle.

Generally speaking, there are three study design options (i.e., research designs) for concordance studies that operationalize the matching principle. This section will briefly describe the between-subjects, within-subjects, and matched-subjects options and provide guidance on design choice.

As described below, the between-subjects and matched-subjects designs are the most commonly used concordance and DIF studies options, with the former being most widely used and discussed.

2.3.1 Basic Concepts and Terminology

Rewriting the matching principle with research design in mind, let us imagine two test delivery settings: test delivered at a test centre and the other wherein the test is delivered online. *Table 2.2* depicts the three study design options. The matching idea enters into the definition of concordance in the requirement that test takers be identical or matched on the attribute(s) being measured prior to comparing their item or component test scores. The notation and design features in *Table 2.2* will be described when each option is discussed.

Table 2.2. Depiction of the Three Concordance Study Design Options

Research Design	***Test Delivered at a Test Centre***	***Test Delivered Online***
Between Subjects	Y_A	Y_B
Within Subjects	Y_A	Y_A
Matched Subjects	Y_A	$Y_{(B \approx A)}$

2.3.2 Option #1: Between Subjects Design

The portrayal of the between-subjects design in *Table 2.2* is displayed below. In the between-subjects design, the matching principle states that for any two test takers from the two test delivery settings (denoted Y_A and Y_B) who are identical on the attribute(s) being measured by the test item or component, the item or component is concordant if the probability of attaining any particular score on the item is the same for the two individuals. The between-subjects design uses a separate test taker sample, denoted A and B below, for each test delivery setting, and differences are then measured between groups.

Research Design	***Test Delivered at a Test Centre***	***Test Delivered Online***
Between Subjects	Y_A	Y_B

- As described in the "Implications of The Matching Principle" sub-section of *Table 2.1*, it is evident that concordance studies have been conceived with an implicit between-subjects design without randomization from its historic beginnings. It is this non-randomized between-subjects design that called for the matching principle (Angoff, 1993).

- This implicit between-subjects setting partly reflects the field of psychometrics' apparent historical aversion to randomized experiments, but far more likely that DIF and test bias studies are more generally, of which a concordance study is an instantiation, are conducted after the fact in live operational testing settings.
    - The advantage of this study design is that test takers are completing the test in an *in vivo* setting with maximal motivation and test performance.
- As described in *Sections 2.1 and 2.2*, the matching principle is operationalized in between subjects study designs of concordance by some combination of covarying (or conditioning on) matching variable(s) using variants of analysis of multi-way tables or generalized linear models, or by integrating out the test score distribution using latent variable models such as IRT or confirmatory or exploratory multi-group factor analysis.
- In this sense, the analysis of between-subjects concordance study designs shares a lot in common with analysis of covariance (ANCOVA) or attribute-by-treatment interaction (ATI) methods. Building on this similarity, it is important to recognize that nearly all concordance (DIF) methods are applied in what would be called an observational or quasi-experimental between-subjects study design. So one must keep in mind all of the commonly known caveats around making causal claims of grouping variable effects in observational studies involving intact groups.

2.3.3 Option #2: Within Subjects Design

The portrayal of the between-subjects design in *Table 2.2* is displayed below. In the within-subjects design, the matching principle states that for any test taker (denoted Y_A) completes the test in both test delivery settings, the item or component is concordant if the probability of attaining any particular score on the item is the same in both settings. The within-subjects design uses one sample of test takers that complete the test in both test delivery settings. Differences are then measured by comparing the same test taker's scores observed under both the test delivery settings.

Research Design	***Test Delivered at a Test Centre***	***Test Delivered Online***
Within Subjects	Y_A	Y_A

- One can consider the within-subjects design as the ultimate in matching with the same person in both test administration settings.
- An advantage is that one needs far fewer test takers in the within-subjects design than between-groups or matched-groups designs for comparable statistical efficiency and power.
    - It is noteworthy that, in many cases, within-subjects designs are statistically more powerful because the individual differences are controlled by having the same respondent in both test delivery conditions.

This can be conceived as a reduction in the standard error of the comparison among means or proportions.

- Order of which testing condition may confound within-subject comparisons, but this can be mitigated by randomizing order (*viz* randomize whether a test taker first completes the test at a test centre or online).
- Heeding the cautions in *Section 2.2* describing the DLD framework, a possible limitation of withing-subjects design needs to be considered because test takers who are willing to sign on for two testing sessions may represent a distinct and non-exchangeable subset of all test takers for whom one wants to make an inference about concordance.
    - In the same vein, if a test taker knows their test result from their first test session, they may not be maximally motivated during the second.

2.3.4 Option #3: Matched Subjects Design

The portrayal of the matched-subjects design in *Table 2.2* is displayed below. In the matched-subjects design, the matching principle states that for any two test takers from the two test delivery settings (denoted Y_A and $Y_{(B \approx A)}$) who are statistically matched on the attribute(s) being measured by the test item or component, the item or component is concordant if the probability of attaining any particular score on the item is the same for the two individuals. The notation $Y_{(B \approx A)}$ is meant to convey that the test takers in sample B are adjusted (balanced) statistically to ideally be equivalent to those in sample A.

Research Design	***Test Delivered at a Test Centre***	***Test Delivered Online***
Matched Subjects	Y_A	$Y_{(B \approx A)}$

Advances in the statistical theory of matched-subjects designs (e.g., Holland, 1986; Rubin 1974, 1977, 1978) has led to the mathematical theory propensity score methods (Rosenbaum and Rubin 1983) that focus on "interventions" (such as the two modes of test delivery) the comparability of nonexperimental comparison groups commonly found in between-subjects concordance study designs in terms of "pre-intervention" variables.

The central statistical concept is the propensity score, which quantifies the probability of assignment to the test centre or online test delivery settings conditional on covariates. One can control for differences between the test delivery settings in the non-experimental comparison groups by the estimated propensity score, a single variable ranging [0,1]. In essence, using propensity score methods, one can balance the two groups and replicate the experimental treatment effect for a range of specifications and estimators.

The matched-subjects design is, in essence, a compromise between the between-subjects and within-subjects design, taking advantage of the strengths of each design by mitigating the concerns described in *Section 2.3.3* about:

(i) possible carry-over effects from the test order,

(ii) the concern that those test takers willing to sign on for two testing sessions may represent a distinct and non-exchangeable subset of all test takers for whom one wants to make an inference about concordance, and

(iii) possible confounding motivation effects based on the test order.

This matched samples approach offers an alternative to covarying or conditioning, creating two balanced (or equivalent) groups based on the observed covariates. In essence, one creates a matched group based on specific measured covariates.

Like the between-subjects design, the matched-subjects design uses two samples of test takers that complete the test in both test delivery settings. However, in the matched-subjects design, the two groups of test takers are pre-processed to balance the two groups on key variables before differences are measured by comparing the same test taker's observed under both the test delivery settings.

An assumption underlying the propensity score matching method is the ignorable treatment assignment assumption or selection on observables; see Holland (1986); Rubin (1974, 1977, 1978). This assumption translates to a condition in which the assignment to test delivery condition depends only on the observable pre-assignment variables. As Rubin has noted on several occasions, although this is a strong assumption, one can demonstrate that propensity score methods are an informative starting point because they quickly reveal the extent to which the testing conditions and comparison groups overlap in terms of pre-assignment variables.

Borrowing from the conceptual connections to DIF studies, the results of a series of studies suggest that the propensity score approach is a promising strategy for use in the design of concordance studies because it can be used for balancing pre-test differences between groups and achieving an effect akin to the random assignment if the key covariates are collected. Dorans and Holland (1993) suggested that propensity score matching may be a useful solution instead of directly matching multiple observed variables. Lee and Geisinger (2014) adopted propensity scores to control contextual sources when examining gender DIF. In a focused program of research on the use of propensity score methods in DIF studies, Zumbo and his colleagues:

- successfully adapted and refined propensity score methods for DIF analyses (Chen, Liu, & Zumbo, 2020; Liu, Zumbo, Gustafson, Huang, Kroc, & Wu, 2016), demonstrating their power and utility, and
- evaluated the performance of propensity score approaches for DIF analysis through Monte Carlo simulation methods to assess bias, mean square error, Type I error, and power under different levels of effect size and a variety of model

misspecification conditions, including different types and missing patterns of covariates (Liu, Kim, Wu, Gustafson, Kroc, & Zumbo, 2019).

2.3.5 Choice of Research Design – There is No Perfect Design; Rather, it is a Tradeoff of Advantages and Disadvantages

Ideally, in all three research design settings, the score on an item or component that lacks concordance will depend not only on the attribute being measured but also on the individual's mode of administration under consideration. In many situations, any of the three design options may be possible. Three interrelated guidelines should determine the trade-offs in the choice of the research design options

(1) considerations of the issues raised in *Sections 2.2 and 2.3* from Zumbo's DLD framework, and the logic and principles of concordance, respectively,
(2) the balance of priorities for internal and external validity,
   - internal validity – the degree to which the results are attributable to the mode of test administration and not some other rival explanation,
   - external validity – the extent to which the results of the concordance study can be generalized to test takers from the target population who did not participate in the concordance study, and
(3) the cost constraints for running the study.

The approach advocated herein is the use of multimethod strategies that allow one to mitigate the limitations of each design option.

Therefore, there is no one right way to conduct a concordance study; investigators make trade-offs and decide whether to use a within- or between-subjects design depending on the research circumstances, based on how and what they want to study.
The three design options can be categorized into two varieties of between-subjects designs and a within-subjects design in broad terms.

Research Design	**How the Design is Implemented**	**The Main Advantages and Disadvantages**
<u>Between-subjects</u> A. Observational or quasi-experimental between-subjects study design	• Concordance analyses are conducted with covariate matching or integrating out the latent variable ***based on test response in typical live operational testing settings***	• The main advantage of the between-subjects study design is that test takers are completing the test in an ***in vivo*** setting with ***maximal motivation and test performance***

Research Design	How the Design is Implemented	The Main Advantages and Disadvantages
B. [Matched-design] Observational or quasi-experimental between-subjects study design applying statistical matched-groups methods	Concordance analyses are conducted using a statistical matching method such as those based on propensity score analysis using available covariates ***based on test response in typical live operational testing settings***	• The main disadvantages are - that the two groups of test takers may be non-equivalent, and - that the effectiveness of the statistical analyses to adjust for this nonequivalence invoke an ideal setting that ***makes strong assumptions of the data and design***, such as having the correct covariates in (A) and the correct as well as a sufficient number of covariates in (B) to form a reliable propensity score.
<u>Within-Subjects design</u> <u>Within-Subjects Design</u> (continued	Concordance analyses are conducted after each test taker completes the test in both test delivery settings. There are fewer test takers, but they need to complete the test twice.	The major advantage is in ideal conditions when there are no confounding effects of order. A test taker completes the two test administrations or carry-over effects; ***this design is more statistically efficient and powerful because each test taker serves as their control***. - That is, within-subjects design minimize the random noise and prevents the introduction of any underlying systematic individual differences from contaminating the experimental findings a) The major ***disadvantage stems from the same source as the main advantage; a test taker takes both tests and therefore serves as their control***. - Therefore, the sample of those test-takers willing to sign on for two testing sessions may represent a distinct and non-exchangeable subset of all test takers for whom one

Research Design	How the Design is Implemented	The Main Advantages and Disadvantages
		wants to make an inference about concordance. - There are possible confounding motivation effects based on the test order because if a test taker does well on their first test-taking session, they may not be as motivated to do well on the second. - The within-subjects design does not reflect a typical test taking setting; therefore, an arrangement would have to be made with the test users about how they will resolve the fact that they have two test scores. - Also, it would need to be decided ahead of time if test takers are informed of their performance after the first session. As such, this may heighten the concern about the eventual sample being comprised of a non-exchangeable sample compared to the target test taker population.

2.4 Transitioning to Part II – A Concordance Study of the *CAEL* Online

To this point in the report, we have achieved the first of our three objectives: to ***describe a theoretical framework*** for unpacking the assumptions that support the validity of claims made from alternate online tests, whether they are an indicator or otherwise. The theoretical framework informs the multimethod research methodology and connects the research to the validity of the inferences and claims made from the test.

We next turn to the final two objectives to ***describe and demonstrate the novel multimethod research methodology*** and ***report on the findings* of a study** investigating whether the *CAEL* test scores delivered at a Test Centre are concordant with those delivered remotely (online). These last two objectives exemplify how one can adapt and apply the novel framework and research methodology.

To have this report be self-contained, ***Part II*** begins with a brief description of the *CAEL*. This description also informs the methodological and statistical choices made in *Chapter 4*, applying the novel multimethod approach to plan the concordance study of the *CAEL* delivered at test centres and online.

Part II - Applying the Novel Multimethod Approach for Operational and Validity Research to the Design of the Concordance of the *CAEL* Delivered at Test Centres and Online

Chapter 3 Brief Description of the *Canadian Academic English Language (CAEL)* Test

The *Canadian Academic English Language (CAEL)* test is a standardized test designed to measure the English language proficiency of students planning to study in Canadian post-secondary institutions. *CAEL* is fully computer-delivered and available via two methods of Administration - *CAEL* at a test centre and *CAEL* online. *CAEL* online includes the same test format, content, and reporting scale as the *CAEL* Test delivered at Paragon's test centres.

The *CAEL* Test provides an authentic representation of language use in a Canadian academic context. As would be expected in a first-year Canadian university or college classroom, test takers read articles, listen to a lecture, answer questions, and write an essay based on input from the reading and the lecture. Each test taker receives a score report showing their performance on each component as well as an overall score that is the average of the four individual component scores.

3.1 Test Format

CAEL assesses test takers' English language proficiency in an academic context. Test scores are reported on four components—Speaking, Reading, Listening, and Writing. Test takers complete a range of tasks. Some of these tasks will require test takers to use what they have read and listened to answer a question in speaking or writing. *Table 3.1* describes the format and content of each test component.

Table 3.1. Test Format

Component	Item Description	Number of Items	Time Allotted (Minutes)
Speaking	• Two speaking tasks, each based on a short question • One speaking task based on a graph/ diagram/ chart	3	7-10
Integrated Reading	• One or two short reading passages with comprehension questions • One or two long reading passages with comprehension questions • One speaking question, answered using material from a long reading passage	15-26	35-50
Integrated Listening	• One or two short listening passages with comprehension questions • One or two long listening passages with comprehension questions	15-26	25-35

	• One speaking question, answered using material from a long listening passage		
Academic Unit A	• One long reading passage with comprehension questions • One long listening passage on the same topic, with comprehension questions • One writing question requiring an extended response, using material from both the long reading passage and the long listening passage	23-31	60-70
Academic Unit B	• One long reading passage with comprehension questions • One long listening passage on the same topic, with comprehension questions • One writing question requiring a short response, using material from the long reading passage or the long listening passage	23-31	40-45

*Unscored Items: Each test contains unscored items used for test development. These unscored items can be found anywhere within each test and will have the same format as the scored items.

3.2 Scoring and Reporting of Results

The multiple-choice items are scored by computer. Each correct answer contributes proportionately to the final score, and no points are deducted for wrong answers. According to a scale established by Paragon, the Speaking and Writing components are each evaluated by at least four certified raters.

Test takers receive a score report including scores for the Speaking, Reading, Listening, and Writing components and an overall score. *CAEL* scores are reported on a 9-band scale from 10 - 90 with accompanying descriptors of what the performance represents. The overall score is calculated as an average of the four component scores rounded to the nearest band level. *Table 3.2* presents descriptions of test taker proficiency at each band level.

Table 3.2. *CAEL* Band Scores and Interpretation

CAEL Band	Descriptor
10-20	Low Beginner: Communicates with limited ability
30	High Beginner: Expresses basic ideas about familiar topics in routine settings
40	Intermediate: Demonstrates some ability to comprehend and articulate complex ideas and arguments typical of academic or professional settings
50	High Intermediate: Exhibits some competence in academic or professional settings; communication may break down in places
60	Advanced: Displays competence in academic or professional settings
70	Adept: Uses generally accurate language in most settings; some limitations in flexibility are evident
80-90	Expert: Demonstrates a high level of competence, accuracy, and effectiveness in academic/professional settings

For a detailed review of the test format, content, reporting scale, and score interpretation of *CAEL*, please visit the Paragon website (*CAEL* Test Reports: https://www.paragontesting.ca/about-research/test-reports/*CAEL*-test-reports/) and the *CAEL* official website (https://www.*CAEL*.ca/take-*CAEL*/overview/.

3.3 Online Test Delivery and Test Proctoring

The *CAEL Online* is the same test that is delivered at one of Paragon's test centres, and the format, content, and reporting scale remain unchanged. *CAEL Online* can be safely taken from home and is overseen by a live, human proctor through Examity, who will provide online proctoring and identity verification and monitoring for the entire test session. Paragon Testing staff also undertake a comprehensive secondary identity verification check after the test is administered and uses an extensive set of its own data forensics to identify any fraudulent activity.

The data reported herein were collected between June and October 2020. During this time period, Paragon Testing had partnered with *Examity* (URL: https://www.examity.com/) to conduct remote proctoring for Paragon during the *CAEL Online* test delivery, which will verify test taker's identity during the check-in procedure and will monitor test takers' behaviour during the test session. Since 2013, Examity has been working with colleges, universities, employers, and certification providers to maintain exam and institutional integrity.

Chapter 4 Applying the Novel MultiMethod Approach to Plan the Concordance Study of the *CAEL*

The logic and principles articulated in *Chapter 2* for designing the concordance study of the *CAEL* delivered at test centres and online are expressed in three parts.

- It has been well established in the psychometric literature since the early 1980s that comparing average test scores, average item scores, or average pass rates for the two groups are considered uninterpretable and flawed psychometric evidence because they confound concordance with true differences in item performance (i.e., impact). The same goes for simply comparing group psychometric properties such as test reliability for each group.
  A formal definition of statistical methods has replaced these unconditional approaches to investigate concordance based on the matching principle (see, for example, Angoff, 1993).
- Over and above the need for a research design founded on Angoff's matching principle, no optimal research method and design applies in all settings. Instead, a multimethod approach guided by (a) Zumbo's DLD framework, (b) establishing the Bona Fides for using the test scores blending Zumbo's model and Kane's argument-based approach in *Section 2.1*, (c) the guiding principles described in *Section 2.2*, and (d) the research design options in *Section 2.3* the multiple strands of statistical and psychometric evidence is organized in terms of *CAEL* test taker comparability, and *CAEL* test score comparability.

4.1 The Kinds of Online Tests (*Section 1.1.1*)

As described in *Chapter 3* and *Section 1.1.1,* the *CAEL CE* is an online alternative mode of test delivery with the same test format, content, and reporting scale as the *CAEL* test delivered at test centres. As such, by design, the *CAEL* is delivered online alongside the *CAEL* delivered at a test centre.

4.2 Applying the Theoretical Framework That Forms the Principles and Logic of Our Methodology to Investigate Online Testing (*Section 2.2*)

The equity principle in *Section 2.2.1* for an online alternative mode of test delivery with the same test format, content, and reporting scale, states that it should be a matter of indifference to a test taker or a test user as to which of two modes of test administration (test centre or online) the test taker chooses to take the test.
As described in *Table 2.1*, the implication of the equity principle is that test scores' are interchangeable across the two modes of administration. The first part of the evidence supporting this principle is in the test design. As we see in *Chapter 3*, the *CAEL*'s online and test centre versions were built to a common blueprint and designed to measure the

same constructs for the same use. Nearly everything about the tests themselves is the same except for the test delivery, at a test centre or online.

4.3 Unit of Analysis – The Component Band Score (*Section 2.2*)

The test use principle in *Table 2.1* states that comparison across test administrations to establish the level of concordance should be established on the scale on which scores are reported. **The implication of this principle for studies of the concordance of the *CAEL* is that the comparisons are made on the band scores on each of the (four) components of the language test**.

4.4 Concordance Study Design (*Section 2.3*)

As described in *Table 2.1*, the test use principle implies that there should be a high degree of interchangeability (exchangeability) between the reported score from the test centre and online tests.
The interest was comparing the test score comparability using the common between-subjects research design; we were interested in comparing test performance in an *in vivo* setting with maximal motivation and test performance. This design choice prioritizes making claims about live test performance but risks that the groups formed by test takers choose their desired mode of test delivery.

As implied by Zumbo's (2007a) Draper-Lindley-De Finetti (DLD) framework described in *Section 2.1.1*, *2.2*, and *Figure 2.4*, there should be a homogeneity or equivalence for (sub) populations of online test takers and those at a test centre. This comparability will be investigated in *Chapter 5*.

4.5 Evidence for Score Comparability

The two tests should produce comparable scores in terms of evidence and criteria of test score exchangeability arising from the test use principle. The evidence assembled to address score comparability includes (i) similar joint distribution (e.g., covariance matrix) of the scores on the test components investigating in *Chapter 6*, (ii) component score comparability using a between the modes of test delivery in *Chapters 7, 8, and 9*. *Chapter 7,* titled "Visualizing Comparability: Comparing Groups of Test Takers Using a Novel Kernel Smoothing Band Score Function," introduces an approach based on integrating out the score distribution to match the scores arising from the two modes of test administration. *Chapter 8* uses a variant of generalized linear model DIF analyses to investigate the score comparability. *Chapter 9*, on the other hand, investigates the score comparability using a propensity score-matched sample. In Chapters 7 through 9, in addition to the conventional comparison of mode of test delivery, meaningful subgroups of test takers defined by gender or first language were also investigated. Finally, *Chapter 10* reports an investigation of the comparative decision consistency and decision accuracy

at key cut-scores determined by test use- viz evidence of the comparative measurement uncertainty.

4.6 Transitioning to Parts III to V of the Report

The multimethod approach's strength stems from using the principled and logical approach described in *Chapter 2* to carefully weave together relevant psychometric evidence to inform our understanding of the *CAEL* language competencies- or, if you wish, construct.

As described above, the evidence assembled about the (a) *CAEL* test taker comparability in *Part III* of this report and (b) *CAEL* test score comparability in *Part IV e*stablish the inferential bounds of the claims one can make when both the test centre and online versions are in used side-by-side.

Part V of this report will weave together the evidence about test taker comparability and the five strands of evidence regarding the comparability of the test scores addressing whether the online and test centre tests measure the same language skills.

Part III - Investigating the Question of *CAEL Test Taker* Comparability Arising from the Between-Subjects Design

Chapter 5 Test Taker Comparability

The current study design is the standard and widely used approach to investigate psychometric comparability typical of DIF or concordance studies. To remind the reader of the study design choices (see *Chapter 4*), we used a two-group between-subjects design without random allocation; test takers chose their preferred mode of test delivery. One group completed the *CAEL* Online with remote proctoring. The other group completed the *CAEL* at a test centre with onsite proctoring in the usual manner.

- **The testing sessions for both groups during the same time period, between June and October 2020**.

The observational or quasi-experimental between-subjects study design allows us to compare the test performance of the test centre and online modes of test delivery under the setting of normal motivation for a test taker to perform as well as they can. However, the resultant two groups of test takers may be nonequivalent.

Therefore, the purposes of this chapter are to

1) describe the study sample reported in this concordance study,
2) report on analyses to investigate the equivalence of the two test taker groups based on their reported age, gender, and the first language based on the logic and principles described in *Section 2.1.1, 2.2*, and *Figure 2.4*, and
3) report data on the equivalence of the two taker groups to past test data, noting, of course, that the 2020 data were collected during a global pandemic.

There are, of course, many other variables that relate to the test taking experience; however, as a practical concern of not over-burdening the test taker and data privacy policies regarding the storage of test taker data, as standard operational practice, Paragon Testing collects test taker reported age, gender, and first language. Other recorded test taker data is not of the form that allows them to be treated as statistical covariates.

It should be noted that to account for any possible pre-existing group differences, in *Chapters 8 to 10,* we matched the sample of test takers who took *CAEL* Online with the pool of test takers who took *CAEL* at a test centre using these same variables (test taker's reported age, gender, and first language) using covariance analysis methods in *Chapter 8*, and a propensity score-matched sample method in *Chapters 9* and *10*. An advantage of our multimethod approach is that we borrow strength from the various methods.

5.1 Sample Size and Composition

The focus of the analysis was the *CAEL* test taker data between June and October 2020. A total of **1,455 test takers** completed the *CAEL* either at a **test centre (n = 765)** or **online (n = 690)**. The table below contains the sample sizes of the mode of test delivery by gender of the test taker.

5.1.1 Adequacy of the Sample Size for Planned Analyses

The question naturally arises as to how many test takers should be involved in a between-subjects concordance study design. The two main kinds of statistical analyses to address the question are kernel smoothing methods reported in *Chapter 7* and generalized linear model (GLIM) DIF methods reported in *Chapter 8*.

- As described in *Section 7.6*, a minimum of 200 test takers per sub-group is recommended for the kernel smoothing methods (e.g., Ramsay & Silverman, 2002, 2005).
- As described in *Section 8.2.2*, Scott et al. (2009) and Herrera and Gómez (2008) recommend a minimum of 250 test takers in either sub-group to use conventional GLIM DIF analyses. In addition, if multiple matching and more complex GLIM DIF models are used, Bujang, Sa'at, and Tg Abu Bakar Sidik, and Lim (2018) recommend that the overall sample size of (100+50p), where p refers to the number of independent variables in the final GLIM model.

On the balance of the recommendations for the two main statistical methods for concordance analysis of test centre and online test delivery, we would need a minimum of 250 test takers per group.

Our total sample of 1,455 test takers with 765 in the test centre group and 690 in the online group, meets and exceeds the requirements above for GLIM DIF analyses. However, our sample size limits the analyses into subgroups in *Chapter 7* using the novel nonparametric kernel smoothing band score function to condition a latent matching variable to integrate the score distribution. Likewise, with our sample size, one needs to approach creating matched samples with caution using a statistical approach such as propensity score matching in *Chapters 9 and 10*.

5.1.2 Comparability of the Composition of the Mode of Test Delivery Groups by Gender of the Test Taker

We tested whether the frequency of cases for the two modes of test delivery varied among combinations of levels of the gender of the test taker—*Table 5.1* lists the samples for the 2x2 table and the column and row totals.

A Chi-square test of goodness of fit was used to test the hypothesis that the total sample is distributed evenly among all levels of the relevant factor. The resulting $\chi^2(1) = 0.52$, p=0.47, allows us to conclude that female and male test takers are distributed evenly among the two modes of test delivery.

- **Conclusion:** This evidence supports the conclusion that overall the two groups of test takers are exchangeable (comparable) from the point of view of the gender of the test taker.

Table 5.1. Sample sizes for Mode of Test Delivery by Gender of the Test Taker

	Mode of Test Delivery		
Gender of the Test Taker	Online	Test Centre	*Row Totals*
Female	326	347	673
Male	364	418	782
Column Totals	690	765	Overall= 1,455

5.1.3 Comparability of the Mode of Test Delivery Groups by the Age (in years) of Test Takers

We tested whether the distribution of the test takers' age for the two modes of test delivery differed among combinations of levels of the gender of the test taker. *Table 5.2* lists the means, minimum and maximum values, and standard deviations for the two test taker groups.

A t-test, with Welch's correction allowing for unequal variances between groups, was used to test the hypothesis of equality means of the two groups. The resulting t(1447)= -0.946, p=0.34, and corresponding Cohen's d effect size of -0.0496. The result is a statistically non-significant p-value and a Cohen's d indicating that the means of the two groups differ by 0.0496 of a standard deviation-far below any criteria for meaningful differences (Kirk, 2001). These results allow us to conclude that the average ages of the test centre and online test takers are equal.

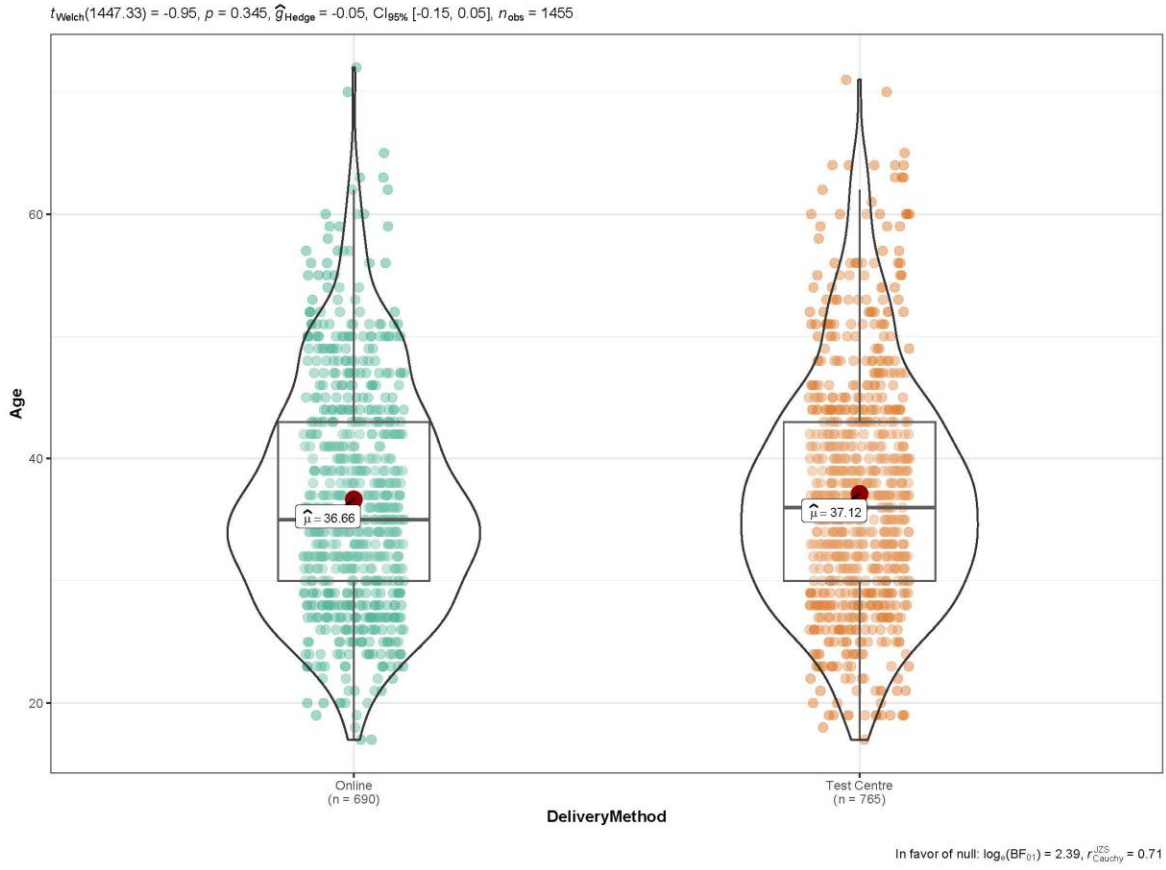
Given that the means of groups may be equal, the distributional shape may be different *Figure 5.1*. displays the violin plots (Hintze & Nelson, 1998), allowing us to compare the age (in years) of the two modes of test delivery. Violin plots display numeric data that combine the advantage of a box plot and a kernel density plot. In the violin plot, the box in the center, from the bottom upward, displays the first quartile, median (second quartile), and third quartiles. The area within the box, bounded by the 25^{th} and 75^{th} percentiles, represents the midspread, middle 50%, or H-spread, of the score distribution and is a robust measure of statistical dispersion. In addition, the entire score distribution is displayed using the rotated kernel density plot on each side in the shape of a violin. By visual inspection of the violin plots in *Figure 5.1*, we can conclude that the two test taker groups have similar medians and midspread.

- **Conclusion:** The evidence from the t-test and the violin plots supports the conclusion that overall, the two groups of test takers are exchangeable (comparable) from the point of view of the test taker's age.

Table 5.2. Descriptive Statistics of the Age (in years) of the Test Taker by Mode of Test Delivery

	Mode of Test Delivery	
	Online	Test Centre
Mean	36.7	37.1
Minimum and Maximum	17-72	17-71
Standard Deviations	9.13	9.51
Sample Sizes	N=690	N=765

Figure 5.1. Violin Plots to Compare the Age (in years) of the Test Taker for the Two Groups of Mode of Test Delivery



5.1.4 Comparability of the Composition of the Mode of Test Delivery Groups by Test Takers' Reported First Language – A Language Families Approach

Test takers self-reported their "first language" by responding to the question "What is your first language?". *Section 5.3* resolves two issues with responses to this question. First, the concept of "first language" is ambiguous, and it leads to a great number of response categories (i.e., 68 first languages) reported in the *CAEL* test taker sample. Many

individual languages, many with small sample sizes, are not feasible as covariates in concordance studies. *Section 5.3* describes a detailed analysis based on contemporary linguistic theory for language classification that focuses on linguistic considerations such as systematic sound correspondences in basic vocabulary and patterned grammatical evidence in the languages being compared—resulting in a two-level Indo-European (IE) vs. Non-Indo-European (Non-IE) distinction as the grouping variable.

We tested whether the frequency of cases for the two modes of test delivery varied among combinations of levels of language family (first language) of the test taker—*Table 5.2* lists the samples for the 2x2 table and the column and row totals.

A Chi-square test of goodness of fit was used to test the hypothesis that the total sample is distributed evenly among all levels of the relevant factor. The resulting $\chi^2(1) = 0.63$, p=0.43, allows us to conclude that two language family groups are distributed evenly among the two modes of test delivery.

- **Conclusion:** This evidence supports the conclusion that overall the two groups of test takers are exchangeable (comparable) from the point of view of the language family group (first language) of the test taker.

Table 5.2. Sample size Mode of Test Delivery by First Language (Language Family Group) of the Test Taker

	Mode of Test Delivery		
First Language (Language Family Group)	Online	Test Centre	*Row Totals*
Indo-European	507	576	1083
Non-Indo-European	183	189	372
Column Totals	690	765	Overall= 1,455

5.1.5 Comparability of the Composition of the Mode of Test Delivery Groups by Gender, Age, First Language (Language Family)

To this point, we have investigated the equivalence of the two groups reflecting test centre and online modes of test delivery separately in terms of:

- the number of female and male test takers,
- the distribution of the test takers' age, and
- their reported first language-operationalized as either Indo-European or Non-Indo-European Language Family Group.

These three effects can be considered as main effects in a nonexperimental fully-crossed 2x2x2 factorial design. We next use this factorial design concept to investigate possible two-way and three-way interactions- where differences in the age of test takers in the test centre and online groups may depend on their gender or first language.

The descriptive statistics for the eight cells of the fully-crossed 2x2x2 factorial design are reported in *Table 5.3*. The average ages (in years) are, for the most part, similar for seven of the eight cells of the design. The oldest average age group comprises male test takers who reported a non-Indo-European first language family group and chose to take the *CAEL* online.

The results of the ANOVA are displayed in *Figure 5.2*. The only statistically significant age differences are the main effects of gender (means F=36.4, M=38.4) and first language (means IE=36.3, NIE=38.4), and both of those effects are in the range of a small effect $\eta^2 = 0.01$ (Kirk, 2001).

Table 5.3. Descriptive Statistics for the Comparison of Age (in years) of the Mode of Test Delivery by Gender, Test Takers' Reported First Language and Age

Indo-European Language Family Group		
	Mode of Test Delivery	
Gender	Online	Test Centre
Female	Mean= 35.6 StDev= 8.86 Min-Max= 20-62 N= 228	Mean= 35.7 StDev= 8.40 Min-Max= 19-64 N= 254
Male	Mean= 36.6 StDev= 9.32 Min-Max= 19-72 N= 279	Mean= 37.3 StDev= 9.82 Min-Max= 17-71 N= 322
Non-Indo-European Language Family Group		
	Mode of Test Delivery	
Gender	Online	Test Centre
Female	Mean= 36.1 StDev= 8.50 Min-Max= 17-62 N= 98	Mean= 38.1 StDev= 8.72 Min-Max= 20-60 N= 93
Male	Mean= 40.3 StDev= 9.20 Min-Max= 18-60 N= 85	Mean= 39.3 StDev= 11.40 Min-Max= 18-65 N= 96

Figure 5.2. ANOVA Decomposition Table for the Comparison of Age (in years) of the Mode of Test Delivery by Gender, Test Takers' Reported First Language and Age

ANOVA - Age

	Sum of Squares	df	Mean Square	F	p	η^2
DeliveryMethod	57.37	1	57.37	0.6690	0.4135	0.000
Gender	1132.77	1	1132.77	13.2084	2.885e-4	0.009
lang_family	1263.31	1	1263.31	14.7304	1.294e-4	0.010
DeliveryMethod ✻ Gender	97.78	1	97.78	1.1402	0.2858	0.001
DeliveryMethod ✻ lang_family	1.30	1	1.30	0.0151	0.9021	0.000
Gender ✻ lang_family	145.80	1	145.80	1.7000	0.1925	0.001
DeliveryMethod ✻ Gender ✻ lang_family	229.61	1	229.61	2.6773	0.1020	0.002
Residuals	124097.45	1447	85.76			

Figure 5.3 displays the marginal mean plots of the test takers' age (in years), including the standard error for each mean representing uncertainty in the estimate for the 2x2x2 fully-crossed factorial design. As noted earlier from the table of means, the average oldest group comprises male test takers who reported a non-Indo-European first language family group and chose to take the *CAEL* online. The reader should be cautioned of two points. First, the scale of the age axis zooms into a 5-year age range, and as one sees in *Figures 5.4* and *5.5*, what appear as large differences in the marginal means plots are much small when the full data points and their range is taken into account. Second, in support of the first point, it should be noted that none of the interactions were statistically significant in the factorial ANOVA, likely due to the substantial inherent variability visible in *Figure 5.3* in the standard errors and *Figures 5.4* and *5.5* in the data distributions.

Figure 5.3 The Marginal Mean Plots of the Factorial Design (standard errors)

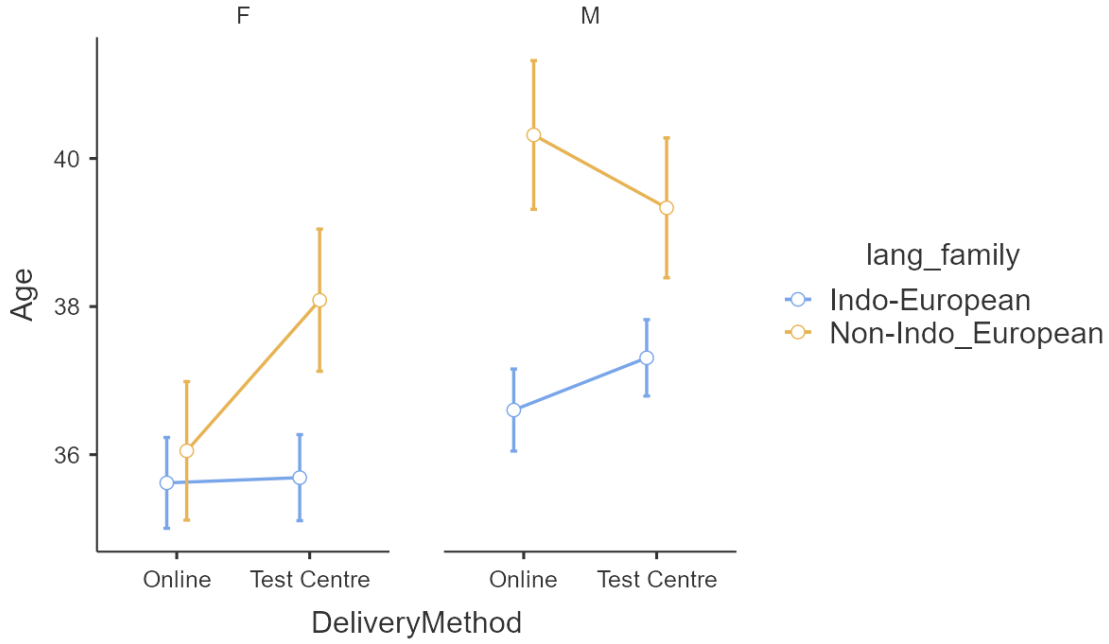


Figure 5.4 The Marginal Distribution and Mean Plots of the Factorial Design

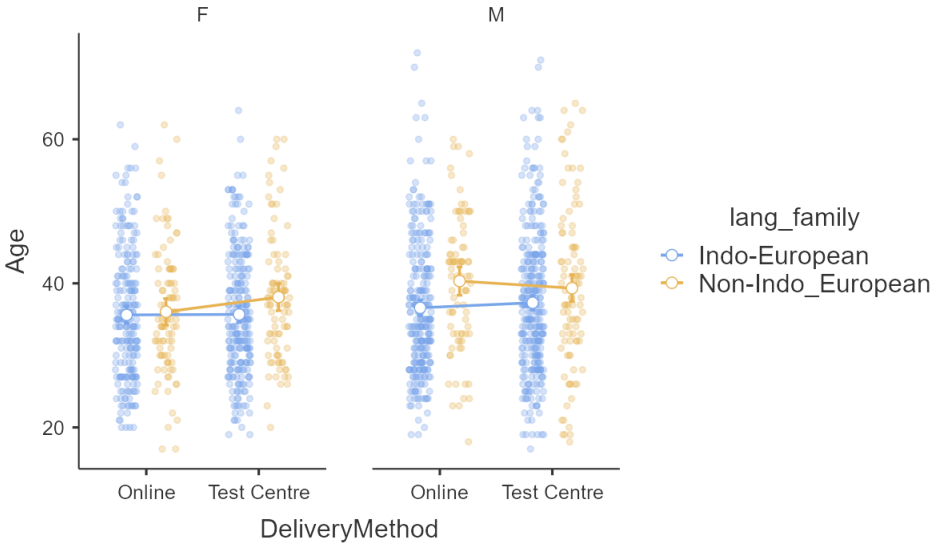


Figure 5.5. Violin Plot for the Comparison of Age (in years) for the Mode of Test Delivery by Gender and by the Test Takers' Reported First Language (continued next page)

A. For Female Test Takers

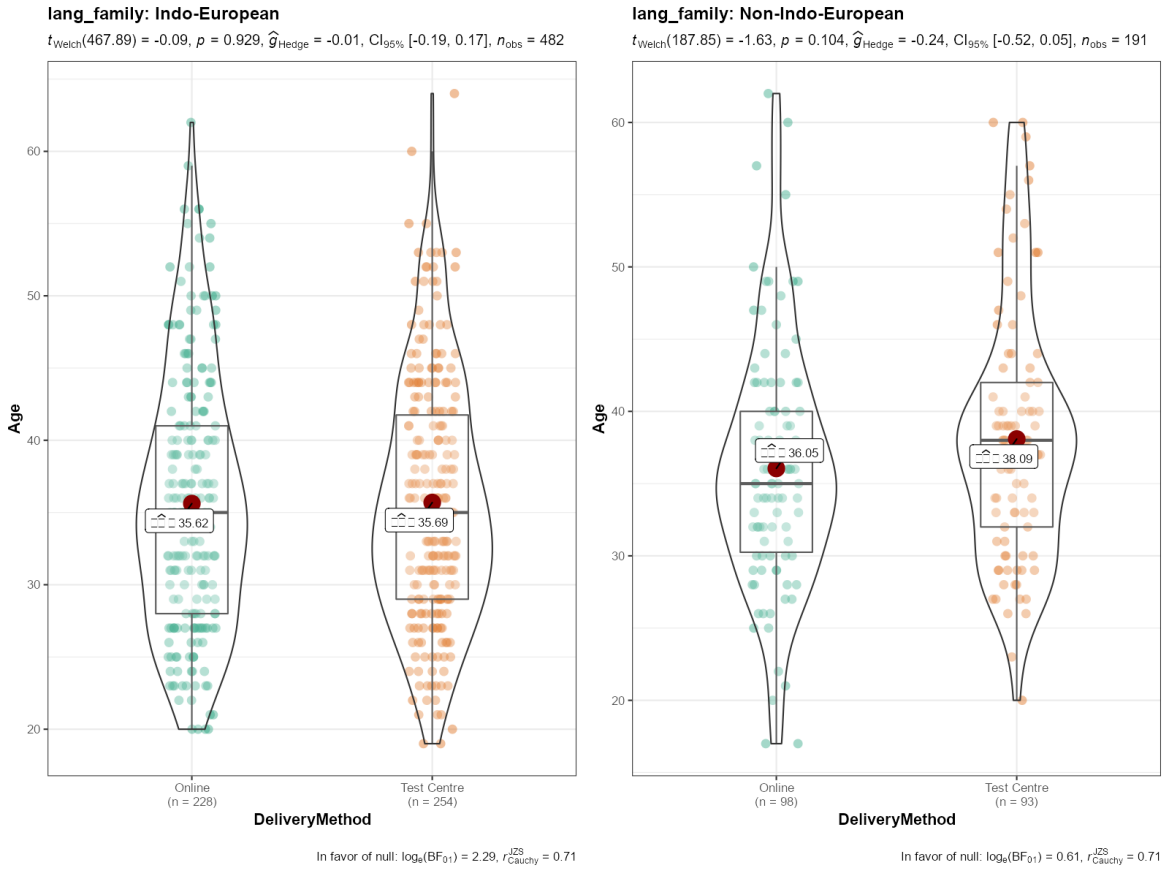
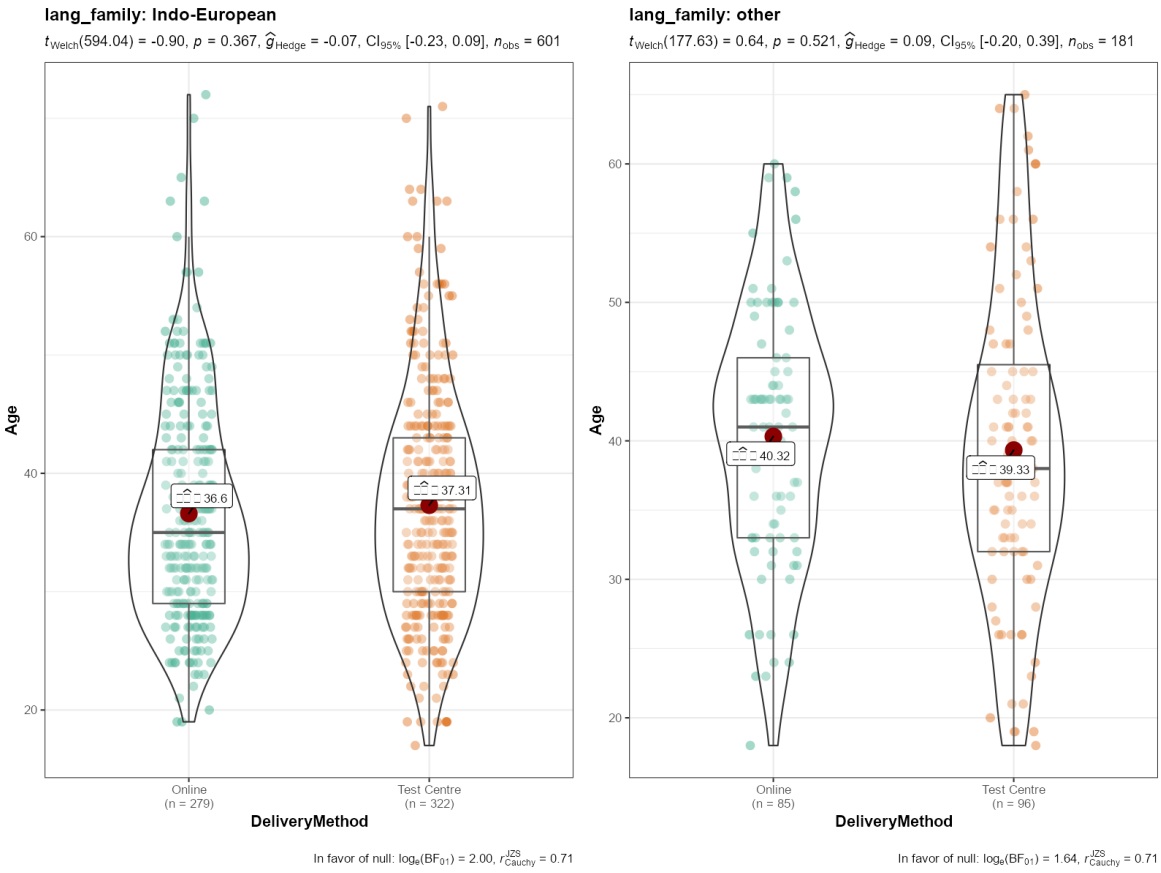


Figure 5.5. Violin Plot for the Comparison of Age (in years) for the Mode of Test Delivery by Gender and by the Test Takers' Reported First Language (continued)

B. For Male Test Takers



- **Conclusion:** The evidence reported in this section supports the conclusion that overall, the eight subgroups in the factorial design mostly have a small difference in average age except for male test takers who reported a non-Indo-European first language family group and chose to take the *CAEL* online.
    - It is also noteworthy that the distributional shapes vary, as does the inherent variability in each group.
    - Although the mean differences were not statistically significant and the within-group variability is, for the most part, similar age should be considered amount the key matching covariates in *Chapters 8* through *10*.

5.1.6 Comparison of the Age Distribution and also the Band Scores on the Four Components Earned for Test Takers in 2019 Test Centre, 2020 Test Centre, and 2020 Online

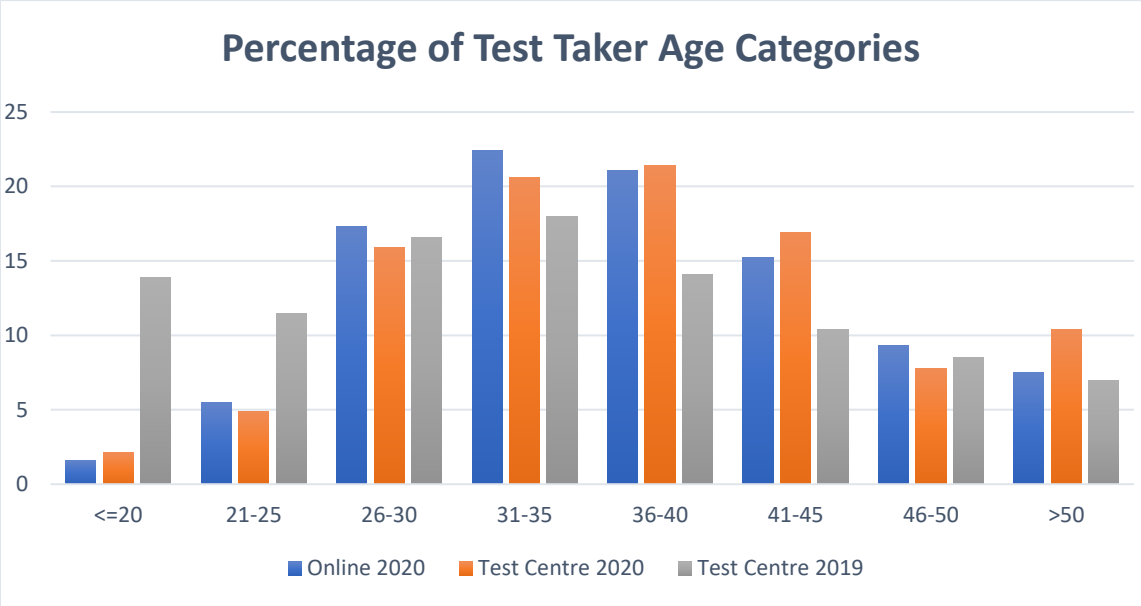
Given that the 2020 data were collected during the (ongoing) global pandemic, a natural question arises about the equivalence of the two groups of test takers who completed the *CAEL* at a test centre or online between June and October 2020 to past test data. The comparative prior test data information came from Paragon Testing's *CAEL CE Annual Report of 2019 Test Takers* URL: https://www.paragontesting.ca/wp-content/uploads/2020/03/*CAEL*-Test-Report-2019.pdf

Two points are important to keep in mind regarding this comparison.

- In terms of the logic and principles described in *Section 2.1.1, Section 2.2*, and exchangeability conditions described in *Figure 2.4*, the equivalence of the two groups of test takers in 2020 has primacy over the equivalence with past data because the within-study equivalence is important to claims of internal validity in the quasi-experimental design. However, the equivalence with past test taker populations speaks to the matter of external validity of the study results.
- Statistical hypothesis tests of the equality of proportions over three time periods are not conducted because tests of multiplicity problems, these hypothesis tests typically do not offer posthoc tests, and concerns over non-fixed marginals and totals. The data will be displayed and provided in tabular form in *Figures 5.6* and *5.7*.

- **Conclusion:** The pattern of findings reported in *Figures 5.6* and *5.7* shows:
  a) Both of the 2020 test taker groups were, on average, younger than the test takers in 2019.
     - There was a greater prevalence of younger test takers in 2019; particularly, more test takers in the ≤ 25 year-old age categories in 2019 than in the two 2020 samples. Consequently, there were fewer test takers in 2019 in the older 31-45 age categories.
  b) The pattern of test score results, band scores, was the same in 2019 and the two 2020 samples- this pattern holds for all four language components.

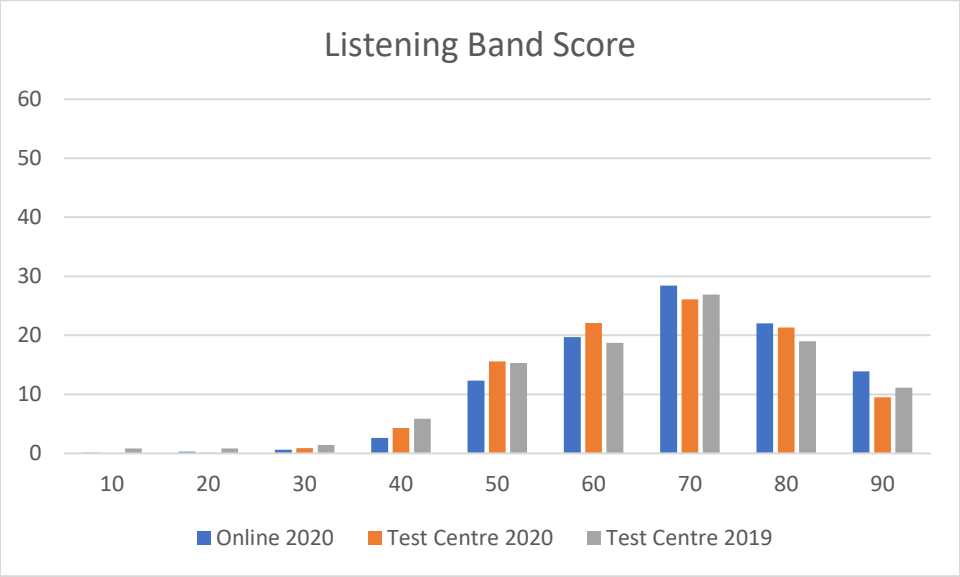
Figure 5.6. Age Distribution for Test Centre 2019, Test Centre 2020, and Online 2020 – Table of Percentages Below the Chart



Age Category	Online 2020	Test Centre 2020	Test Centre 2019
<=20	1.6	2.1	13.9
21-25	5.5	4.9	11.5
26-30	17.3	15.9	16.6
31-35	22.4	20.6	18.0
36-40	21.1	21.4	14.1
41-45	15.2	16.9	10.4
46-50	9.3	7.8	8.5
>50	7.5	10.4	7.0

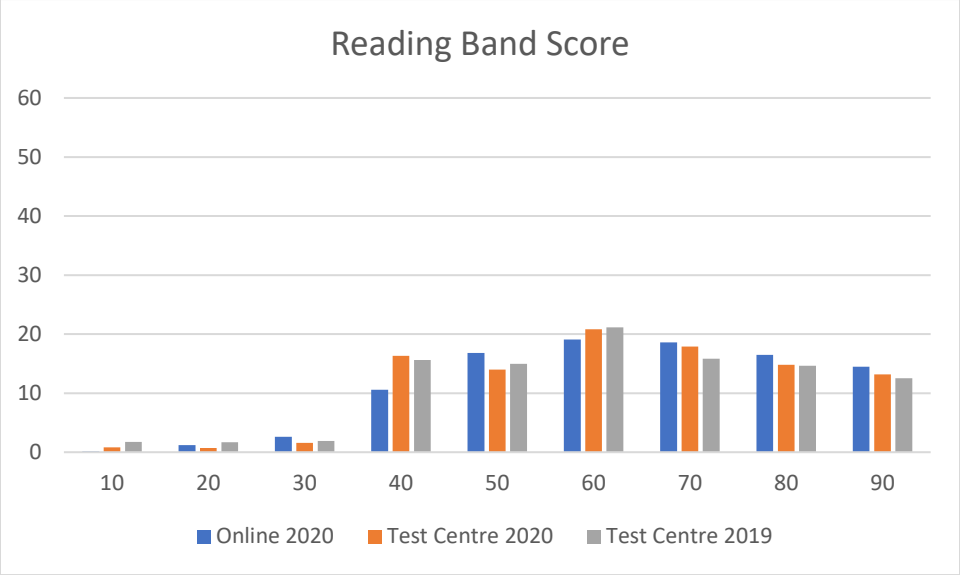
Figure 5.7. Distribution of Band Scores on the Four Components Earned by Test Takers in 2019 Test Centre, 2020 Test Centre, and 2020 Online – Table of Percentages Below the Chart

A. Listening Component – Percentage test takers in each of the band scores



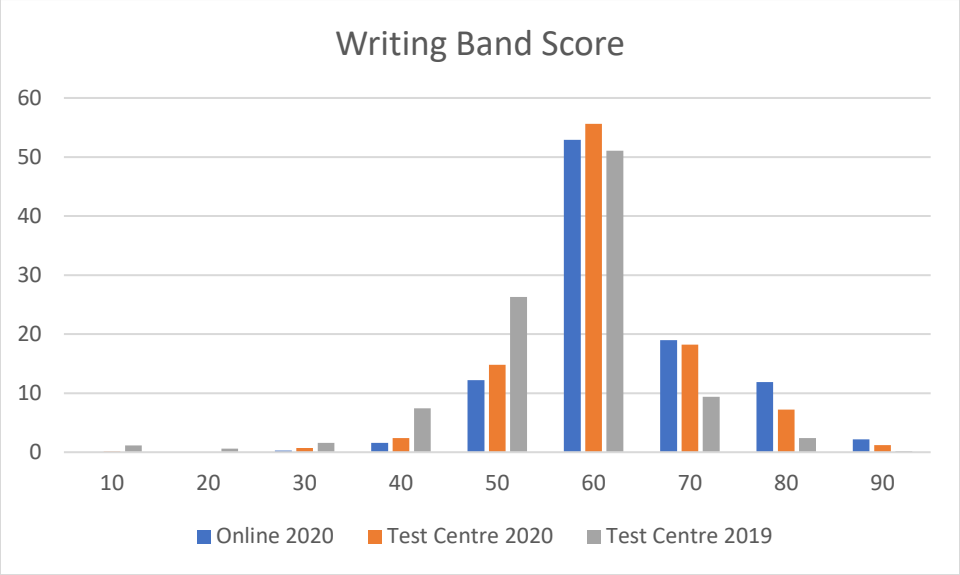
Band Score	Online 2020	Test Centre 2020	Test Centre 2019
10	0.1	0.0	0.8
20	0.3	0.1	0.8
30	0.6	0.9	1.4
40	2.6	4.3	5.9
50	12.3	15.6	15.3
60	19.7	22.1	18.7
70	28.4	26.1	26.9
80	22.0	21.3	19.0
90	13.9	9.5	11.2

B. Reading Component – Percentage test takers in each of the band scores



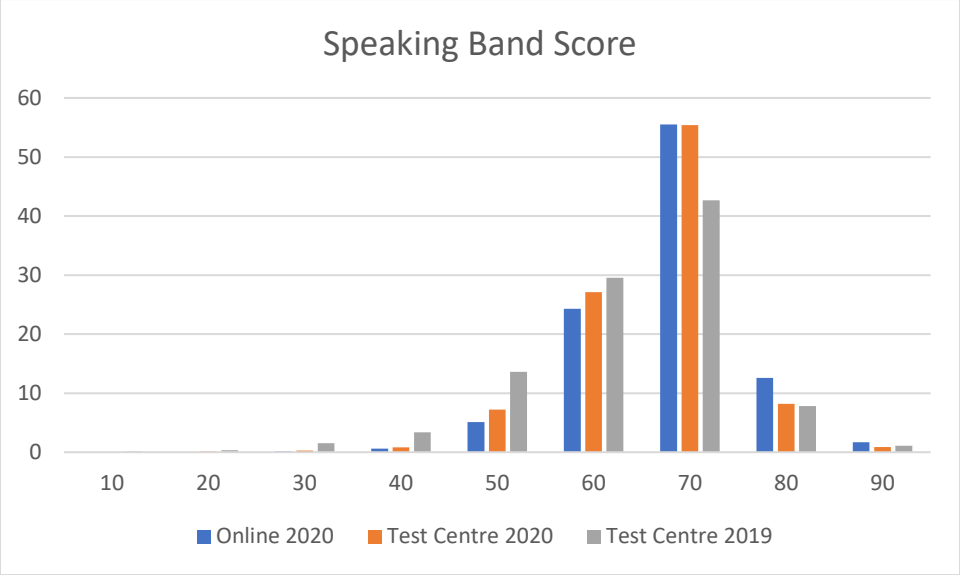
Band Score	Online 2020	Test Centre 2020	Test Centre 2019
10	0.1	0.8	1.73
20	1.2	0.7	1.67
30	2.6	1.6	1.88
40	10.6	16.3	15.59
50	16.8	14	14.97
60	19.1	20.8	21.14
70	18.6	17.9	15.86
80	16.5	14.8	14.65
90	14.5	13.2	12.51

C. Writing Component – Percentage test takers in each of the band scores



Band Score	Online 2020	Test Centre 2020	Test Centre 2019
10	0	0.1	1.15
20	0	0	0.58
30	0.3	0.7	1.57
40	1.6	2.4	7.43
50	12.2	14.8	26.27
60	52.9	55.6	51.07
70	19	18.2	9.37
80	11.9	7.2	2.41
90	2.2	1.2	0.16

D. Speaking Component – Percentage test takers in each of the band scores



Band Score	Online 2020	Test Centre 2020	Test Centre 2019
10	0	0	0.1
20	0	0.1	0.31
30	0.1	0.3	1.52
40	0.6	0.8	3.35
50	5.1	7.2	13.61
60	24.3	27.1	29.57
70	55.5	55.4	42.65
80	12.6	8.2	7.8
90	1.7	0.9	1.1

5.2 A Consideration of Test Takers' Reported First Language (L1): Language Family as a Grouping Variable

Test takers' self-reported first languages (L1s) are collected as responses to the question "What is your first language?". Due to the great number of L1s (n = 68) reported in the *CAEL* test taker sample, it is not feasible to use individual languages as a grouping variable for subsequent statistical analysis of comparability. We investigated methods to construct language groups based on test takers' self-reported L1s to reduce the number of levels. One possible way for grouping L1s is by language family, a phylogenetic unit analogical to a biological family, whose members are descendants from a common parental language or proto-language (Pereltsvaig, 2012).

Contemporary methods for language classification focus on linguistic considerations such as systematic sound correspondences in basic vocabulary and patterned grammatical evidence in the languages being compared. Empirically, however, such linguistically-motivated classifications may not be completely independent of those drawn from non-

linguistic, broad anthropological interests such as national borders and ethnic groups (Campbell & Poser, 2008). Therefore, language classifications may bear more than linguistic relevance but be well taken as proxies for an amalgamation of social and cultural factors and experiences.

With these caveats in mind, we introduce the process of investigation, which leads us to the decision that classifying test takers' self-reported L1s into Indo-European vs. Non-Indo-European groups is the most feasible solution for our current purpose.

5.2.1 Procedures

The L1s are mapped to language families according to Ethnologue (Eberhard, Simons, & Fennig, 2020), a printed and online resource that provides statistics and other information on the living languages of the world.

The L1 labels go through a process of cleaning before they are mapped to a language family to reduce redundancies and resolve ambiguity, e.g., *Farsi* and *Persian (Farsi)* correspond to separate L1 labels, and *Chinese*, *Cantonese (Chinese),* and *Mandarin (Chinese)* are listed as separate L1s. However, no further detail is accessible to help identify the specific variety of Chinese intended. Therefore, for L1 labels that lack specific information to be mapped to one iso-693 language code or are only mapped to a more general group of families, an iso-693 of one individual language in the same group is assigned the purpose of mapping. The "Other" label provides no information to be mapped to a language or family and will be shown as such in the classification.

The cleaned L1 labels are then mapped to an iso-693 code, a unique identifier for all languages in the world (SIL, 2007). Finally, each iso-693 code associated with an L1 label is mapped to the corresponding language classification listed in the Ethnologue (Eberhard et al., 2020).

5.2.2 Results

Language families are structured like phylogenetic trees with varying levels and spread: the top node represents the proto-language from which all other languages in the families descended. Each node may branch out further to include descendent languages or sub-language families. For instance, the Indo-European language family subsumes five large sub-families, including the Germanic language family from which English, German, and Dutch descended. *Table 5.4* shows the number of test takers and L1s for all tests and online and at test centre tests for each language family represented in the sample. For ease of comparison, only the top level of each language family is shown.

Table 5.4 Language Family and Count of Test Takers and L1s

	Overall		Number of Test Takers		Number of L1s	
Language Family	**Number of Test Takers**	**Number of L1s**	**Online**	**Test Centre**	**Online**	**Test Centre**
Indo-European	1,083	29	507	576	23	26
Sino-Tibetan	106	3	57	49	3	3
Afro-Asiatic	94	6	56	38	4	5
Austronesian	59	6	20	39	4	5
Niger-Congo	29	9	10	19	4	8
Dravidian	28	4	17	11	4	3
Koreanic	24	1	8	16	1	1
Austro-Asiatic	15	1	8	7	1	1
Turkic	11	4	4	7	3	2
Kra-Dai	2	1		2		1
(Other)	1	1	1		1	
Creole	1	1	1		1	
Nakh-Daghestanian	1	1		1		1
Japonic	1	1	1		1	
Grand Total	**1,455**	**68**	**690**	**765**	**50**	**56**

The table shows no significant disparity in the distribution of language families between the online and test centre modes of delivery. However, the language families vary greatly in size (in terms of the number of individual languages), levels of categorization, and representativeness in the test taker sample. For instance, the largest language family, Indo-European, subsumes 1,083 (74% of 1,455) Test takers and 29 (42% of 68) L1s. While having only 29 Test takers (2% of 1,455) overall, the Nigerian-Congo family contains 9 (13% of 68) L1s. Some language families are dominated by one or a few languages. In the case of language isolates such as Japanese and Korean, which are not descendants of other languages, each constitutes an individual language family (Japonic and Koreanic). For most families and sub-families, dominant languages are far more represented in the sample than others (e.g., 97.5% of test takers from the Germanic language group are English L1s). For a detailed breakdown of language families and individual languages, see *Appendix Section 5.4*.

5.2.3 Recommended Grouping

Due to the issues discussed above, the most appropriate classification method will need to consider linguistic genealogy and representativeness in the test taker sample and relevance to the *CAEL* test. This will inevitably lead to a pragmatic classification that places unequal prominence of languages in our analysis, ignores certain within-group variability, or assumes homogeneity that may not be entirely "linguistically sound." However, this will remain true whether we use L1 or any type of L1 grouping as a variable. Caution is recommended in our generalization and interpretation of the results.

We have adopted the two-level Indo-European (IE) vs. Non-Indo-European (Non-IE) distinction as the grouping variable for the current analysis. This means that the majority of *CAEL* test takers will fall into the IE category for the present and incoming test takers unless there is a significant change in the test taker population.

The IE vs. Non-IE classification method proposed is merely one possible way of grouping. Alternative classifications may be adopted depending on the focus of research and characteristics of the test taker sample. For instance, one may consider breaking the Indo-European and non-Indo-European languages into a few groups, depending on the representativeness of test takers in the sample, the number of resulting categories, and other special considerations (e.g., English L1 as a separate group). *Table 5.5* presents two possible classifications with a count of test taker numbers.

Table 5.5 The IE vs. Non-IE Grouping and an Alternative Classification

Class 1	**Number of Test takers**	**Class 2**	**Number of Test takers**
IE	1,083	Indo-Aryan	642
		Iranian	97
		Germanic	241
		Other IE	103
Non-IE	372	Sino-Tibetan/Chinese	106
		Other Non-IE	266
Grand Total			1,455

5.3 Overall Conclusions About the Comparability of Test Takers

Our total sample of 1,455 test takers with 765 in the test centre group and 690 in the online group, meets and exceeds the requirements above for DIF analyses. However, caution should be heeded when considering sub-group analyses based on the three main covariates of gender, first language, and age.

The test taker groups are distributed evenly according to the test centre and online groups for gender and the first language of the test takers.

The test centre and online groups of test takers are equivalent from the point of view of the test taker's age. Overall, the eight subgroups defined by the mode of test delivery, gender, and first language categories showed a statistically nonsignificant or small effect size in the factorial ANOVA, suggesting that the groups are of equal average age.

Overall, the online and test centre test takers are comparable relative to the covariates at our avail. The 2020 test taker groups were, on average, younger than the test takers in 2019.

We will close this section by considering the implications of the findings in this chapter regarding the comparability of the test centre and online test takers for the investigation of test score comparability in the remaining chapters of this report. The differences in the average age were not statistically significant for the test centre and online test takers, and the within-group age variability is, for the most part, similar. Nonetheless, it would be prudent to consider age among the key matching covariates where possible in *Chapters 8* and *9,* which involve testing (generalized) linear models.

5.4 Appendix Chapter 5- Test-Taker Count in Each Language Family (Detailed Breakdown, Including Individual L1s)

Level 0 = blue; Level 1 = green; Level 2 = yellow

Language Family	Sum of Number of Test Takers
Indo-European	**1083**
Indo-Iranian	**739**
Indo-Aryan	642
Panjabi; Punjabi	331
Hindi	180
Urdu	55
Gujarati	45
Bengali	20
Nepali	4
Sinhalese	3
Marathi	2
Sindhi	2
Iranian	97
Farsi	67
Persian (Farsi)	20
Pushto	8
Kurdish	2
Germanic	**241**
West	241
English	235
German	3
Dutch; Flemish	3
Italic	**72**
Romance	72
Spanish	39
Portuguese	12
French	11
Spanish; Castilian	5
Romanian	4
Italian	1
Balto-Slavic	**24**
Slavic	24
Russian	14
Ukrainian	6
Polish	1
Slovenian	1
Croatian	1

Language Family	Sum of Number of Test Takers
Serbian	1
Albanian	**7**
Tosk	7
Albanian	7
Sino-Tibetan	**106**
Chinese	**106**
(blank)	106
Chinese	64
Mandarin (Chinese)	32
Cantonese (Chinese)	10
Afro-Asiatic	**94**
Semitic	**91**
Central	82
Arabic	81
Hebrew	1
South	9
Tigrinya	6
Amharic	3
Cushitic	**3**
East	3
Somali	2
Afar	1
Austronesian	**59**
Malayo-Polynesian	**59**
Greater Central Philippine	56
Tagalog	46
Philippine (Other)	8
Cebuano	2
Malayo-Chamic	2
Malay	1
Indonesian	1
Northern Luzon	1
Iloko	1
Niger-Congo	**29**
Atlantic-Congo	**29**
Volta-Congo	28
Igbo	10
Yoruba	8
Swahili	3
Twi	2

Language Family	Sum of Number of Test Takers
Akan	2
Bini	1
Tiv	1
Shona	1
Atlantic	1
Fulah	1
Dravidian	**28**
Southern	**20**
Tamil-Kannada	20
Tamil	10
Malayalam	9
Kannada	1
South-Central	**8**
Telugu	8
Telugu	8
Koreanic	**24**
(blank)	**24**
(blank)	24
Korean	24
Austro-Asiatic	**15**
Mon-Khmer	**15**
Viet-Muong	15
Vietnamese	15
Turkic	**11**
Southern	**9**
Turkish	7
Turkish	7
Azerbaijani	2
Azerbaijani	2
Western	**1**
Aralo-Caspian	1
Kazakh	1
Eastern	**1**
(blank)	1
Uzbek	1
Kra-Dai	**2**
Kam-Tai	**2**
Tai	2
Thai	2
(blank)	**1**

Language Family	Sum of Number of Test Takers
Creole	**1**
French based	**1**
(blank)	1
Creoles and pidgins, French-based (Other)	1
Nakh-Daghestanian	**1**
Nakh	**1**
Chechen-Ingush	1
Chechen	1
Japonic	**1**
(blank)	**1**
(blank)	1
Japanese	1
Grand Total	**1455**

Part IV - Five Strands of Evidence to Investigate *Test Score* Comparability

Test Score Interchangeability or Equivalence of Test Centre and Remote Test Administration

Chapter 6 Covariance Analysis- Dispersion Matrices and Comparative Factor Analyses.

This chapter reports on an investigation of the comparative multivariate structure of the four components of the *CAEL* delivered at test centres and online. The three sections of this chapter investigate the comparative multivariate structure of the four component scores of the two test taker groups with applying progressively more structure to the multivariate data.

The equality of comparison of unstructured dispersion matrices, *Section 6.2*, provides initial information about the latent structure of the assessment. Latent variable dimensionality is an essential next step, *Sections 6.3* and *6.4* because the same test could be unidimensional for one examinee population and not for another. The reader should be reminded that a test is designed to reflect test taker performance on a latent variable and minor secondary, latent dimensions. It is important in terms of validity theory, see *Chapter 2*, when considering assessment and test data, the classification of latent dimensions into those intended-to-be-measured, such as language ability (called essential, dominant, or major) dimensions and the unintended-to-be-measured (called an inessential, weak, nuisance, or minor) dimensions. This concept is rooted in the factor analytic tradition: see, for example, Tucker, Koopman, and Linn (1969) for a factor analytic model distinguishing between minor (and hence inessential) factors and major factors. However, this distinction implicitly manifests itself in Cronbach and others' characterization of construct relevant and irrelevant variance when considering test validation.

Overall Conclusions

The analyses of the covariance matrices consistently supported the equality of the covariance matrices for Test Centre and Online band scores.
This is evidence in support of the comparability of test centre and online test performance because the equality dispersion matrices, (i.e., the covariance matrices), essential unidimensionality, and metric invariance are indicators that the two sub-populations of test takers (by mode of administration) are interchangeable in terms of their joint distributions.

In other words, this finding supports the concordance of the Test Centre and Online modes of test delivery in terms of the band scores of the four language domains.

6.1 Sample Data and Treating Band Scores as Continous vs Ordered Categorical Variables

The data used herein are described in *Chapter 5*; *CAEL* test taker data collected between June and October 2020. There were 1,455 test takers; 765 completed the *CAEL* at a test centre and 690 online.

The focus continues to be on the band score because band scores' comparability is essential to *CAEL*'s concordance across modes of test delivery- that is, this is central to treating the test centre and online *CAEL* score results interchangeably.

For the analyses reported in *Sections 6.2*, the four language component variables had to be treated as continuous variables because these statistical methods have not yet been fully adapted for ordered categorical variables. However, Rhemtulla, Brosseau-Liard, and Savalei (2012) noted that the analysis of covariance matrices treating ordered categorical variables as continuous is acceptable with even as few as 6–7 outcome categories. Our course of nine-point random variables for each of the four *CAEL* language components meets this condition to investigate the equality of covariance matrices. The limitation of constraint to continuous variables is released in *Sections 6.3* and *6.4*, an underlying variable (polychoric correlation) framework was used in those cases for the ordered categorical data. However, in the remaining chapters, it is important to note that this matter of needing to treat the band scores variables as continuous only arises again in *Chapter 9* because the statistical methods in *Chapters 7, 8*, and *10* allow for ordered categorical or binary outcome variables.

6.2 Equality of Covariance Matrices

This report focuses on the equality of the (unstructured) covariance matrices of the four band scores for the Test Centre and Online groups of test takers. The hypothesis is represented as:

$$\sum_{TestCentre} = \sum_{Online},$$

where $\sum_j$ denotes the dispersion matrix for group *j*. The hypothesis is that the equality of the two dispersion matrices was investigated using three approaches, all of whom treat the four components as continuous variables. *Section 6.2.1* reports on (i) Box's M-test for homogeneity of covariance matrices, including statistics based on eigenvalues of each of the covariance matrices, and (ii) a plot of the log-determinants of the covariance matrices. *Section 6.2.2*, on the other hand, reports a graphical analog of minimally sufficient statistics to investigate the equal covariance ellipse- using a scatterplot of the covariance ellipse.

6.2.1 Results – Statistical Tests of the Hypothesis of Equal Dispersion Matrices

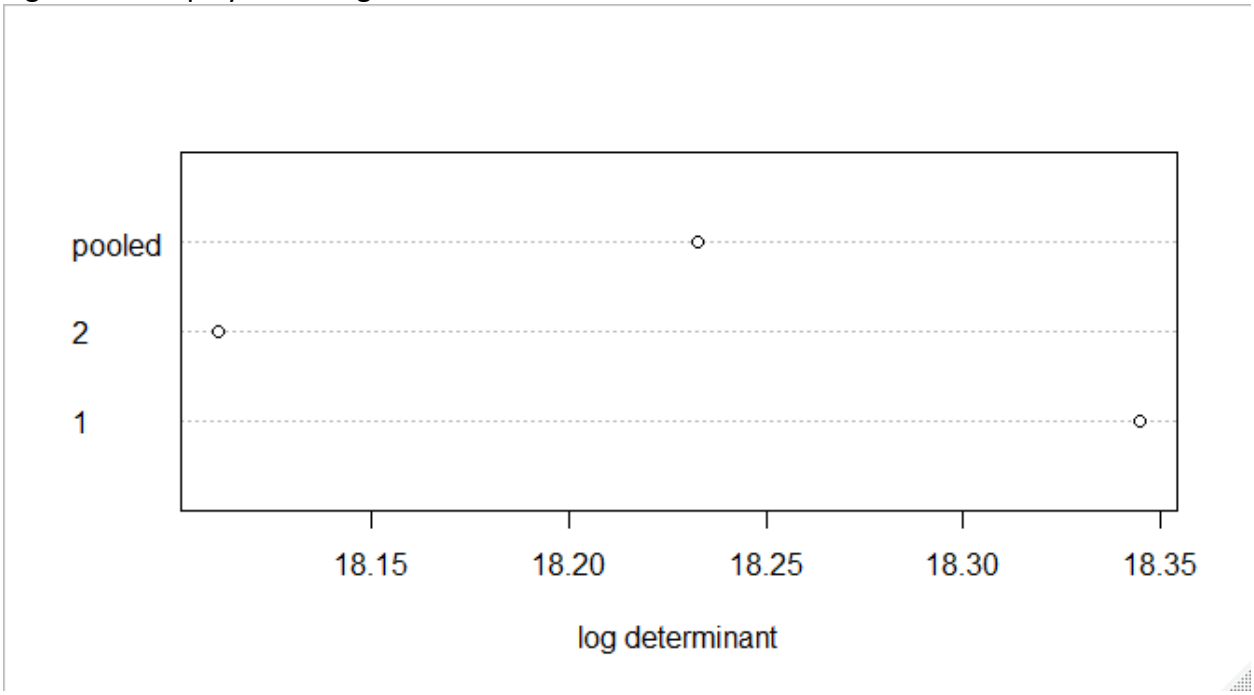
Box's M-test for Homogeneity of Covariance Matrices
Chi-Sq (approx.) = 15.698, df = 10, p-value = 0.1086

```
log of Covariance determinants:
       1        2   pooled
18.34501 18.11087 18.23273

Eigenvalues:
          1         2    pooled
1 479.27010 496.32528 488.04837
2  75.07033  72.57811  73.76108
3  57.68015  57.35297  57.55947
4  44.67449  35.50766  39.99145

Statistics based on eigenvalues:
                    1            2       pooled
product   9.271172e+07 7.335843e+07 8.286556e+07
sum       6.566951e+02 6.617640e+02 6.593604e+02
precision 1.813945e+01 1.628880e+01 1.724586e+01
max       4.792701e+02 4.963253e+02 4.880484e+02
```

Figure 6.1. Display of the log-determinants of the covariance matrices

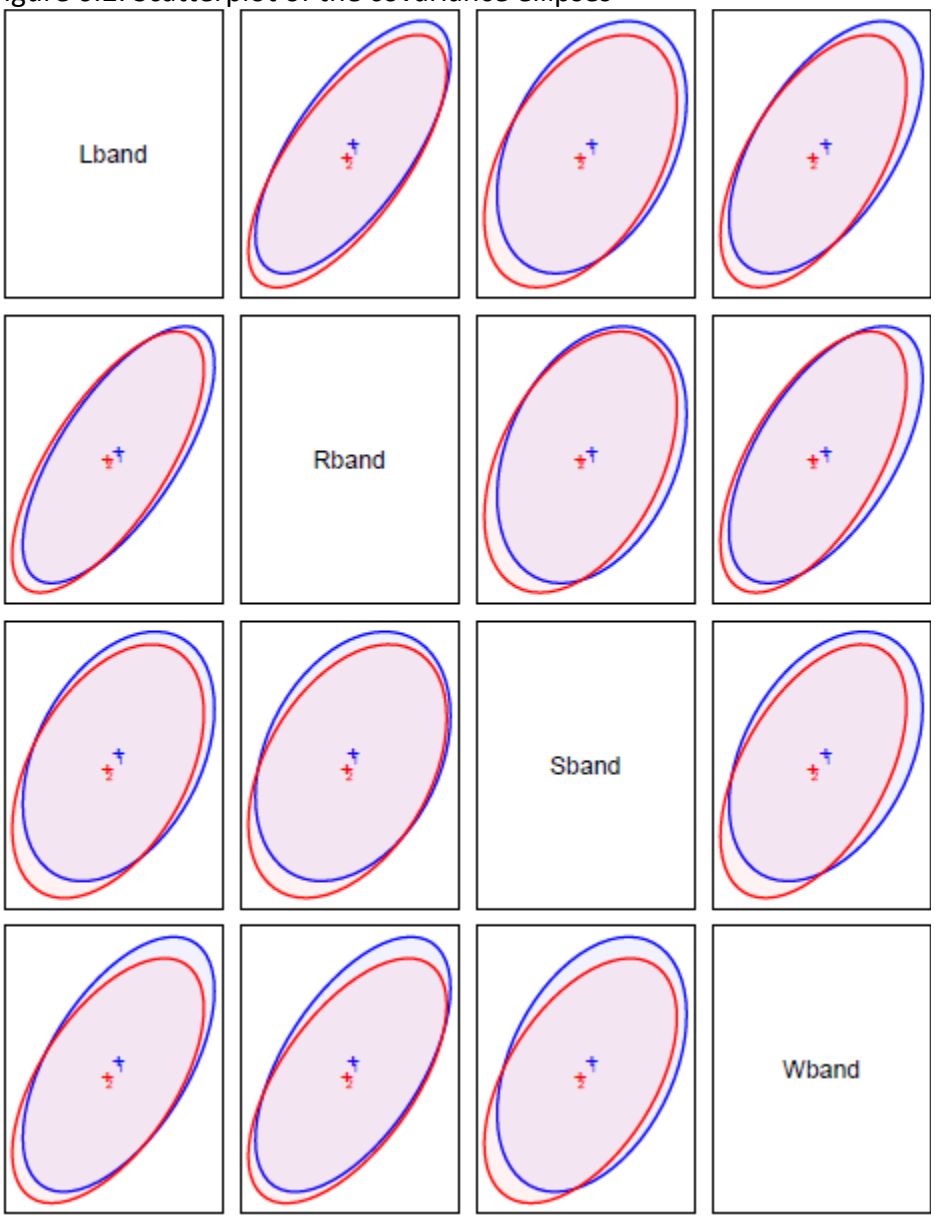


6.2.2 Results – Visualizing Tests of Equality for Covariance Matrices

These data visualization methods rely on characterizing statistical methods through elliptical geometry (e.g., Friendly, Monette, & Fox, 2011). Friendly and Sigal (2014) provide a thoroughgoing review of visualizing multivariate linear models. In the

multivariate plots, all the ellipses are centred at the origin to focus only on the size and shape of the within-group covariances to be directly compared visually (for *He* R code to create these scatterplots of the covariance ellipses, see, Fox, Friendly, & Monette, 2021).

Figure 6.2. Scatterplot of the covariance ellipses



6.3 Multi-Group Essential Unidimensionality

As widely accepted in the psychometric literature, a factor analysis framework was used to investigate the dimensionality assumption. As Schmitt (2011) notes, the first essential

step in factor analysis is determining the appropriate number of factors using parallel analysis methodology. Essentially, parallel analysis simulates datasets with the same number of variables and the sample size as the original data, but the variables are uncorrelated. These data sets generated from independent variables serve as a frame of reference to decide on the number of latent variables (factors) to include in the eventual factor analysis model. A correlation matrix is computed from the randomly generated dataset, and the eigenvalues of the correlation matrix are computed for each one. When the eigenvalues from the simulated random data are larger than the eigenvalues from our sample data, the factors are mostly random (Crawford, Green, Levy, Lo, Scott, Svetina, & Thompson, 2010; Horn, 1965).

The exhibit below lists the initial eigenvalues and the parallel analysis results of the band scores, with 50 simulated iterations, with reference to the average of the simulation and 95th percentiles. The eigenvalues for our sample were computed based on polychoric correlation matrices using *Mplus* version 8.4.

- One can see in *Table 6.1* that only the first eigenvalue meets the condition wherein the sample value is larger than either of the simulated reference values for both the test centre and online test taker groups.
- ***The parallel analysis results support the assumption of essential unidimensionality for <u>both</u> the test centre and online versions of the CAEL***.

Table 6.1. Results of the Parallel Analysis for Test Centre and Online Groups

A. Test Centre

EIGENVALUES FOR SAMPLE POLYCHORIC CORRELATION MATRIX			
2.830	0.599	0.329	0.243
AVERAGE OF THE EIGENVALUES FROM THE SIMULATED REFERENCE DATA			
1.082	1.023	0.969	0.927
95TH PERCENTILE OF THE EIGENVALUES FROM THE SIMULATED REFERENCE DATA			
1.128	1.058	0.995	0.957

Online

EIGENVALUES FOR SAMPLE POLYCHORIC CORRELATION MATRIX			
2.663	0.705	0.369	0.264
AVERAGE OF THE EIGENVALUES FROM THE SIMULATED REFERENCE DATA			
1.082	1.027	0.974	0.917
95TH PERCENTILE OF THE EIGENVALUES FROM THE SIMULATED REFERENCE DATA			
1.148	1.064	1.003	0.946

6.4 Multigroup Factor Analysis

In conventional item factor analysis, one may investigate *strict unidimensionality* using confirmatory factor analysis or *essential unidimensionality* using exploratory factor analysis methods. The latter is a sufficient condition for creating an aggregate total score

of the sort needed when scoring items using IRT or an observed total score in test operations.

Reise, Widaman, and Pugh (1993) and Wu, Li, and Zumbo (2007) describe the multigroup factor analysis statistical model and the implications of lack of measurement invariance for assessment data. A multigroup factor analysis model for a test taker (denoted i) belonging to group g (g=1,2, …, k) on observed variable j (j=1, 2, …, p) and latent variable l (l=1,2, …, q) can be written as follows:

$$y_{ijg} = \nu_{jg} + \sum_{l=1}^{q} \lambda_{j|g} \eta_{i|g} + \varepsilon_{ijg}.$$

Moreover, ν_{ij} denotes the intercept, $\lambda_{j|g}$ denotes the loading of the observed variable on the factor l, $\eta_{i|g}$ denotes the score of test taker i on factor l, and ε_{ijg} denotes the residual score of that person. The regression form of the factor model can be re-expressed with a latent outcome variable (akin to a probit model or logit depending on the estimator) by partitioning the continuous distribution of y^{*}_{ijg} into C categories (c = 0, 1, …, C-1) of the ordered categorical outcome variable.

Given our focus on characterizing latent variable dimensionality rather than test scoring, we compared three increasingly constrained confirmatory factor analysis (CFA) models with each other. Recall that we analyzed the band scores on the four language components to test for measurement invariance rather than the test items as described in *Chapter 4*.

Configural invariance model: A multi-group CFA model fitted without any equality constraints; represents a model of the same factor pattern across all groups; the functional form $\nu_{jg} + \sum_{l=1}^{q} \lambda_{j|g} \eta_{i|g}$ is the same across the groups, but the values of the coefficients are not constrained to be so.

Metric invariance model: A constrained version of the configural invariance model; however, the factor loadings $\lambda_{j|g}$ are equal across groups.

Scalar invariance model: A constrained version of the metric invariance model; however, the factor loadings $\lambda_{j|g}$ are equal across groups, and intercepts ν_{ij} are equal across groups.

6.4.1 Results of the Multigroup Factor Analysis

The model parameters and fit were obtained using *Mplus* version 8.4, using the (MLR) robust maximum likelihood estimator parameterized as a constrained latent class analysis in Mplus. In this parameterization, Mplus uses an EM algorithm (algo= EM for a pure EM algorithm) with accelerations using QN (quasi-Newton) or FS (Fisher scoring) steps when

EM is slow. In our case, a direct likelihood optimization (algo= ODLL) was also required because the robust Chi-square difference test was initially found to be negative.

There are no straightforward quantifiers of fit (e.g., CFI, TLI, or RMSEA) in this model parameterization and estimator; therefore, the fit of the configural model was assessed indirectly by examining comparative parameter values (relative to their standard errors).

As described by Wu, Li, and Zumbo (2007) and several other sources, to test metric invariance, one using a Chi-square difference test; if the difference test is not statistically significant, then metric invariance is established, and thus we can move to the next step, scalar invariance. If the difference test was statistically significant, then there is a lack of metric invariance, and there is no need to test strict invariance.

- Using the direct likelihood optimization, the test of the metric against the configural model resulted in a χ^2 (2) = 0.059, p = 0.9708, which supports metric invariance- equality of factor loadings.
- Unfortunately, the test for scalar invariance test was not interpretable using direct likelihood optimization, leaving that question unanswerable. We have a hint as to what may be happening because of the skewed observed order categorical variables resulting in many cases of sparse cells resulting in extreme values in the interative process.
    - Returning to the estimator using the pure EM algorithm to test the scalar against the metric models resulted in a statistically significant Chi-squared difference test χ^2 (29) = 59.64, p = 0.0007, which does not support scalar invariance.

6.5 Conclusions

The analyses of the covariance matrices consistently supported the equality of the covariance matrices of the band scores for the four language test components for the test centre and online. The findings support the conclusion of the essential unidimensionality of the band scores for the test centre and online groups. Equality of the factor loadings by mode of test administration is interchangeable in terms of their joint distributions of the band scores of the four language domains.

Although this evidence supports the concordance of the dispersion matrices for the two groups of test takers, the metric invariance would have directly impacted test scoring if the scores were item or task response data rather than band scores. The equality of the covariance matrices and the essential unidimensionality of the band scores from the two modes of test delivery are the most important findings supporting the statistical analysis in the remaining three chapters in this part of the report.

Chapter 7 Visualizing Comparability: Comparing Groups of Test Takers Using a Novel Kernel Smoothing Band Score Function

7.1 Overview- The main purposes of this chapter and overall conclusions

The chapter will begin with a brief description of the statistical framework focusing on the new kernel smoothed band score function, which is the new method's driving engine. Next, the kernel smoothed band score function is described as part of an algorithm to compare *CAEL* band scores from the four language components to investigate the comparability test centre and online. ***It is noteworthy that the focus of analysis is each of the four language component band scores because test users' decisions are made on band scores and not individual test item responses***. As described in *Chapter 3*, based on the logic and principles introduced in *Chapter 2*, the comparability of band scores is essential to the concordance of the use of the *CAEL*— this is central to treating the test centre and online *CAEL* results as interchangeable.

There are two main purposes of the report.

i. To introduce a novel statistical psychometric method to help analysts determine the comparability of the test centre and online versions of a test by visualizing the comparability with a novel graphical tool, the kernel Smoothing Band Score Function. Analysts and other interested readers will be able to compare test centre and online test performance.

ii. To apply this novel statistical methodology to the sample of test takers described and studied in *Chapter 5* who completed the *CAEL* at a test centre or online. The method allows for comparing the component (Reading, Listening, Writing, and Speaking) band scores using an intuitive graphical display.

Conclusions from Visualizing Concordance: Comparing Groups of Test Takers Using a Novel Kernel Smoothing Band Score Function

1. The novel graphical nonparametric method is introduced and demonstrated.
2. The ***test centre and online versions of the CAEL are shown to be concordant, fully comparable*** using the novel kernel smoothed band score function methodology - i.e., concordant for the mode of test delivery, reported gender of the test taker, gender by mode of test delivery, and self-reported first language.

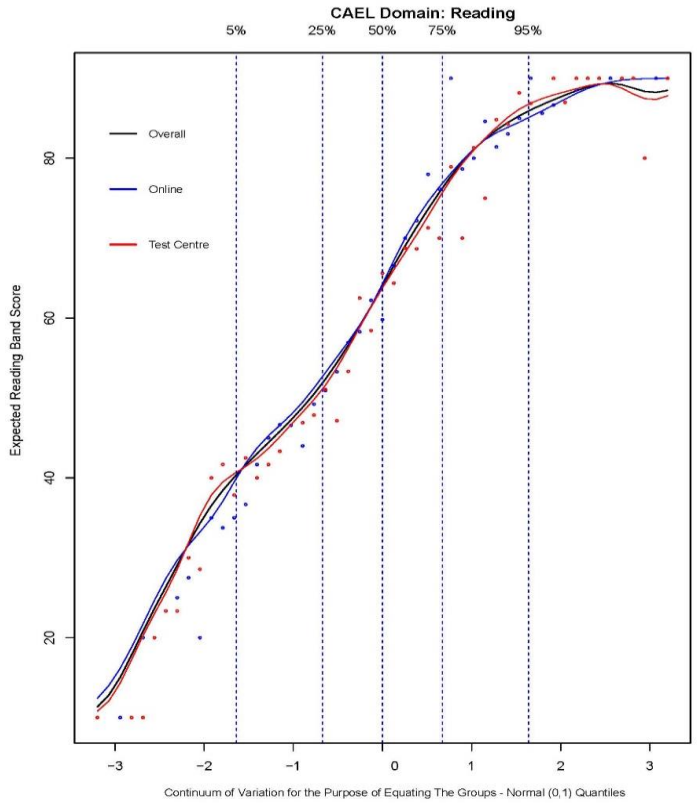
7.2 Visualizing Comparability of the *CAEL* Delivered at a Test Centre as Compared to Online

7.2.1 Description of the Graphs Used to Investigate the Comparability

The next exhibit is the graph displaying the comparability of the band scores of the reading component of the *CAEL* Delivered at a test centre as compared to online.

- The y-axis of this graph is the expected reading band score, which ranges from 10 to 90.
- The x-axis is a continuum of variation, a latent variable, if you wish, constructed from the four band scores arising from the *CAEL* language components. This continuum is constructed to facilitate comparing the band scores for those test takers at a test centre and online at various levels of the continuum. The continuum is interpreted by considering the 5%, 25%, 50%, 75%, and 95% quantiles of the continuum. The median is the 50% quantile, which is the point on the continuum where 50% of the distribution is below. It is important to note that the 5% and 95% quantile values mark those points on the distribution of the continuum of variation that demarcate the bottom and top 5% of the test taker data.

The coloured lines represent the test takers' performance on the reading component. ***One can see below that the lines overlap for nearly all the continuum of variation, indicating the CAEL reading component scores do not differ for test centre and Online test takers***.



Given the limited data at either end of the continuum, it is recommended that any apparent differences that have large uncertainty should be interpreted cautiously. It is safest to interpret the differences between the functions in the 5% and 95% quantiles because of the sparse data in that range of the equating continuum - the x-axis. Therefore, x-axis plots of the functions that follow will include the range of -2.0 to 2.0 of the *Normal* 0, 1 (mean of zero and standard deviation of one) quantiles of the estimated variation continuum $\hat{\vartheta}$.

The analyst is encouraged to compare the lines (which depict the novel kernel smoothed band functions for each group and overall) across the continuum of variation ***because the two groups of test takers are equated at each of those points***.

7.3 Four Noteworthy Strengths of the Novel Kernel Smoothed Band Function Approach

Briefly, there are four unique strengths of the novel kernel smoothed band score function methodology.

- ***Sample Size***. First, although this methodology can be used with large sample sizes (in the order of 5,000 or more test takers), the concept of band score modelling is founded on the data analytic principles of functional data analysis (e.g., Ramsay & Silverman, 2002, 2005; Simonoff, 1996) that was designed with moderate-to-small scale sample sizes in mind—for example, as small as 200 test takers per group.

- ***A Novel Lens into the Differences Between Groups Equated on a Continuum of variation*.** Second, other commonly used methods such as logistic regression differential item functioning (DIF) methods (Zumbo, 2003, 2008) model on the observed differences in response proportions of each, in this case, language component band score whereas the *kernel smoothed band score function methodology* focuses on the differences between the band score functions across a continuum of variation. The band score function traces the relation between an instrumental continuum of variation constructed to equate the groups being compared and the statistically expected band score level.
- ***A Graphical Display to Assist in the Interpretation***. The band score functions are not represented as functional equations of the sort one sees in typical linear or logistic modelling. Rather, they are of the complexity that requires that they be displayed graphically. These functions are meant to closely follow the data rather than forcing an equation, or function, onto the data space. Moreover, as Molenaar (2001) notes, using a kernel smoothing approach is *comparatively easier to explain because it relies on a graphical display* and performs better under replication.
- ***The Groups Being Compared are Equated Rather Than Trying to Statistically or Experimentally Match the Groups***. This form of equating is a direct application of the matching principle described in *Chapter 2*. In short, other approaches to investigating the comparability of language component level band scores for test takers at the test centre and online rely on a kind of "matching" to study the group differences. Groups of test takers matched on the key variable being measured should perform equally well on each language component. Similarly, experiments also match but do so by randomization- assuming full randomization is practically possible and that the homogenous sets of test takers are fully exchangeable. By focusing on determining the area between the curves (that is, the kernel smoothed band functions formally defined in *Section 7.4*) of the two groups at various levels of the continuum of variation, the two groups are equated integrating out the overall language performance.

7.4 Statistical Framework: Definitions, Assumptions, and Modeling

Given that a novel adaption of an empirical nonparametric method is used herein, we provide a brief description of the statistical methodology.

Definition 1: Let us define a band component score x_{ij} for each component, *i* (where *i*=1,2,3,4; reading, listening, writing, speaking), and test taker, *j*. It should be noted that the component (sample space) of the random variable x_{ij} is defined on an ordered

set I with one of k outcomes, which in our case is k = 10, 20, 30, .., 90 representing the band level score.

Definition 2: We aim to build on developments in functional data analysis, which deals with the analysis and theory of data in the form of functions, images and shapes, or more general objects. Functional data are inherently high- or infinite-dimensional. Building on the work of (Altman 1992, Härdle 1990, Simonoff, 1996) and particularly Ramsay (1991), the central idea of this method is to specify a kernel smoothing model (which we denote as the ***band score function***) which describes the conditional probability of test taker, *j*, earning a band score $x_{ij} = m$ within component *i*,

$$\text{Band Score Function : = } \quad P_g(I = m|\vartheta) = \sum_{j=1}^{n} w_{ij}(\vartheta) x_{ij}, \tag{1}$$

where ϑ denotes a continuum of variation expressed as a normal quantile whose instrumental purpose is to compare the band score functions for the g groups of test takers such as those taking the *CAEL* online or others taking the *CAEL* at a test centre. Likewise, w are the so-called Nadaraya-Watson weights defined in terms of a kernel function, which is governed by (a) the distance between the point where the individual is located on the underlying continuum ϑ_j, and the point where the curve is being estimated, ϑ_q; (b) a bandwidth parameter, h, which determines the sensitivity with which the value of the curve is estimated; and (c) the type of kernel, *K(u)*, used.

In essence, kernel smoothing takes a weighted average at each point of the band score function, where the kernel function determines the weights. Furthermore, it should be noted that kernel smoothing does not insist upon monotonicity, in our case concerning ϑ_j; therefore, we use kernel smoothing to investigate the form of the band score function. Importantly, the choice of bandwidth value denoted *h* controls the trade-off between bias and sampling variation. As such, the bandwidth is also described as the smoothing parameter, controlling the amount of smoothness (in terms of bias-variance trade-off). As Ramsay and Silverman (2005) remind us, low values of *h* produce estimated functions with large variance and small bias, and high values of *h* produce estimated functions with a small variance but large bias. However, it is important to keep in mind that the recommended choice of bandwidth includes considering the sample size. In our multisample case, this minimum sample size is the number of test takers in each sub-group. The estimation and plotting were conducted using KernSmoothIRT (Mazza, Punzo, & McGuire, 2014) in R, using the Silverman rule for bandwidth value.

7.4.1 Assumption of Essential Unidimensionality

As widely accepted in the psychometric literature, a factor analysis framework was used to investigate the dimensionality assumption using parallel analysis methodology. As described in *Chapter 6*, the initial eigenvalues and the results of the parallel analysis of the band scores ***support the assumption of essential unidimensionality***.

7.5 Using Kernel Smoothed Band Score Function to Compare the Band Scores for Component Functions of the *CAEL* (Online) and *CAEL* (Test Centre)

As the ***band score function***, equation (1) describes the probability of earning band score option *m* from language component *i* is directly determined by the observed values x_j obtained from *n* test takers. It is these observed band score values x_j, which figure directly in estimating the band score function. The band score function expressed in equation (1) has the distinct advantage of kernel smoothing wherein only requiring minimum *model* assumptions are necessary, and equation (1) is, in essence, just a weighted average of the data. The minimal assumptions are that a continuum of variation ϑ is of interest to the analyst wherein:

1) there is empirical evidence that the band component score x_{ij} is sufficiently unidimensional to justify a single continuum of variation,
2) the choice of the scale of the sample estimated continuum of variation $\hat{\vartheta}$ is arbitrary, since in this context, only rank order considerations make sense (Bartholomew 1983; Ramsay 1991, p. 614), and
3) typically, the continuum of variation, ϑ is interpreted as the underlying latent trait (e.g., language competency) which the test attempts to measure. However, in use herein, the continuum of variation's primary instrumental purpose is to facility equating test takers from the *g* groups in terms of ϑ so that the qualitative differences (the band scores) between these groups can be examined.

> The advantage of this method compared to conventional DIF methods (e.g., logistic regression DIF) is that we are not "covarying" or matching the test takers, per se, but rather we have constructed a (latent) continuum of variation for the instrumental purpose of being able to equate and compare test takers at various levels of a continuum constructed by the four components tested by *CAEL*

In the current use, the purpose of the continuum of variation, ϑ, is not to locate test takers on that continuum and provide test scores, which requires conceptual analysis and justification for the individual differences on the continuum of variation. Instead, the sole purpose of ϑ is to equate test takers so who obtain similar scores on ϑ with the expectation that they should, on average earn attain individual band scores equally. If differences in attaining band score levels are observed in individuals with equivalent levels ϑ, it can be argued that there is a qualitative difference between them. Only after equating individuals in terms of their continuum of variation ϑ can qualitative (item response) differences between groups be examined.

As Zumbo (2003, 2007b, 2013, 2015) notes, in its essence, approaches like the one described herein are focused on determining the area between the curves (or, equivalently, if this were IRT DIF, comparing the IRT parameters) of the two groups. Therefore, it is noteworthy that, unlike the conventional DIF methods involving contingency tables or regression modelling methods, the current approach does not match the groups' conditioning based on the total score. The question of "matching" only comes up if one computes the difference function between the groups conditionally, as in the MH or LogR DIF approaches.

Comparing the estimates of the band score function in equation (1) is an unconditional analysis because it implicitly assumes that the continuum of variation has been, in a sense, "integrated out." The mathematical expression "integrated out" is commonly used in some DIF literature and is used in the sense that one computes the area between the functions across the distribution of the estimated continuum of variation $\hat{\vartheta}$. The emphasis on $\hat{\vartheta}$ is important at this point because it highlights that the empirical band score function,

$$\mathrm{P}_g(I=m\left|\hat{\vartheta}\right.)=\sum_{j=1}^{n} w_{ij}\left(\hat{\vartheta}\right)x_{ij},$$

may be compared in ranges of the continuum of variation where there are few data points and hence wide standard errors. Caution should be exercised in interpreting the band score function in the upper and lower tails of the distribution of the estimated variation continuum $\hat{\vartheta}$.

7.6 Minimum Sample Sizes for the Various Graphical Comparisons

As is widely understood in the psychometric literature, an advantage of kernel smoothing methods is relatively smaller sample size requirements compared to 2-PL and 3-PL IRT methods and the use of graphical visualization that facilitate straightforward interpretation. However, two interrelated points temper these advantages.

- First, one should avoid interpreting differences between the group empirical band score functions below 5% and above the 95% quantiles because there is less data

for estimating the curve in these regions. Thus, there is less precision in the estimates.

- Second, the recommendation in the research literature for the minimum sample size is not univocal but is in the range of 200-400 test takers. That would translate to a minimum of 200 to 400 test takers per sub-group in any sub-group comparison in our setting. For example, Ramsay and Silverman (2002, 2005) and Simonoff (1996) state that these kernel smoothing methods are designed for

> Please ***interpret with caution*** any differences between the group empirical band score functions below the 5% and above the 95% quantiles because there is less data for estimating the curve in these regions and thus there is less precision in the estimates.
> Any pointwise confidence bands of the difference between the function would be substantially larger in these regions.
>
> This problem in the tails of the functions exacerbated as the test taker sample is sub-divided into further (smaller) sub-groups such as delivery mode by the gender of the test taker.
>
> **It is safest to interpret the differences between the 5% and 95% quantiles.**

sample sizes as small as 200 test takers per group. Ramsay (2000) and Molenaar (2001) state that this psychometric technique is appropriate to be used with small sample sizes of a minimum of 300-400 test takers (Molenaar, 2001; Ramsay 2000). The differing values are partly related to the fact that sample size is related to the bandwidth value described in *Section 7.4*.

7.6.1 Study Data

As described in *Chapter 5*, the focus of the analysis was the *CAEL* test taker data between June and October 2020. As one can see in Tables 5.1 and 5.2, the following comparisons meeting the minimum requirement of 200 test takers:

i. the main effects of mode of test delivery,
ii. the main effect of gender,
iii. the interaction of mode of test delivery by gender, and
iv. the main effect of the first language.

However, neither the mode of test delivery by the first language interaction nor the three-way interaction meets the condition of a minimum of 200 test takers per cell of the design.

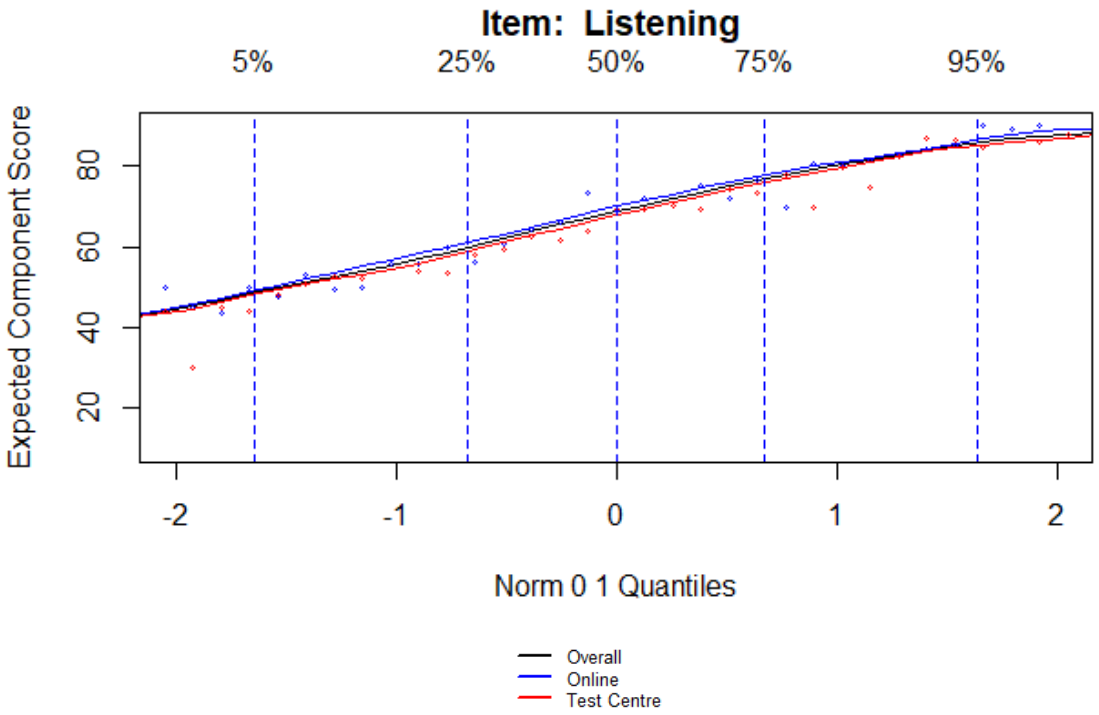
7.7 Graphs of the Four Component Band Scores - Test Centre Compared to Online (Main Effect of Mode of Test Delivery)

The kernel smoothed band score function methodology focuses on the differences between the band score functions across a continuum of variation. The band score function traces the relation between an instrumental continuum of variation constructed to equate the groups being compared and the statistically expected band score level. A non-discernible to small difference (or area) between the lines is evidence for concordance for any graph of the band score functions.

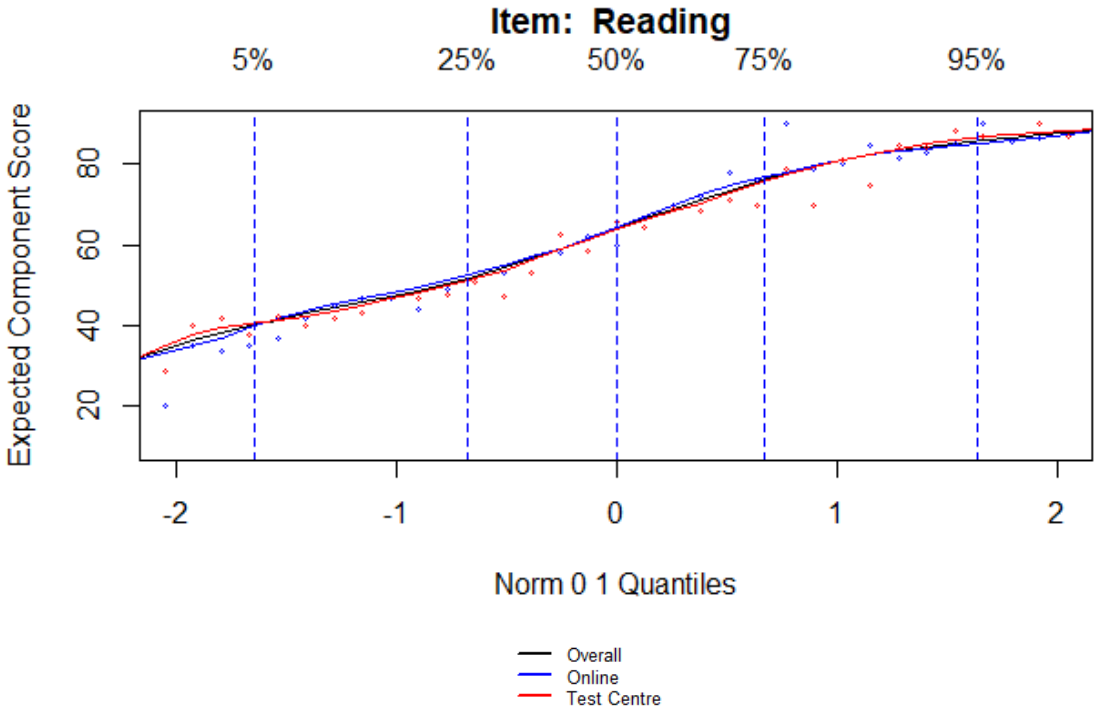
Please recall that caution should be exercised in interpreting the band score functions in the lower tail of the distribution, below the 5% quantile, of the estimated variation continuum because of the sparse data in that range of the equating continuum. Therefore, the x-axis plots of the functions that follow will include the range of -2.0 to 2.0 of the *Normal* 0, 1 (mean of zero and standard deviation of one) quantiles of the estimated variation continuum $\hat{\vartheta}$.

The remainder of this section presents the graphs that allow the investigation of concordance for the mode of test delivery, gender of the test taker, *the comparison of concordance* for the mode of test delivery for male and female test takers, and finally, the concordance for the test taker's first language as categorized by language family.

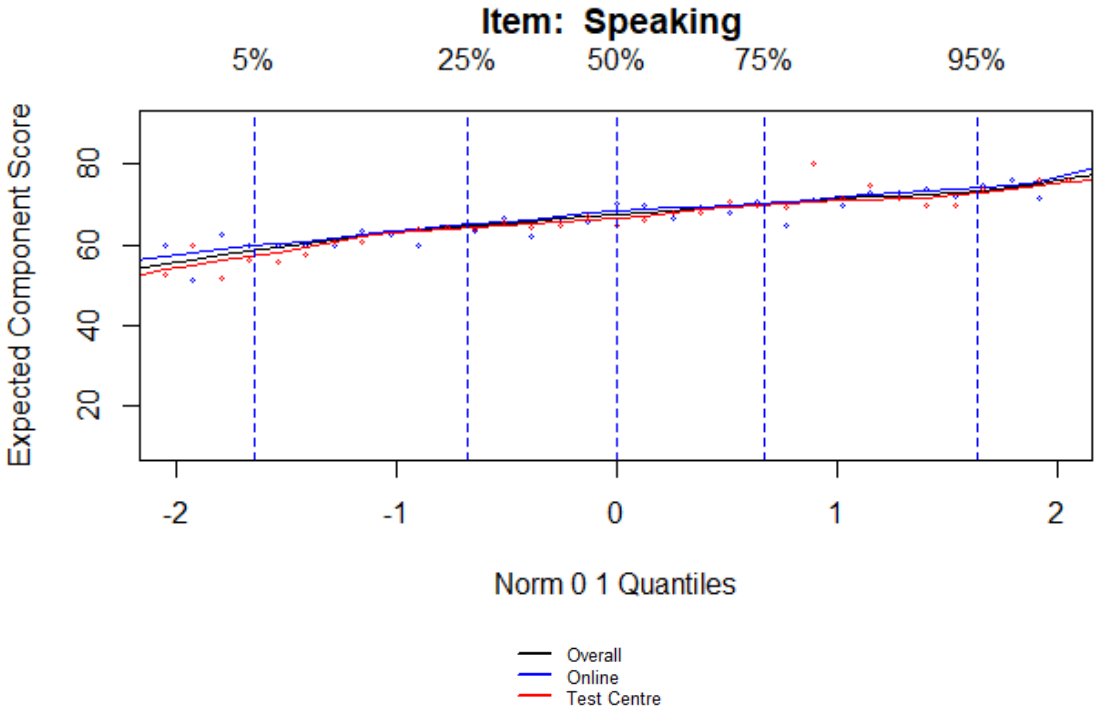
7.7.1 Comparability of the Band Scores on the Listening Component



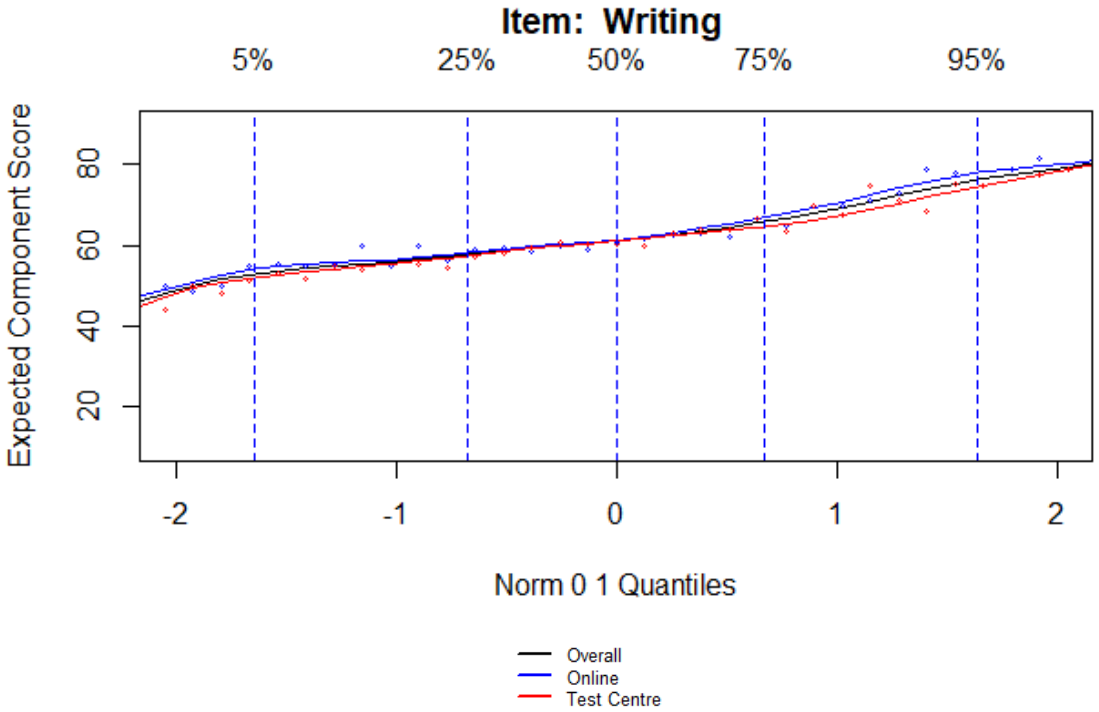
7.7.2 Comparability of the Band Scores on the Reading Component



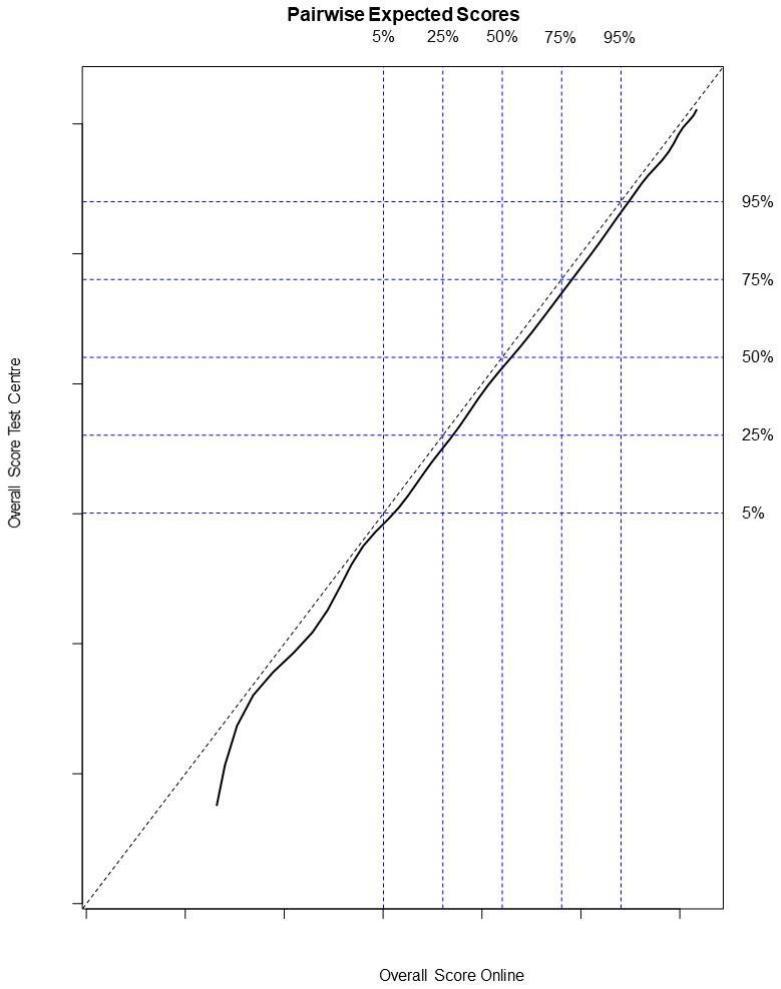
7.7.3 Comparability of the Band Scores on the Speaking Component



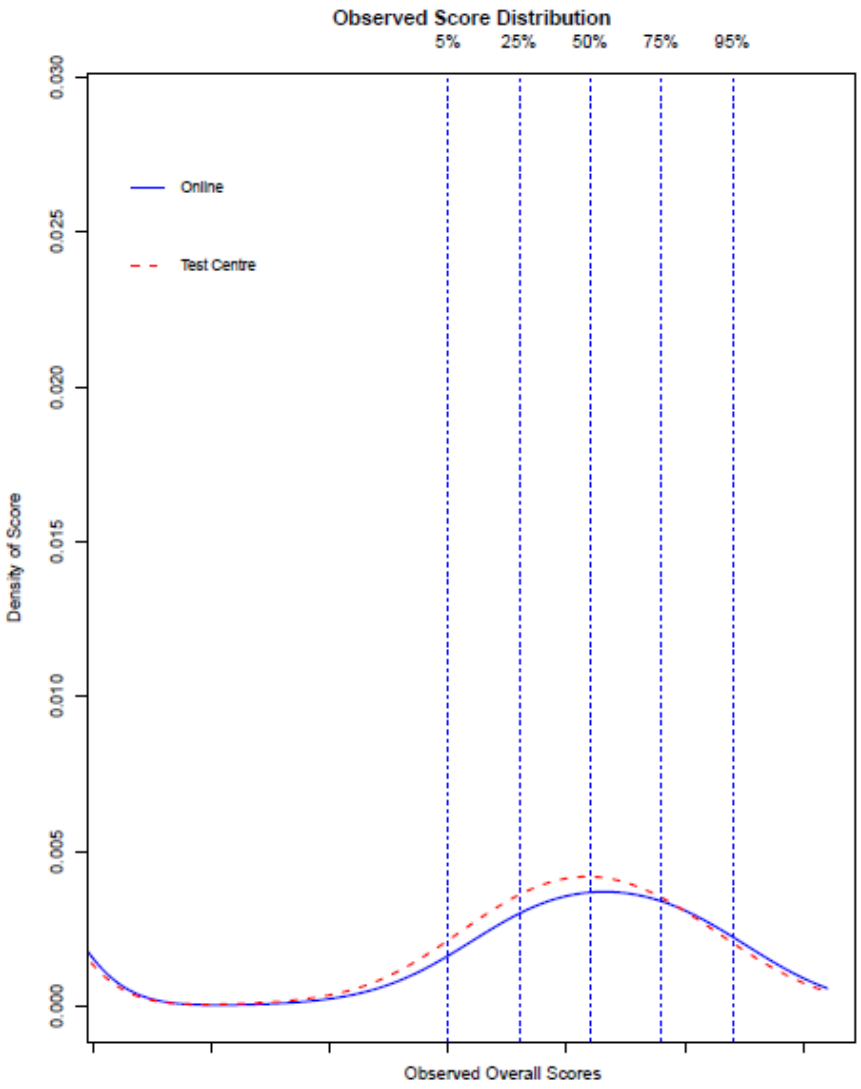
7.7.4 Comparability of the Band Scores on the Writing Component



7.7.5 Overall Comparability of the Band Scores

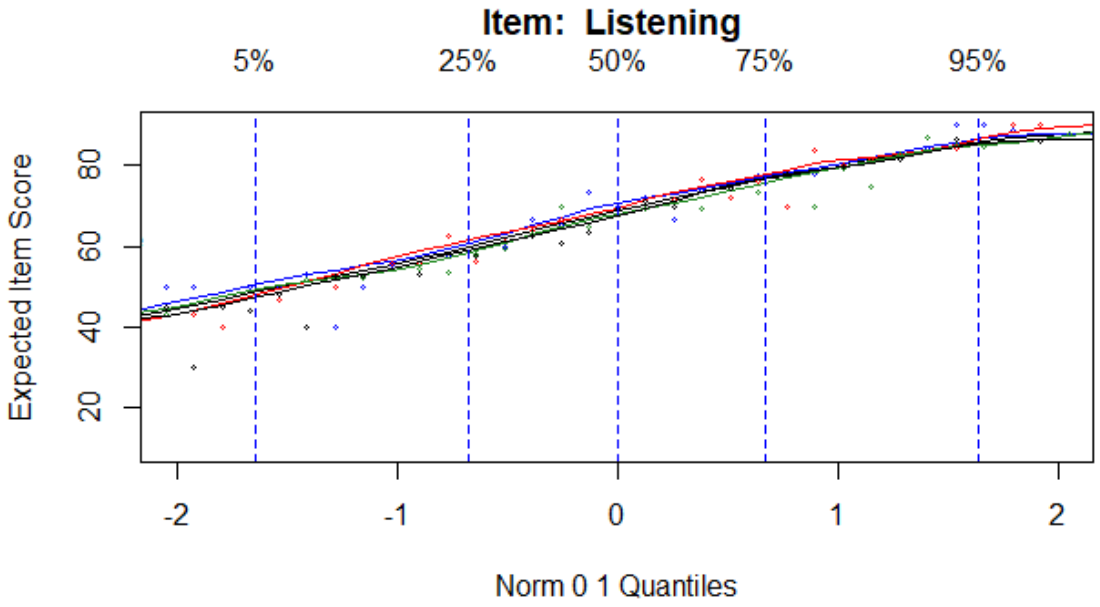


7.7.6 Comparability (Density) Distributions of the Observed Overall Band Score



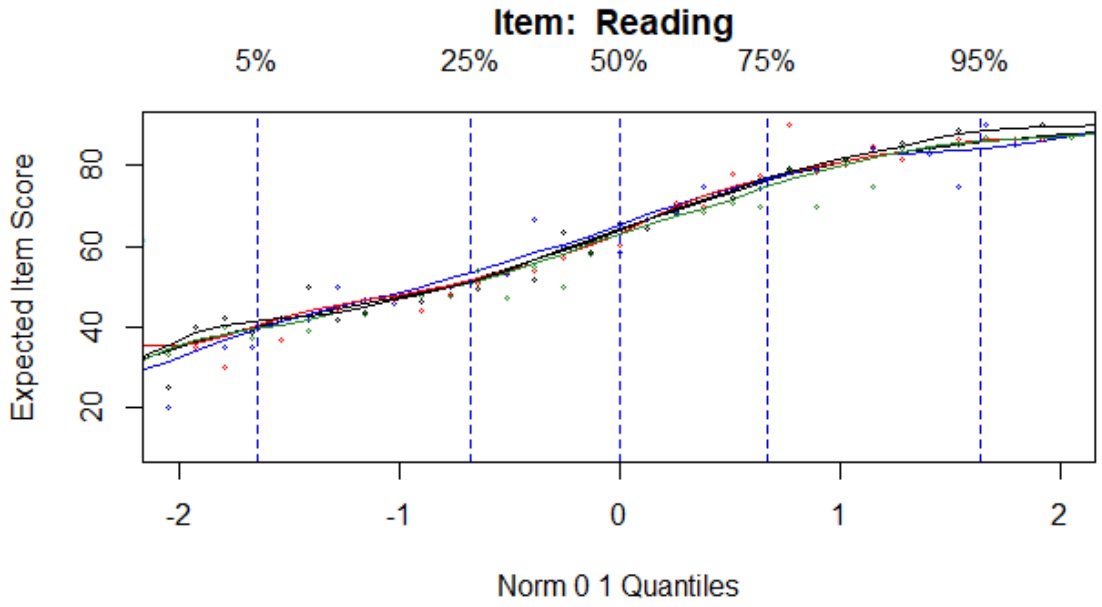
7.8 Graphs of the Four Component Band Scores - Test Centre Compared to Online for Female and Male Test Takers (Mode by Gender)

7.8.1 Delivery Mode by Gender of Test Taker Comparability of the Band Scores on the Listening Component



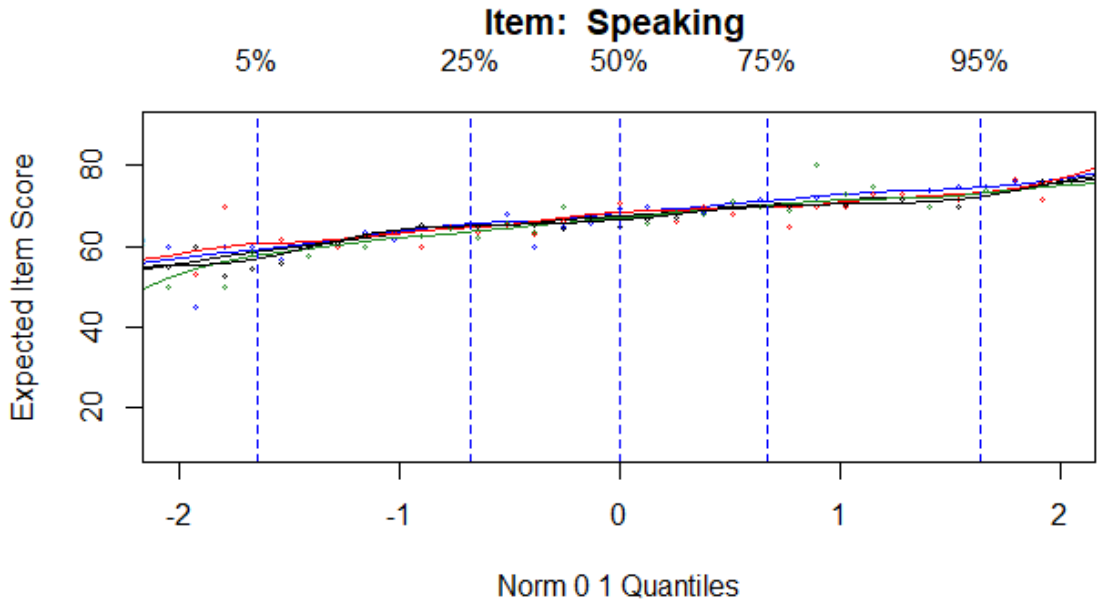
---- Overall
---- Female Online
---- Male Online
---- Male Test Centre
---- Female Test Centre

7.8.2 Delivery Mode by Gender of Test Taker Comparability of the Band Scores on the Reading Component



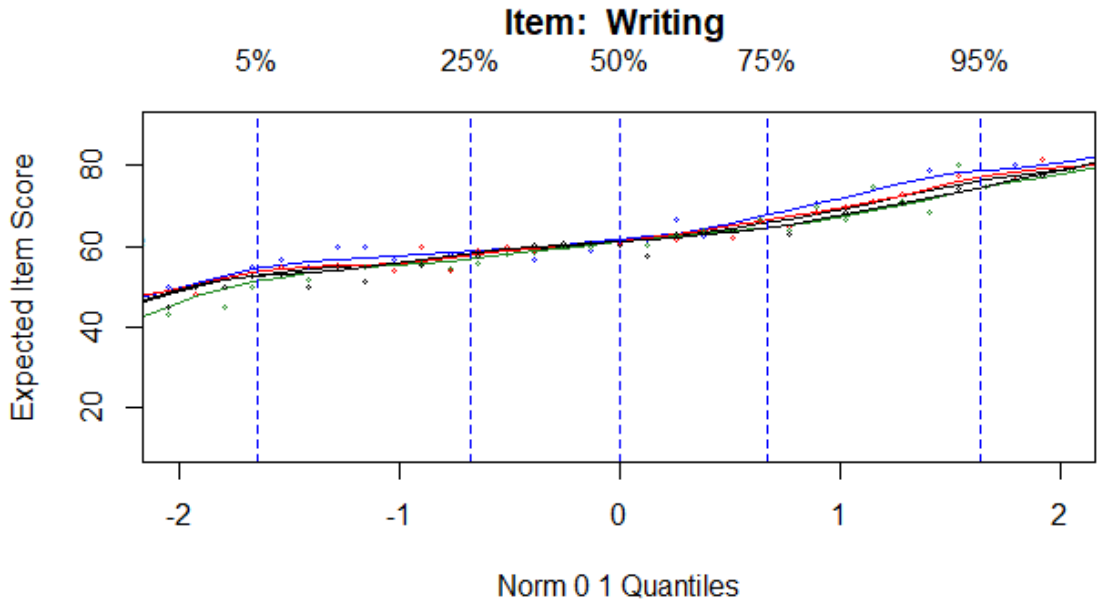
---- Overall
---- Female Online
---- Male Online
---- Male Test Centre
---- Female Test Centre

7.8.3 Delivery Mode by Gender of Test Taker Comparability of the Band Scores on the Speaking Component



---- Overall
---- Female Online
---- Male Online
---- Male Test Centre
---- Female Test Centre

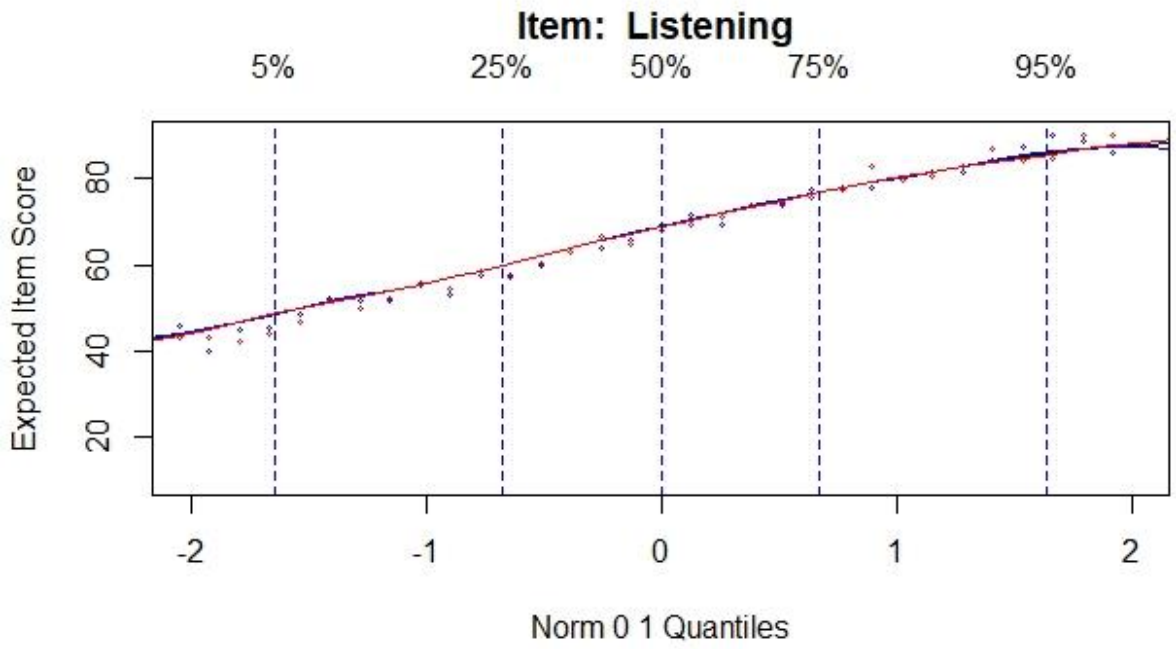
7.8.4 Delivery Mode by Gender of Test Taker Comparability of the Band Scores on the Writing Component



---- Overall
---- Female Online
---- Male Online
---- Male Test Centre
---- Female Test Centre

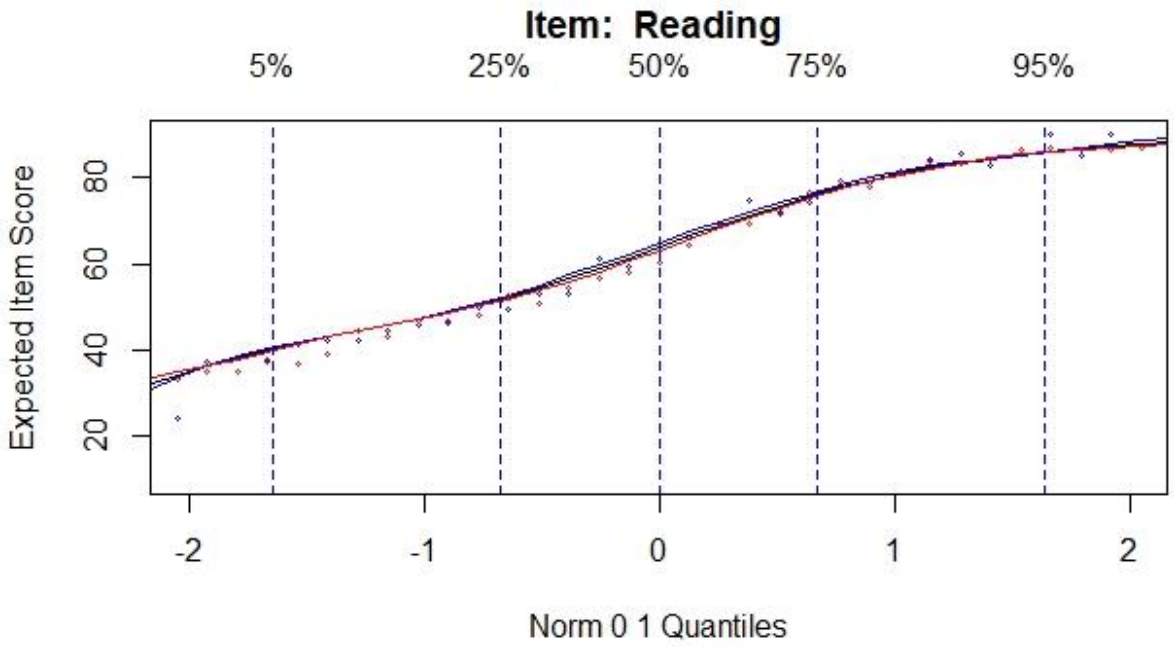
7.9 Graphs of the Four Component Band Scores – Gender DIF for Female and Male Test Takers (Main Effect of Gender)

7.9.1 Comparability of the Band Scores by Gender of the Test Taker for the Listening Component



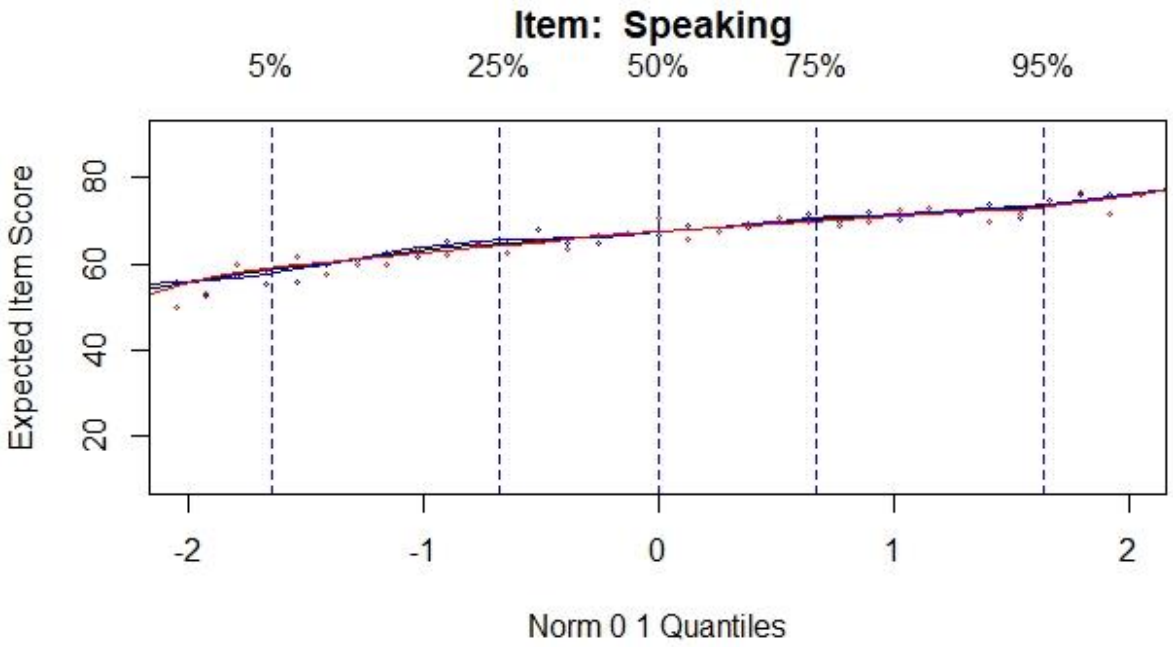
---- Overall
---- Female
---- Male

7.9.2 Comparability of the Band Scores by Gender of the Test Taker for the Reading Component



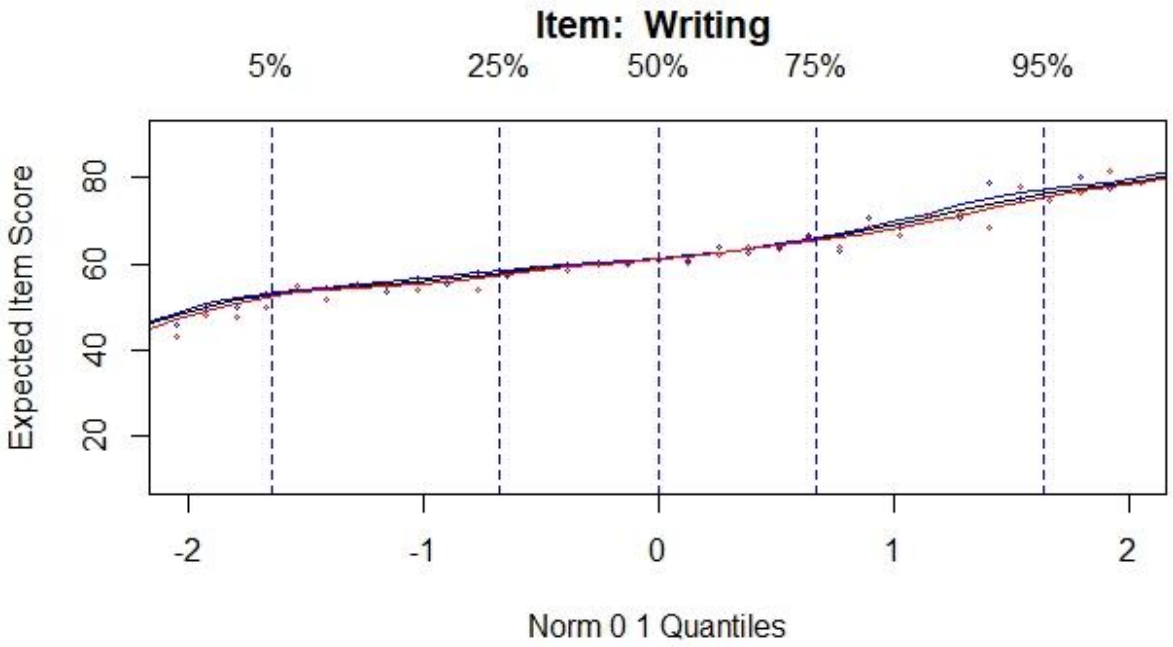
---- Overall
---- Female
---- Male

7.9.3 Comparability of the Band Scores by Gender of the Test Taker for the Speaking Component



---- Overall
---- Female
---- Male

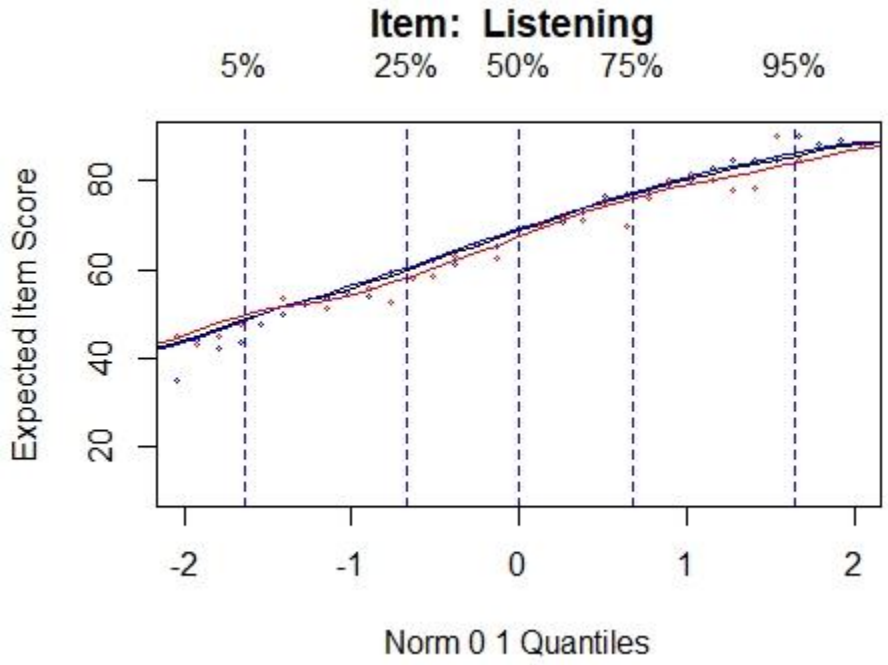
7.9.4 Comparability of the Band Scores by Gender of the Test Taker for the Writing Component



---- Overall
---- Female
---- Male

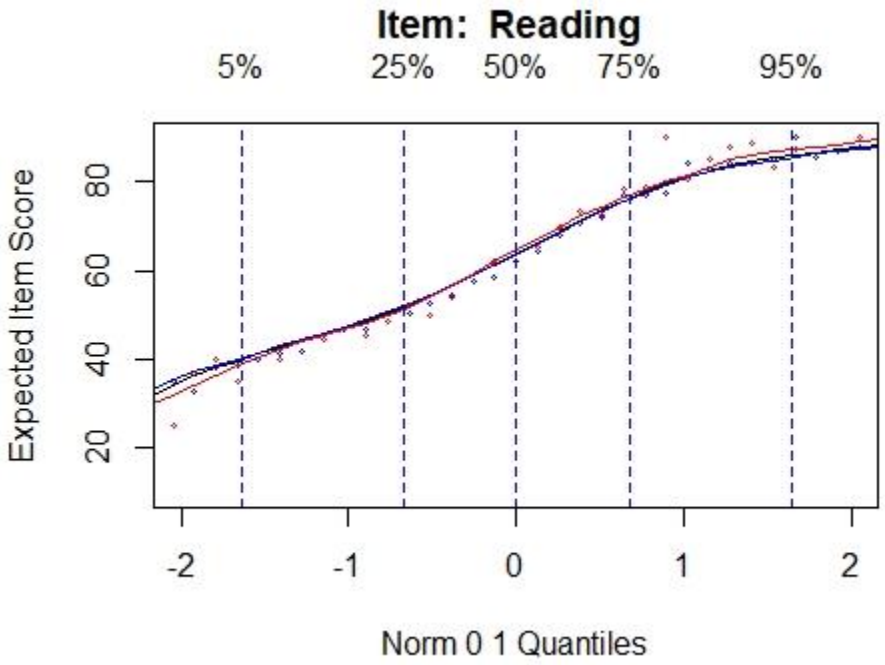
7.10 Graphs of the Four Component Band Scores – Test Taker's Report First Language (Language Family) DIF (Main Effect of First Language)

7.10.1 Comparability of the Band Scores by First Language of the Test Taker for the Listening Component



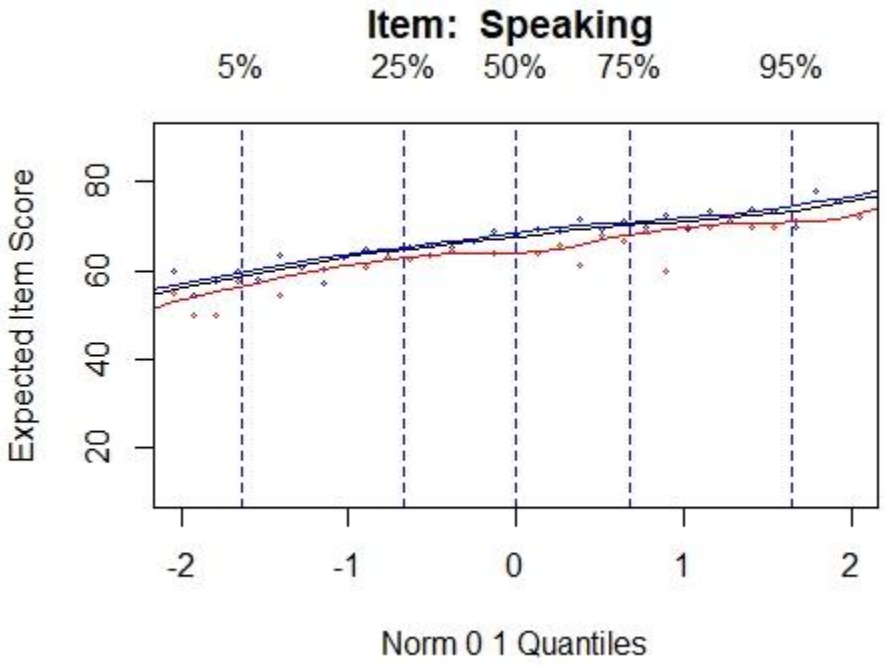
---- Overall
---- Indo-European
---- Non-Indo-European

7.10.2 Comparability of the Band Scores by First Language of the Test Taker for the Reading Component



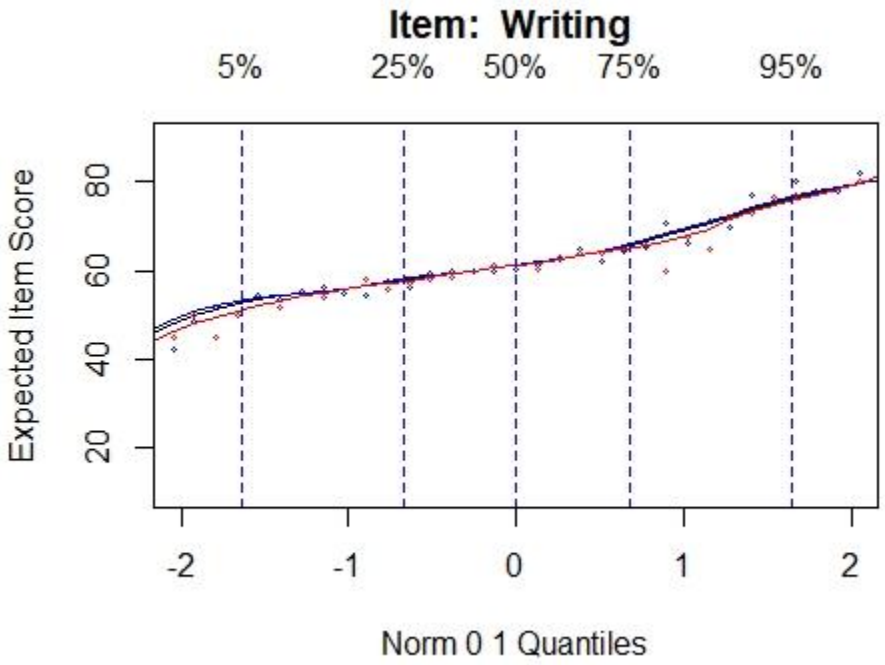
---- Overall
---- Indo-European
---- Non-Indo-European

7.10.3 Comparability of the Band Scores by First Language of the Test Taker for the Speaking Component



---- Overall
---- Indo-European
---- Non-Indo-European

7.10.4 Comparability of the Band Scores by First Language of the Test Taker for the Writing Component



---- Overall
---- Indo-European
---- Non-Indo-European

7.11 Conclusions

Sections 7.7 to *7.10* display the concordance findings that the kernel smoothed band functions, equated by the intended-to-be-assessed language ability, are indistinguishable for the mode of test delivery, reported gender of the test taker, gender by mode of test delivery, and self-reported first language.

The test centre and online versions of the *CAEL* are shown to be concordant, fully comparable using the novel kernel smoothed band score function methodology that allows the analyst to compare (equated) band score performance of test takers at various levels of *CAEL*'s language components.

Chapter 8 Generalized Linear Model Approaches – DIF Analyses of Test Centre vs Online *CAEL* Test Performance

8.1 Overview and Conclusions

Differential item functioning (DIF) is one technique used to help ensure the fairness of tests administered in different modalities- such as our case of Test Centre vs Online.

DIF is a statistical finding wherein test takers of equal language competency as assessed by *CAEL* while belonging to distinct subpopulations (e.g., those who take the *CAEL* at a Test Centre and those who take it Online) perform in detectably different ways on a band score for one of the four language domains.

The generalized linear regression model methods for DIF detection provide a common statistical framework for varied DIF questions (Gadermann, Chen, Emerson, & Zumbo, 2018; Zumbo, 2008).

It is important to note that the DIF analyses focused on the band scores (of the four language domains). The focus on band scores was important because test users' decisions are made on band scores and not individual test item responses. **In the end, it is the comparability of band scores that is essential to the concordance of the use of the *CAEL*— this is central to treating the Test Centre, and Online *CAEL* results as interchangeable**.

Overall Conclusions

The DIF analyses showed that test takers of equal language competency as measured by the CAEL would earn statistically similar band scores whether they took the CAEL at a Test Centre or Online.

DIF analyses showed negligible or statistically nonsignificant effects of delivery mode. This was further supported when taking test taker gender and self-reported first language family (Indo-European vs Non-Indo-European) into account.

8.2 Statistical Method

A GLIM model was specified to test the concordance by mode of test delivery and the moderating effect of gender and first language on the mode of test delivery. The latter moderators allow us to investigate if the mode of test delivery concordance differs based on gender or first language. In addition, as suggested in *Chapter 5*, the test taker's age will also be covaried.

8.2.1 Minimum Sample Size

A recent paper by Scott et al. (2009) demonstrated by Monte Carlo simulation that the power and Type I error rates of the ordinal logistic regression (GLIM) DIF method that a minimum of 200-300 test takers per group for adequate (80%) statistical power for a test of uniform DIF and considerably more for non-uniform DIF. Scott et al.'s and Herrera and Gómez's (2008) recommendations state a minimum of 500 test takers in total for the research study, with a minimum of 250 test takers in either group.

In addition, if multiple matching and more complex GLIM DIF models are used, Bujang, Sa'at, and Tg Abu Bakar Sidik, and Lim CJ (2018) recommend that the overall sample size of (100+50*p*), where *p* refers to the number of independent variables in the final GLIM model.

8.2.2 Specification of the family of GLIM DIF Models

As Zumbo (2007, 2008, 2013) notes, the question of "matching" arises if one computes the difference function between the groups conditionally, as in the MH or logistic regression DIF approaches. DIF methods statistically match by covariance analysis by conditioning on the groups' total score on a test.

Zumbo (1999) extended the logistic regression procedure for polytomous scored items or test band scores, using a proportional-odds cumulative logit regression model (see, e.g., McCullagh, 1980). In this model, the logit for a person j to respond to the scoring category k or below is expressed as:

$$\text{logit}\left[P\left(Y_j \le k \middle| X_j, G_j\right)\right] = \log\left[\frac{P\left(Y_j \le k\right)}{P\left(Y_j > k\right)}\right]$$
$$= a_{k-1} + b_1 X_j + b_2 G_j + b_3\left(X_j G_j\right).$$

The additional notation for this proportional-odds cumulative logit regression model is:

- Y_j denotes the item response for person *j*;

- k is the response category;
- X_j is the matching variable(s) and covariates for person j;
- G_j is the design matrix for groups denoting the differences in group membership;
- XG is the interaction of matching variable(s) and the design matrix;
- a_{k-1} denotes the intercept terms and can be interpreted as the log-odds of falling into or below category k when the predictor variables equal zero; and
- the b_1, b_2, and b_3 are the regression coefficients.

It should be noted that $P(Y_j \leq k)$ is the probability of person j responding less than or equal to category k, and likewise, $P(Y_j > k)$ is the probability of person j responding greater than category k. One can interpret this logistic regression model as a linear regression of predictor variables on an unobservable continuously distributed random variable y^*. Zumbo (1999) notes that one can get an R-squared index for ordinal logistic regression (see Latila, 1993; McKelvey & Zavoina, 1975), and he proposed using that index as an effect size quantifier.

Several points are noteworthy in the DIF analyses of Test Centre vs Online *CAEL* Test Performance.

- The findings in *Chapter 6* regarding the essential unidimensionality and metric invariance support the use of the band scores to interpret the logistic regression DIF tests.
- The frequency plots in *Section 8.4* show that the ordinal logistic regression model of the family of generalized linear models is most appropriate for the DIF analyses- the band scores are the outcome (DV) in these models.
- Multiple matching, rather than a total test score, was used because the object of analysis was a band score on one of the four language domains. The multiple matching involved covarying on the estimated true scores of the left-out language domains- i.e., the three language domains that complement the outcome or y-variable. For example, suppose the listening band score was the object of DIF analysis. In that case, the estimated equated true scores for reading, writing, and speaking serve as covariates.
    - In addition, as noted in the conclusions of *Chapter 5*, given the findings regarding test taker comparability, it would be prudent to consider age among the key matching covariates when fitting and testing concordance using the (generalized) linear models.

  Therefore, the specified model will have four covariates: three estimated true scores of the left-out language domain and age.

- In addition to the main effect of mode of test delivery, the main effects of gender and first language will be included to allow for the specification of tests of the moderating effect of gender and first language on the mode of test delivery.

- The resulting GLIM model includes nine independent variables:
    - three estimated true scores of the left-out language domain and age (4)
    - main effects of mode of test delivery, gender, and first language (3)
    - two two-way interaction terms for the moderating effects (2).
    - The GLIM model for the listening component would be:

$$\text{logit}\left[P\left(Y(\textit{Listening})_j \le k \middle| R_j, S_j, W_j, Age_j, M_j, G_j, FL_j\right)\right] = \log\left[\frac{P\left(Y_j \le k\right)}{P\left(Y_j > k\right)}\right]$$
$$= a_{k-1} + \left(b_1 \mathrm{R}_j + b_2 S_j + b_3 W_j + b_4 Age_j\right) + \left(b_5 M_j + b_6 G_j + b_7 FL_j\right) + \left(b_8 M_j G_j + b_9 M_j FL_j\right)$$
$$= a_{k-1} + (\text{matching covariates}) + (\text{uniform DIF of M, G, FL}) + (\text{moderated M DIF}).$$

In the GLIM model[2], the first left parentheses contain the matching covariates denoted R, S, and W for the reading, speaking, and writing estimated true scores of the left-out language domains; age denotes the test taker's age in the first set of parentheses. The middle parentheses contain M, G and FL, which denote the contrast vectors for the uniform DIF effects of mode of test delivery, gender, and first language. The right parentheses reflect the gender (G) and first language (FL) moderation of the uniform DIF effect of mode of test delivery (M).

- Applying the Bujang et al. (2018) criteria for minimum sample size, the GLIM DIF model requires a minimum of 550 test takers for adequate model fit, which is easily met and exceeded by the data at hand described in *Chapter 5*.
- The GLIM model described above was fit for each of the four language domains' band scores separately.
- Using the modelling strategy first described in Zumbo (1999) and extended by him in 2008, a statistically significant ($p<.05$) change in the chi-square test statistic with the appropriate degree of freedom for the main effects signals the presence of uniform DIF and the three-degree freedom for non-uniform DIF.

As Zumbo (1999, 2008) described and recently in Gadermann, Chen, Emerson, and Zumbo (2018), the generalized linear model DIF involves three models: model 1 includes only the matching covariates; model 2 adds the uniform DIF effects, and model 3 adds the moderated DIF effects.

To test for the statistical effects, in addition to statistical significance, DIF results were classified into three categories by Jodoin and Gierl's (2001) effect size criteria: category A, **negligible or nonsignificant** (change in Nagelkerke R-squared **< 0.035**); category B, **moderate** (change in Nagelkerke R-squared between **0.035 and 0.070**); or category C, **large** (change in Nagelkerke R-squared **> 0.070**).

[2] Erratum: Although described correctly in the text and implemented in the analysis, an earlier version of the monograph contained typographical errors in the equation describing the GLIM model for listening.

8.3 Results – Concordance of the *CAEL* (Online) and *CAEL* (Test Centre) and Its Moderation by the Test Takers Gender and First Language

8.3.1 Listening

- The exhibits below contain the statistical results of fitting the nine variable GLIM model described above. The comparison of models 1 and 2 provides the statistical test of uniform DIF and the corresponding effect size. The 3-df test of uniform DIF was statistically significant, $\chi^2(3) = 9.067, p<0.05$; however, the difference in R-squared is 0.002. Therefore the uniform DIF effect is labelled as category A, negligible or nonsignificant.
- The comparison of models 2 and 3 provides the statistical test of the moderated uniform mode of test delivery DIF and the corresponding effect size. The 2-df test of moderate uniform DIF was statistically non-significant, $\chi^2(2) = 0.541$, p=0.763, and the difference in R-squared is 0.001.
- **We can conclude that the listening component shows concordance for the mode of test delivery, including that this mode effect is not moderated by test taker gender or first language.**

Ordinal Logistic Regression

Model Fit Measures

						Overall Model Test		
Model	Deviance	AIC	R^2_{McF}	R^2_{CS}	R^2_N	χ^2	df	p
1	3784	3808	0.240	0.0871	0.275	1193	4	0.000
2	3775	3805	0.242	0.0877	0.277	1203	7	0.000
3	3775	3809	0.242	0.0878	0.278	1203	9	0.000

Note. The dependent variable 'L_Band' has the following order: 10 | 20 | 30 | 40 | 50 | 60 | 70 | 80 | 90

Model Comparisons

Comparison					
Model		Model	χ^2	df	p
1	-	2	9.067	3	0.02842
2	-	3	0.541	2	0.76288

8.3.2 Reading

- The exhibits below contain the statistical results of fitting the nine variable GLIM model described above.
    - The 3-df test of uniform DIF was statistically significant, $\chi^2(3) = 12.42$, p<0.05; however, the difference in R-squared is 0.002. Therefore the uniform DIF effect is labelled as category A, negligible or nonsignificant. This conclusion is supported by the individual parameter estimates, significance tests, and odds ratio effect sizes.
    - The 2-df test of moderate uniform DIF was statistically non-significant, $\chi^2(2) = 1.59$, p=0.453, and the difference in R-squared is 0.001.
- **We can conclude that the reading component shows concordance for the mode of test delivery, including that this mode effect is not moderated by test taker gender or first language.**

Ordinal Logistic Regression

Model Fit Measures

						Overall Model Test		
Model	Deviance	AIC	R^2_{McF}	R^2_{CS}	R^2_N	χ^2	df	p
1	4326	4350	0.218	0.0882	0.256	1209	4	0.000
2	4314	4344	0.221	0.0891	0.258	1222	7	0.000
3	4312	4346	0.221	0.0892	0.259	1223	9	0.000

Note. The dependent variable 'R_Band' has the following order: 10 | 20 | 30 | 40 | 50 | 60 | 70 | 80 | 90

Model Comparisons

Comparison					
Model		Model	χ^2	df	p
1	-	2	12.42	3	0.006075
2	-	3	1.59	2	0.452464

8.3.3 Speaking

- The exhibits below contain the statistical results of fitting the nine variable GLIM model described above.
- The 3-df test of uniform DIF was statistically significant, $\chi^2(3) = 41.39$, $p<0.05$; however, the difference in R-squared is 0.013. Therefore the uniform DIF effect is labelled as category A, negligible or nonsignificant.
- The 2-df test of moderate uniform DIF was statistically non-significant, $\chi^2(2) = 1.57$, $p=0.455$, and the difference in R-squared is 0.001.
- **We can conclude that the speaking component shows concordance for the mode of test delivery, including that this mode effect is not moderated by test taker gender or first language.**

Ordinal Logistic Regression

Model Fit Measures

						Overall Model Test		
Model	Deviance	AIC	R^2_{McF}	R^2_{CS}	R^2_N	χ^2	df	p
1	3027	3049	0.127	0.0371	0.144	440	4	0.000
2	2986	3014	0.139	0.0405	0.157	481	7	0.000
3	2984	3016	0.139	0.0406	0.158	483	9	0.000

Note. The dependent variable 'S_Band' has the following order: 20 | 30 | 40 | 50 | 60 | 70 | 80 | 90

Model Comparisons

Comparison					
Model		Model	χ^2	df	p
1	-	2	41.39	3	5.397e-9
2	-	3	1.57	2	0.4553

8.3.4 Writing

- The exhibits below contain the statistical results of fitting the nine variable GLIM model described above.
- The 3-df test of uniform DIF was statistically significant, $\chi^2(3) = 9.39$, p<0.05; however, the difference in R-squared is 0.002. Therefore the uniform DIF effect is labelled as category A, negligible or nonsignificant.
- The 2-df test of moderate uniform DIF was statistically non-significant, $\chi^2(2) = 3.02$, p=0.221, and the difference in R-squared is 0.002.

  We can conclude that the speaking component shows concordance for the mode of test delivery, including that this mode effect is not moderated by test taker gender or first language.

Ordinal Logistic Regression

Model Fit Measures

						Overall Model Test		
Model	Deviance	AIC	R^2_{McF}	R^2_{CS}	R^2_N	χ^2	df	p
1	2916	2938	0.237	0.0750	0.268	907	4	0.000
2	2907	2935	0.240	0.0757	0.270	916	7	0.000
3	2904	2936	0.240	0.0759	0.271	919	9	0.000

Note. The dependent variable 'W_Band' has the following order: 10 | 30 | 40 | 50 | 60 | 70 | 80 | 90

Model Comparisons

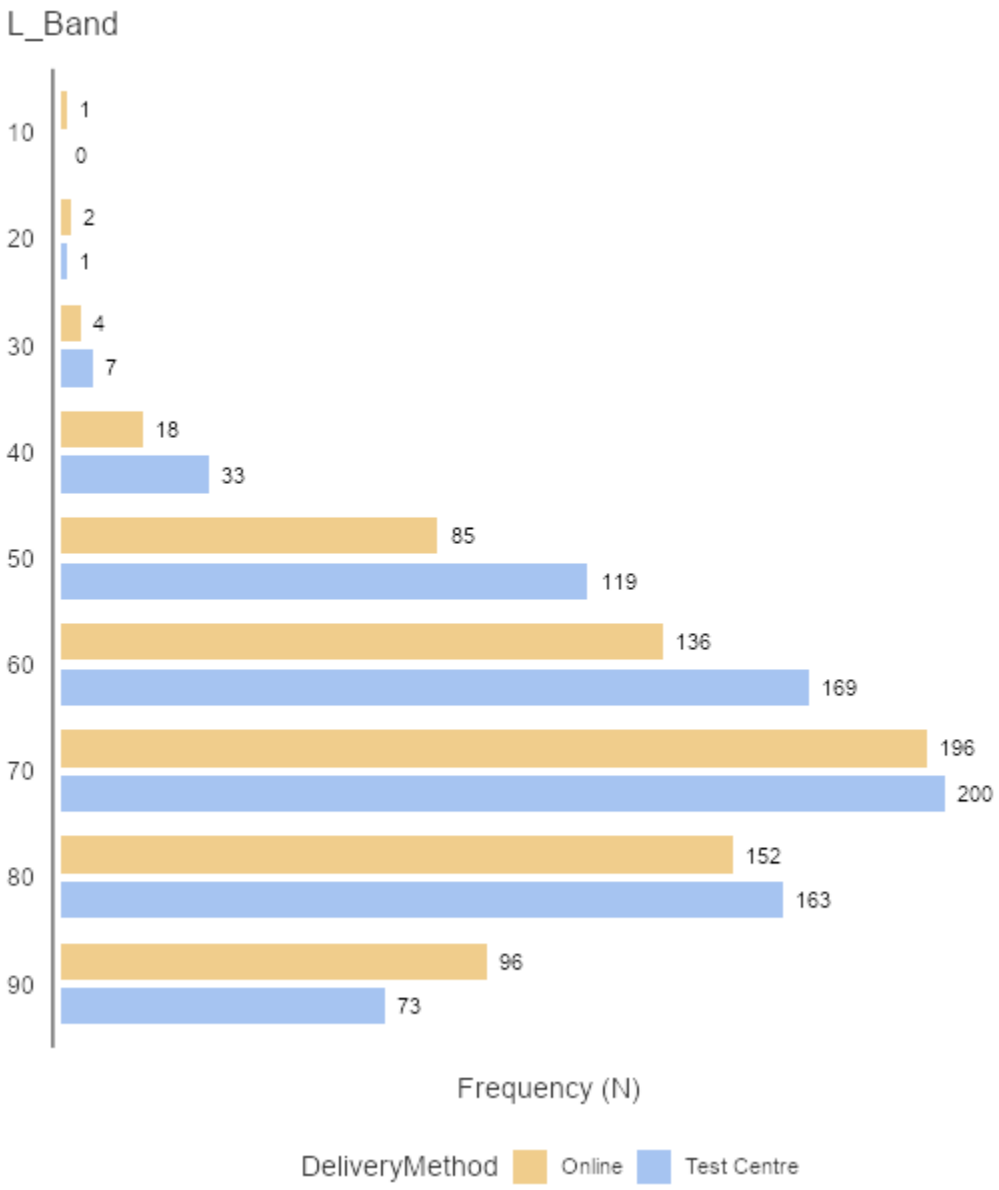
Comparison					
Model		Model	χ^2	df	p
1	-	2	9.39	3	0.02453
2	-	3	3.02	2	0.22128

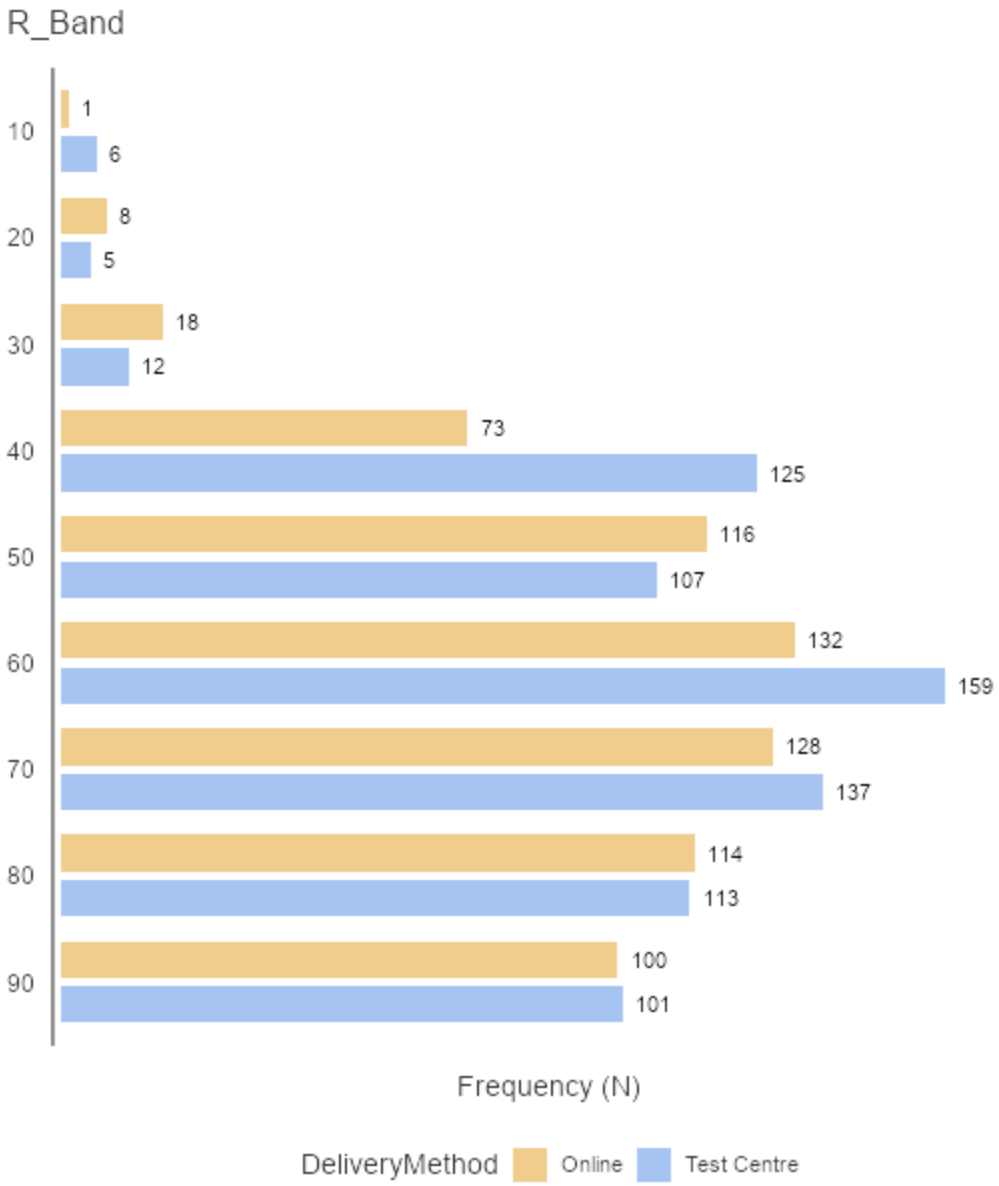
8.4 Conclusions

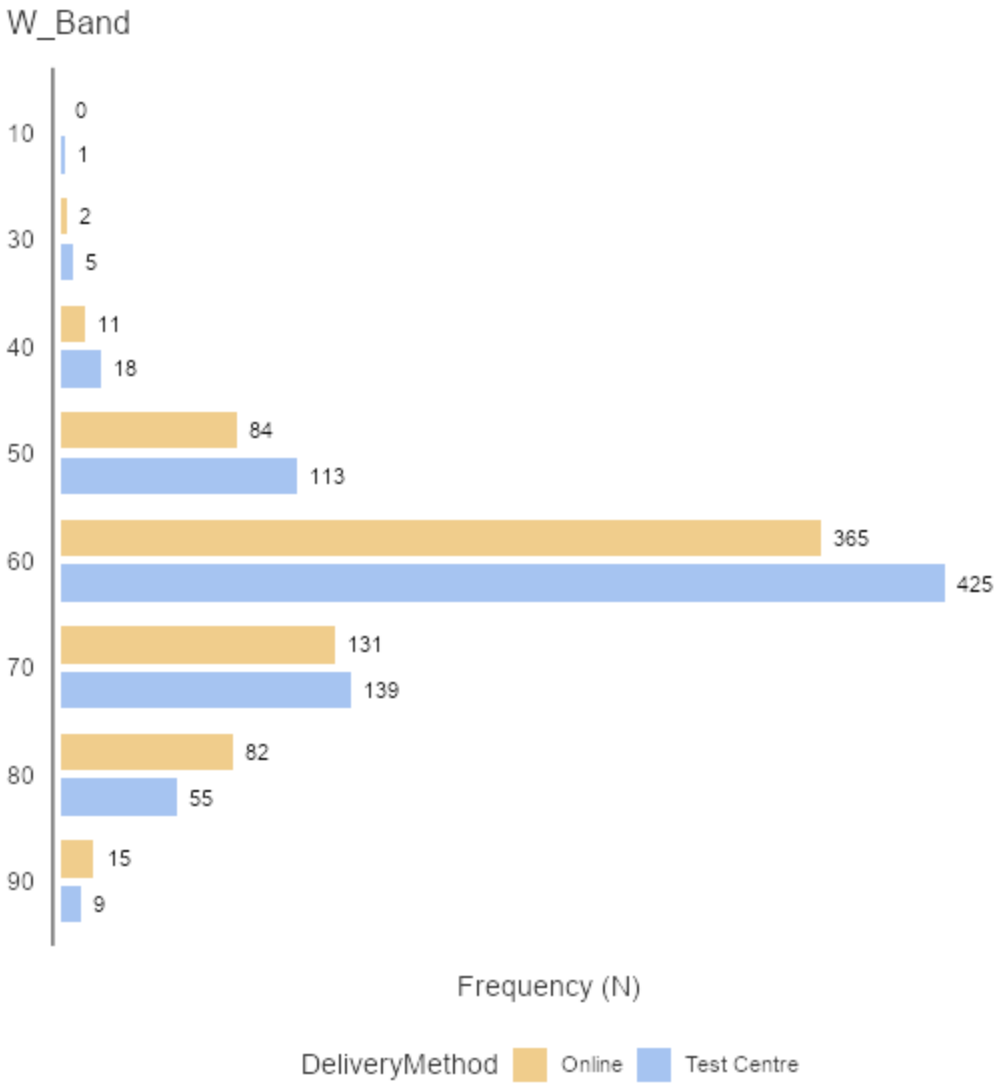
A GLIM DIF model is specified that allows the testing of concordance due to mode of test delivery, test centre compared to online, and that concordance is moderated (i.e., is dependent on) the test taker's gender or their reported first language.

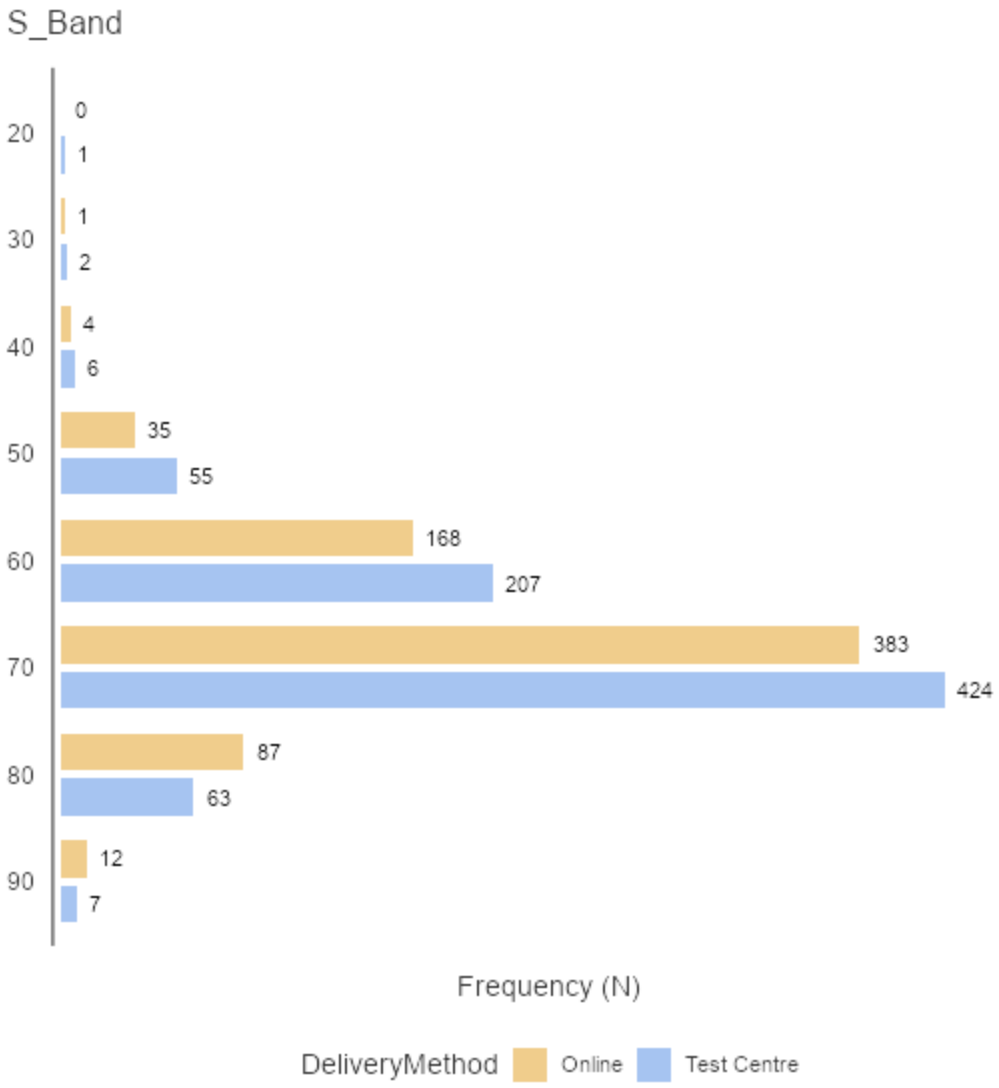
The GLIM DIF analyses showed that test takers of equal language competency measured by the *CAEL* would earn statistically similar band scores whether they took the *CAEL* at a test centre or online.

8.5 Appendix Chapter 8 – Frequency Plots of the Four Band Scores




R_Band
10
20
30
40
50
60
70
80
90
1
6
8
5
18
12
73
125
116
107
132
159
128
137
114
113
100
101
Frequency (N)
DeliveryMethod
Online
Test Centre


W_Band
10
30
40
50
60
70
80
90
0
1
2
5
11
18
84
113
365
425
131
139
82
55
15
9
Frequency (N)
DeliveryMethod
Online
Test Centre


S_Band
20
30
40
50
60
70
80
90
0
1
1
2
4
6
35
55
168
207
383
424
87
63
12
7
Frequency (N)
DeliveryMethod
Online
Test Centre

Chapter 9 Concordance Analysis Using Statistical Matching as an Alternative to Within Subjects Designs or Randomized Experiment

As described in *Section 2.3* and *Section 2.3.3*, the matched-subjects design is, in essence, a compromise between the between-subjects and within-subjects design, taking advantage of the strengths of each design by mitigating the concerns described in *Section 2.3.3* to approximate group equivalence.

Liu, Zumbo, Gustafson, Huang, Kroc, and Wu (2016) describe a statistical matching methodology that does address a concordance "Do the same items function the same in different test administration modes (e.g., test centre vs. online) for all test takers?"

As Liu et al. (2016) state, the most common attempts to approximate group equivalence are matching and covariance adjustment, the latter of which was reported in *Chapter 8*. In concordance studies, matching is a method of selecting units from one of the test delivery groups (for example, those who complete the test at a test centre) who are similar to those in the other group with respect to the observable covariates related to the group membership mechanism. However, exact matching becomes onerous or even impossible when matching on a large number of covariates, especially when several continuous covariates are involved. This concern with the exact matching of a large number of covariates results in the sparse data problem; some units from the treatment group do not have matched units from the control group. Rosenbaum and Rubin (1983) described this problem and described the need to find approximate matching methods instead of exact matching.

Liu et al. went on to state that stratification is an alternative to matching. When applying a stratification method, groups are classified into several strata, and in each stratum, units from one of the test delivery groups such as taker takers at a test centre are comparable to the units from the test takers in the online group (Rosenbaum, 2002). While easier to implement than the exact matching methods, stratification methods may still produce extremely unbalanced groups within certain strata. However, stratification may also run into the sparse data problem as exact matching methods.

An exact matching method allows the comparison to be made from a matched sample and is intended to mirror the processes of an experiment using randomization to groups. Consistent with other sections of this report, the focus of the analysis is the four language component band scores (Reading, Listening, Writing, and Speaking) because test users' decisions are made on band scores.

Conclusions from the Matching Study

Using exact matching method, we created samples that were matched on test takers' demographics (age and gender) and performance on other language components.

The comparison of the matched sample show that the test centre and online versions of the *CAEL* are concordant, i.e., no difference was observed on the band scores for test takers of similar backgrounds who took the two versions of the *CAEL*.

9.1 Sample

The analysis used the data of *CAEL* test takers described in *Chapter 5* for test administrations between June and October 2020. A total of 1,455 test takers completed the *CAEL* either at a test centre (n = 765) or online (n = 690).

9.2 Exact Matching Method

In social and behavioural sciences, matching is frequently used for reducing the confounding effects attributable to unwanted variables in observational studies. Matching is "the process of making a treatment group and a control group comparable with respect to extraneous factors" (Everitt & Skrondal, 2010, p. 271). Extraneous factors, also called nuisance variables or covariates, are unwanted factors that are not of research interest but can potentially influence the outcome variable and or predictor variable (Colman, 2015; Powers & Knapp, 2011; VandenBos, 2015).

Usually, extraneous factors are characteristics of a participant or background variables of a study. For example, to compare female and male test takers' performance on an oral English proficiency test, extraneous factors such as the personality of extroversion or introversion (a personal characteristic) and test locations (a background variable) may obscure the comparison between gender groups. In observational studies, confounding effect attributable to extraneous factors is prevalent but rarely under control due to the absence of experimental control procedures like randomization of participants, elimination of nuisance variables, or holding nuisance variables constant. Hence, statistical control measures are often applied to remedy this situation.

Matching is a popular statistical control method that intends to create groups from observational data with approximate distributions like those in a randomized experiment. Various types of matching have been identified, including exact matching, caliper matching, and frequency matching (e.g., Everitt & Skrondal, 2010; Porta et al., 2014). Nowadays, these methods are readily available and can be implemented using the open-source R platform (R Core Team, 2020). Popular R packages include *MatchIt* (Ho et al., 2011), *Matching* (Sekhon, 2011), *optmatch* (Hansen & Klopfer, 2006), *cem* (Iacus et al., 2020), *stratamatch* (Aikens et al. 2020).

Exact matching refers to the situation where a participant in the treatment group is matched with a participant who has the same values on all covariates (e.g., age, years in education, sex, etc.) in the control group. This matching method allows different types of covariates, including continuous, dichotomous, categorical, and nominal. Although finding individuals who are matched on a handful of covariates (e.g., two or three categorical variables) may be relatively simple, it could soon become impossible with an increasing number of covariates.

Given our analysis goals and the data at hand, we chose exact matching due to its simplicity. Although we used mixed-effects models to compare the group means between *CAEL* at the test center vs. online, statistical power is not a particular concern; rather, we are interested in the value of the mean difference on the reporting scale. Also, only five covariates are available for matching, and four of them are categorical.

It should be noted that this matched sub-sample data, after-matching, was used in *Chapter 10* for the analysis of classification error, DA/DC analysis.

9.3 Results

A matched dataset was created for each comparison.

- The listening score comparison: the test takers were matched using the covariates of gender (binary), age (continuous), reading band scores (max. 9 levels), writing band scores (max. 9 levels), speaking band scores (max. 9 levels).
- The reading score comparison: the test takers were matched using the covariates of gender, age, listening band score, writing band score, and speaking band score.
- The speaking score comparison: test takers were matched using the covariates of gender, age, listening band score, reading band score, and writing band score.
- The writing score comparison: test takers were matched using the covariates of gender, age, listening band score, reading band score, and speaking band score.

Table 9.1 shows the number of test takers retained in each group after matching. As expected, many cases could not find their "exact match" on all the covariates in the other group and, thus, were discarded.

Table 9.1. The sample size for each group before and after matching

The Target Component Score Under Investigation		Test Center	Online
	All	765	690
Listening	*Matched*	*219*	*197*
	Discarded	546	493
Reading	*Matched*	*230*	*210*
	Discarded	535	480
Speaking	*Matched*	*136*	*135*
	Discarded	629	555
Writing	*Matched*	*142*	*136*
	Discarded	623	554

Based on the matched sample, linear mixed-effects models were used to compare the mean component band scores between test center vs. online (coded 0 and 1 in the model, respectively). Linear mixed models treat the band scores as continuous.

Mixed-effects models were chosen to take into account the cluster effect and weights arising from the matching process. In *Table 9.2*, two parameters were reported for each model. The intercept shows the grand mean, and the parameter named "mode of delivery" represents the average score difference between test center vs. online. A positive value means a higher score for the online group. As shown in the table, the mean differences were small for all four components.

Table 9.2. Group difference in band scores after matching

		Estimate	Std. Error	t value	P
Listening	(Intercept)	70.25	0.98	71.50	<.001
	mode of delivery	0.42	0.96	0.44	0.657
Reading	(Intercept)	66.73	1.17	57.28	<.001
	mode of delivery	-1.53	1.13	-1.25	0.179
Speaking*	(Intercept)	67.39	0.34	198.00	<.001
	mode of delivery	0.70	0.49	1.43	0.154
Writing	(Intercept)	63.06	0.83	76.06	<.001
	mode of delivery	0.34	0.82	0.42	0.678

*The mixed-effects model for speaking band scores did not converge properly, so a regular regression (i.e., ignoring the cluster effects due to matching) is reported for this comparison.

9.4 Conclusion

As mentioned earlier, the literature has described many statistical methods for reducing the confounding effects in observational studies, such as exact matching, greedy matching, and optimal matching.

This chapter reported the matching results based on exact matching focusing on the comparability of band scores rather than on the methods for creating matched samples. Using other matching methods such as nearest-neighbour matching and optimal matching could avoid losing too many cases from both groups. To check the consistency of the results across matching methods, we conducted a similar analysis using exact matching, optimal matching (1-to-1), and optimal full matching. The results were consistent.

The results in this chapter highlight the statistical and methodological advances in the last 30 years in the development of alternatives to within-subjects designs; see *Chapters 2* and *3* for a discussion of within and between-subject designs. Matched samples bypass the limitations of within-subject designs for concordance studies and build on the strengths of the conceptual advantages of within and between design.

The matching concordance analyses showed that statistically matched test takers based on language competency measured by the *CAEL*, age, and gender earn statistically similar band scores whether they took the *CAEL* at a test centre or online.

Chapter 10 Measurement Error/Misclassification Analysis from a Statistically Matched Sample: Comparative Decision Consistency and Decision Accuracy

As described in *Section 3.2*, *CAEL* scores are reported on a 9-band scale from 10 - 90 with accompanying descriptors of what the performance represents in *Table 3.2*. When test takers are classified into different proficiency levels based on their test performance, measurement uncertainty (measurement error) is quantified based on the degree to which the classifications are accurate and consistent. Acknowledging that making classification decisions has a certain amount of uncertainty acknowledges that one accepts a certain amount of (minimal) potential classification error rate (Kane, 1994, 2006, 2013).

10.1 Misclassification and Concordance

As Zumbo (2016) states, the potential misclassification may come from measurement error on the test score or what is referred to as construct irrelevant variance or construct underrepresentation in terms of the domain one is testing. For example, construct irrelevant variance may arise from irrelevant task easiness or difficulty that can be traced to, for example, the design of the computer interface while measuring the construct of focus of the test, in this case, language abilities (Kane 2006; Messick 1989; Zumbo 2007a). As described in *Chapter 6*, this distinction of construct relevant and irrelevant variance is related to the description of the latent test data dimensionality into those intended-to-be-measured, such as language ability (in this case construct relevant), dimensions or factors and the unintended-to-be-measured (in this case, the construct irrelevant) dimensions for factors. This classification acknowledges that test and assessment data are inherently multidimensional, which may be related to misclassification.

Whether the construct irrelevant variance may arise from irrelevant task easiness or difficulty that can be traced to, for example, the design of the computer interface while measuring the construct of focus of the test; or inherent multidimensionality of the test data wherein the secondary or minor dimensions are related to features of the test centre or online test delivery the measurement uncertainty needs to be compared for the two modes of test delivery.

- Acknowledging that making classification decisions has a certain amount of uncertainty acknowledges that one accepts a certain amount of (minimal) potential classification error rate that should be similar for the two modes of test delivery.

10.1.1 Evidence of Concordance: Comparative Misclassification Take the Matching Principle Into Account

It is useful at this point to remind ourselves of some of the formalism described in *Section 2.2*; test data can be characterized as the realization of a stochastic event defined on a product space where the components are the probability spaces for items, examinees, and test settings (test centre or online), respectively. In language assessment, we often deal with a profile of test scores reflecting listening, reading, writing, and speaking, so we will use a vector to denote a multidimensional observation on a single test taker unit. Keeping the three-component product space in mind, the interpretation of this test data involving online testing *minimally* requires a judgement of exchangeability (similarity or homogeneity) of a vector of language testing component scores, examinees, and test settings, as well as the specification of a stochastic process that is supposed to have generated the data (Zimmerman & Zumbo, 2001).

As described in *Section 2.2*, next, one needs to decide which of the online testing scenarios described in *Section 1.1* involving either (i) different tests administered online that substitute for the original or (ii) an online version of an existing test. As described in *Chapter 3*, the *CAEL* is an online version of an existing test. How the *CAEL* is used, described in *Chapter 3*, in high-stakes decision-making will lead to the appropriate descriptions of the necessary exchangeability for each of the three orthogonal components characterizing the stochastic event. The description in *Table 2.1* of the implications of Angoff's (1993) matching principle for evidence of concordance allows for the definition of optimal statistical psychometric methods that do not confound concordance with true differences in item performance (i.e., impact). As Millsap, Angoff, Mellenbergh (1989, 1994), and others state, it is essential that we compare individuals from different groups who are matched rather than randomly chosen pairs of individuals. Comparing pass rates or reliability (in our case, misclassification) for each unmatched group is considered uninterpretable and flawed psychometric evidence because they confound concordance with unmatched sample characteristics.

As Zumbo et al. (2002) note, the test purpose and test population help determine which decision-theoretic (e.g., decision accuracy and decision consistency) statistics should be of focus. As described in *Chapter 9*, statistical matching aims to approximate group equivalence. In concordance studies, matching is a method of selecting units from one of the test delivery groups, such as test takers at a test centre who are similar to those in the group of test takers who completed the test online with respect to the observable covariates related to the group membership mechanism.

10.2 Sample Data

The matched data from *Chapter 9*, described in *Table 9.1*, are used to compute the decision consistency and decision accuracy statistics for each group of test takers, those who completed the *CAEL* at a test centre and those who completed it online.

10.3 Decision Accuracy and Decision Consistency Statistics

For the Listening and Reading components, the decision accuracy (DA) and decision consistency (DC) of the classification decisions are calculated using the Rudner method (Rudner, 2001). Under this method, each test taker's estimated theta and its standard error of estimate are used to construct the normal distribution of theta for the test taker. The decision levels that are originally on the true score scale are mapped onto the theta scale. Each decision level is then applied to the distributions of theta to calculate the DA and DC for each test taker. The overall DA and DC for each decision level are calculated by averaging over all test takers.

For the Speaking and Writing components, DA and DC of the classification decisions were calculated using the Livingston and Lewis method (Livingston & Lewis, 1995) via the R package titled *betafunctions*. With this method, the reliability of the scores is used to estimate the effective test length, and the true-score distribution is estimated by fitting a 4-parameter beta model. The conditional distribution of score on an alternate test form, given the true score, is estimated from a binomial distribution based on the previously derived effective test length. Finally, the agreement between classifications on alternate forms is estimated by assuming conditional independence, given the true score.

10.4 Results and Conclusions of Decision Accuracy and Decision Consistency – Reduced Matched Sample Data

Tables 10.1 to *10.4* list the decision accuracy (DA) and decision consistency (DC) results from the matched samples in *Chapter 9*.

A review of the contents of Tables 10.1 to 10.3 reveals that there are no marked differences in DA (and likewise in DC) between the online and test centre modes of delivery for the *CAEL's* listening, reading, and speaking components.

In *Table 10.4*, on the other hand, the sample size for the reduced matched sample of those test takers who completed the *CAEL* online was too small to accurately estimate the DA or DC statistics for all but the 50 and 60 band scores levels. *Table 10.5* lists the DA and DC results for the full (unmatched) test data. Cell-by-cell comparison of the elements of *Tables 10.4* and *10.5* provides a contrast of matching and decision statistics for typical testing samples with the degree of comparability described in *Chapter 5*.

In terms of concordance, the evidence supports that the decision accuracy and consistency are of the band classification decisions are high and consistent across the two modes of test delivery (test centre and online) for the listening, reading, and speaking components. Because of the limited sample sizes for the writing score bands online, the evidence is insufficient for the statistically matched writing component to speak to the concordance of DA and DC for the writing component.

Table 10.1. DA and DC of Listening Decision Levels (reduced matched sample data)

	DA		DC	
Level	Online	Test Centre	Online	Test Centre
10	1.000	1.000	1.000	1.000
20	1.000	0.999	1.000	0.998
30	0.998	0.994	0.997	0.991
40	0.992	0.990	0.987	0.986
50	0.966	0.977	0.950	0.963
60	0.907	0.896	0.867	0.853
70	0.862	0.853	0.808	0.796
80	0.873	0.859	0.819	0.805
90	0.904	0.916	0.868	0.880

Table 10.2. DA and DC of Reading Decision Levels (reduced matched sample data)

	DA		DC	
Level	Online	Test Centre	Online	Test Centre
10	1.000	1.000	1.000	1.000
20	0.999	0.999	0.999	0.998
30	0.995	0.992	0.992	0.988
40	0.985	0.978	0.978	0.967
50	0.921	0.915	0.888	0.877
60	0.874	0.881	0.822	0.832
70	0.837	0.858	0.779	0.802
80	0.860	0.871	0.800	0.818
90	0.895	0.908	0.862	0.875

Table 10.3. DA and DC of Speaking Decision Levels (reduced matched sample data)

	DA		DC	
Level	Online	Test Centre	Online	Test Centre
10	1.000	1.000	1.000	1.000
20	1.000	1.000	1.000	1.000
30	1.000	1.000	1.000	1.000
40	1.000	1.000	1.000	1.000
50	0.999	0.998	0.999	0.997
60	0.959	0.947	0.940	0.925
70	0.854	0.843	0.797	0.783
80	0.996	1.000	0.994	1.000
90	1.000	1.000	1.000	1.000

Table 10.4. DA and DC of Writing Decision Levels (reduced matched sample data)

	DA		DC	
Level	Online	Test Centre	Online	Test Centre
10	--	1.000	--	1.000
20	--	1.000	--	1.000
30	--	1.000	--	0.999
40	--	0.992	--	0.988
50	0.952	0.934	0.913	0.905
60	0.764	0.842	0.686	0.782
70	--	0.873	--	0.823
80	--	0.981	--	0.970
90	--	1.000	--	1.000

Table 10.5. DA and DC of Writing Decision Levels (full sample data)

	DA		DC	
Level	Online	Test Centre	Online	Test Centre
10	1.000	1.000	1.000	1.000
20	1.000	1.000	1.000	1.000
30	0.999	0.999	0.999	0.999
40	0.989	0.990	0.983	0.984
50	0.928	0.925	0.897	0.892
60	0.844	0.831	0.784	0.767
70	0.864	0.881	0.811	0.833
80	0.971	0.988	0.956	0.981
90	1.000	1.000	1.000	1.000

Part V - Bringing the Multimethod Strands Together

Chapter 11 Closing Remarks and Conclusion

11.1 A Rigorous Integrated Method for Concordance Studies

The methodology of designing and interpreting concordance studies in support of the valid use of tests simultaneously offered in multiple modes of test delivery has been spurred by the recent global pandemic forcing test providers to offer alternative modes of test delivery. The multimethod approach introduced in this monograph is a general model for other concordance studies that provides a principled rationale for designing such studies to investigate any delivery modes; for example, a concordance study may investigate a test administered simultaneously at a test centre, remotely online, at pop-up administration centres, and in paper-and-pencil format.

Since the early 1980s, comparing average test scores, average item scores, or average pass rates for test takers who complete a test in one of two delivery modalities are considered uninterpretable and flawed psychometric evidence because they confound concordance with true differences in item performance (i.e., impact). The same goes for simply comparing group psychometric properties such as test reliability for each group. These unconditional approaches were replaced over three decades ago by a formal definition of statistical methods to investigate concordance based on the matching principle (see, for example, Angoff, 1993).

However, an integrated and rigorous concordance methodology is critically needed to support the move to online testing to guide the design choices and interpretation of concordance studies. A multimethod strategy was developed and presented herein to allow for a robust comparison of the test centre and online test performance that far exceeds conventional methods to investigate the comparability of tests- i.e., their concordance.

Compared to typical comparability studies, which generally only involve generalized linear models, this report establishes a more robust sense of comparability by expanding our evidential base into multiple methods to support the comparability of the test takers' scores. *Chapters 1* and *2* of this report provide an integrated theoretical foundation and logic based on a principled approach to research design in support of test validity.

In addition to the matching principle, the rigour and logic of our methodology are grounded in test validity and a framework based on four key principles. First, Angoff's (1993) matching principle allows for the definition of optimal statistical psychometric methods that do not confound concordance with true differences in item performance (i.e., impact). Second, the equity principle states it should be a matter of indifference to a test taker or a test user about which of the two modes of test administration (test centre or online) test takers choose. Third, the test use principle states that the comparison across test administrations should focus on the scale on which scores are reported- for example, band scores on each of the (four) components of a language test rather than an item-by-item comparison. Fourth, there is an overall principle of multiple sources of

evidence (multimethod methodology) that calls for more than one source of evidence supporting the concordance investigation to rule out rival plausible alternative interpretations and ferreting out multiple sources of potentially hidden invalidity.

We highlight the following methodological points for the reader.

- What is being prioritized in the schematic depicting establishing the test validity as the bona fides for using the test scores – Blending Zumbo's Model and Kane's Argument-Based Approach (*Section 2.1.1*)
- The description of *Figure 2.4* characterizing the various strengths of the claims we can make from concordance studies based on Zumbo's DLD framework.
    - Two methodological principles emerge from Zumbo's DLD theoretical framework that forms the basis and logic to investigate the concordance of the test centre and online testing. *Table 2.1* describes the matching, equity and test use principles and their corresponding criteria and evidence.
    - A few points are noteworthy. First, the equity principle is defined differently for the cases of an online version of an existing test than an indicator test that substitutes for the test centre version. In line with Zumbo's DLD framework, the former implies a stricter principle of indifference. The latter is a weaker form of no hindrance or disadvantage for the test taker having to take the indicator test. Of course, the former implies the latter.
    - In addition to the equity and test use principles, there is an overall principle of multiple sources of evidence (multimethod methodology) that calls for more than one source of evidence in support of the concordance investigation to allow to rule out rival plausible alternative interpretations and to ferret out multiple sources of potentially hidden invalidity.
    - The equity principle in *Section 2.2.1* states that it should be a matter of indifference to a test taker or a test user as to which of two modes of test administration (test centre or online) the test taker chooses to take the test.
- Generally speaking, there are three study design options (i.e., research designs) for concordance studies that operationalize the matching principle. *Section 2.3* describes the between-subjects, within-subjects, and matched-subjects options and guides design choice.
    - The between-subjects and matched-subjects designs are the most commonly used concordance and DIF studies options, with the former being most widely used and discussed.

11.2 The Concordance of Test Centre and Online Delivery of the *CAEL*

This novel methodology is applied to investigate the concordance of the *CAEL* delivered at a test centre and online. Between June and October 2020, a sample of 1,455 *CAEL* test takers, 765 test takers who completed the *CAEL* at a test centre, and 690 completed it online. The findings from the six statistical and psychometric methods are consistent. The sample of the test centre and online test takers were equivalent, and the test performance was found to be consistently concordant.

Together, this is strong evidence that the conclusions from the *CAEL* band scores from the test centre and online versions are concordant, fully comparable band score performance of test takers at various levels of *CAEL*'s language domains.

Each of the elements of this method approach brings unique contributions to the body of evidence. What follows are the highlights.

- The DLD framework, see *Figure 2.4*, highlights that one needs to investigate both the exchangeability of the test taker test centre and online samples and the exchangeability of the test scores for a strong claim of concordance.
- The test takers reported first language (L1) considered from a language family point of view – a brief report on the construction of language group recommends classifying test takers' self-reported first languages into Indo-European vs. Non-Indo-European family groups, which is required for language family to be used as a grouping variable in the statistical/psychometric methods for comparability.
- Covariance Matrices – this method investigates a comparison of the joint distribution of the test performance.
- Nonparametric Domain Level Modeling – a novel method was developed to allow the equating of test takers and comparison of band scores by data visualization.
- Generalized Linear Model Approaches – an extended GLIM DIF model is introduced to investigate test delivery concordance and potential moderators of concordance based on the test taker's gender or reported first language.
- Matching Methods – the exact matching method allows the comparison to be made from a matched sample and is intended to mirror the processes of an experiment using randomization to groups.
- Classification Accuracy and Classification Consistency – these findings reflect the actual use of the *CAEL* in practice and compare the classification outcomes for each mode of delivery.

The results of this analysis will serve as evidence to support the comparability of the test centre and online modes of delivery for stakeholders. The multimethod approach is a model for other concordance studies that provides a principled rationale for the design of such studies. A rigorous test of the concordance of importance to test users, test providers, and external stakeholders who rely on valid and comparable test performance and test use across different test administration modalities.

Returning to *Section 2*, one can evaluate the evidence reported herein using the categories described in *Figure 2.4* from Zumbo's DLD framework. A review of the evidence reported in *Chapters 5* through *10* supports the claim that there is a high degree

of exchangeability of test scores and a high degree of exchangeability of test takers. This degree of exchangeability reflects the conditions described in the top left corner of *Figure 2.4*. The test provider, test user, and test taker have good reason to believe that the concordance study allows for general measurement inference because there is evidence to support the test scores' exchangeability and the online test takers' exchangeability with those taking the test at a test centre. ***In this case, the test scores from the online and test centre modes of administration are fully interchangeable***.

11.3 Next Steps

There are three directions for the next steps.

- The covariates included in this study are limited to those available to us. It is likely that other confounding variables still exist, and they could distort the comparison results. Future studies could look into ways to collect more variables and, when possible, include variables that could help explain the choice between test center vs. online tests and variables that are related to test taker performance.
- There is a need to investigate if changing proctoring methods impacts concordance. This will, of course, depend on how different and in what way the alternative proctoring processes differ.
- What is missing from the concordance studies to date is a rigorous comparison of the "test taking experience" and any possible relations to individual difference variables such as test anxiety. Evidence of test taker-reported experience will add important comparative evidence once established that the test scores are concordant and comparable.

References

-- (2010) Speaking Personally—With David Foster, *American Journal of Distance Education, 24:4*, 227-233, DOI: 10.1080/08923647.2010.519945

Addicott, S., & Foster, D. (2017). Security of Technology-Enhanced Assessments. In J. Scott, D. Bartram, & D. Reynolds (Eds.), *Next Generation Technology-Enhanced Assessment: Global Perspectives on Occupational and Workplace Testing (Educational and Psychological Testing in a Global Context, pp. 171-192*). Cambridge: Cambridge University Press.

Aikens, R. C., Rigdon, J., Lee, J., Baiocchi, M., Goldstone, A. B., Chiu, P., Woo, Y. J., & Chen, J. H. (2020). Stratified pilot matching in r: The stratamatch package. *Statistics ArXiv*. https://arxiv.org/abs/2001.02775

Altman, N.S. (1992). An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression. *The American Statistician, 46*, 175-185.

Angoff, W. H. (1993). Perspectives on differential item functioning methodology. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 3–23). Hillsdale, NJ: Erlbaum.

Bartholomew, D.J. (1983). Latent Variable Models for Ordered Categorical Data. *Journal of Econometrics, 22*, 229-243.

Bartram, D. (2009). Commentaries: The international test commission guidelines on computer-based and internet-delivered testing. *Industrial and Organizational Psychology, 2*, 11–13.

Bujang MA, Sa'at N, Tg Abu Bakar Sidik TMI, Lim CJ. (2018). Sample size guidelines for logistic regression from observational studies with large population: Emphasis on the accuracy between statistics and parameters based on real life clinical data. *Malaysian Journal of Medical Science, 25(4)*, 122–130.

Campbell, L., & Poser, W. J. (2008). *Language Classification: History and Method*. Cambridge University Press. https://doi.org/10.1017/CBO9780511486906

Chen, M.Y., Liu, Y., & Zumbo, B.D. (2020). A Propensity Score Method for Investigating Differential Item Functioning in Performance Assessment. *Educational and Psychological Measurement, 80(3)*, 476–498.

Colman, A. M. (2015). *A dictionary of psychology*. Oxford University Press.

Crawford, A. V., Green, S. B., Levy, R., Lo, W.-J., Scott, L., Svetina, D., & Thompson, M. S. (2010). Evaluation of Parallel Analysis Methods for Determining the Number of Factors. *Educational and Psychological Measurement, 70(6)*, 885–901.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*, 281-302.

De Finetti, B. (1975). *Theory of probability (Vol. 2)*. New York: Wiley.

Dorans, N. J., & Holland, P.W. (1993). DIF Detection and Description: Mantel-Haenszel and Standardization. In P. W. Holland, and H. Wainer (Eds.), *Differential Item Functioning* (pp. 35-66). Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

Eberhard, D. M., Simons, G. F., & Fennig, C. D. (Eds.). (2020). *Ethnologue: Languages of the world (23rd ed.), online version*. SIL International. http://www.ethnologue.com

ETS (2020). *ETS introduces at-home solution for TOEFL iBT and GRE General Test amid coronavirus pandemic. ETS Press Release*. Retrieved from: https://news.ets.org/press-releases/ets-introducesat-home-solution-for-toefl-ibt-test-and-gre-general-test-amidcoronavirus-pandemic/

Everitt, B., & Skrondal, A. (2010). *The Cambridge dictionary of statistics* (4th ed). Cambridge University Press.

Fox, J., Friendly, M., & Monette, G. (2021). *heplots: Visualizing Tests in Multivariate Linear Models.* https://CRAN.R-project.org/package=heplots

Friendly, M., Monette, G., & Fox, J. (2011). Elliptical Insights: Understanding Statistical Methods Through Elliptical Geometry. *Statistical Science, 28(1)*, 1–39.

Friendly, M., & Sigal M. (2014). Recent Advances in Visualizing Multivariate Linear Models. *Revista Colombiana de Estadística Current Topics in Statistical Graphics, 37(2)*, 261-283.

Gadermann, A.M., Chen, Y., Emerson, S.D., & Zumbo, B.D. (2018). Examining validity evidence of self-report measures using Differential Item Functioning: An illustration of three methods. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences, 14*, 164–175.

Guttman, L. (1945). A basis for analyzing test-retest reliability. *Psychometrika, 10*, 255–282.

Guttman, L. (1953a). A special review of Harold Gulliksen, Theory of mental tests. *Psychometrika, 18*, 123–130.

Guttman, L. (1953b). Reliability formulas that do not assume experimental independence. *Psychometrika, 18*, 225–239.

Hansen, B. B., & Klopfer, S. O. (2006). Optimal full matching and related designs via network flows. *Journal of Computational and Graphical Statistics, 15(3)*, 609–627.

Herrera, A.N., Gómez, J. (2008). Influence of equal or unequal comparison group sample sizes on the detection of differential item functioning using the Mantel–Haenszel and logistic regression techniques. *Quality and Quantity, 42*, 739-755.

Hintze, J.L., & Nelson, R.D. (1998). Violin plots: A Box Plot-Density Trace Synergism, *The American Statistician, 52(2)*, 181-184,

Ho, D. E., Imai, K., King, G., & Stuart, E. A. (2011). Matchit: Nonparametric preprocessing for parametric causal inference. *Journal of Statistical Software, 42(8)*.

Holland, P. W. (1986). Statistics and Causal Inference. *Journal of the American Statistical Association, 81*, 945-960.

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika, 30 (2)*, 179–185.

Hubley, A. M., & Zumbo, B. D. (2011). Validity and the consequences of test interpretation and use. *Social Indicators Research, 103*, 219–230.

Hubley, A. M., & Zumbo, B. D. (2013). Psychometric characteristics of assessment procedures: An overview. In K. F. Geisinger (Ed.), *APA handbook of testing and assessment in psychology (Vol. 1, pp. 3–19)*. Washington, D.C.: American Psychological Association Press.

Iacus, S. M., King, G., & Porro, G. (2020). *cem: Coarsened Exact Matching* (1.1.20) [Computer software]. https://CRAN.R-project.org/package=cem

International Organization for Standardization. (2007, February 1). *ISO 639-3:2007 Codes for the representation of names of languages — Part 3: Alpha-3 code for comprehensive coverage of languages*. https://www.iso.org/standard/39534.html

Isbell, D. R., & Kremmel, B. (2020). Test Review: Current options in at-home language proficiency tests for making high-stakes decisions. *Language Testing, 37(4)*, 600–619.

Jodoin, M. G., & Gierl, M. J. (2001). Evaluating type I error and power rates using an effect size measure with the logistic regression for DIF detection. *Applied Measurement in Education, 14(4)*, 329–49.

Kane, M. (1994). Validating the performance standards associated with passing scores. *Review of Educational Research, 64*, 425–461.

Kane, M. (2006). Validation. In *Educational Measurement*, (pp. 17-64). 4th ed. Edited by R.L. Brennan. American Council on Education/Praeger, Westport.

Kane, M. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*, 1–73.

Kane, M. (2012). Validating score interpretations and uses. *Language Testing, 29(1)*, 3–17.

Kane, M. (2016) Explicating validity. *Assessment in Education: Principles, Policy & Practice, 23(2)*, 198-211.

Kirk, R.E. (2001). Promoting good statistical practices: Some suggestions. *Educational and Psychological Measurement, 61(2)*, 213-218.

Kroc, E., & Zumbo, B.D. (2020). A Transdisciplinary View of Measurement Error Models and the Variations of X=T+E. *Journal of Mathematical Psychology, 98*, 1-9.

Langenfeld, T. (2020), Internet-Based Proctored Assessment: Security and Fairness Issues. *Educational Measurement: Issues and Practice, 39*: 24-27.

Lee, H., & Geisinger, K. F. (2014). The effect of propensity scores on DIF analysis: Inference on the potential cause of DIF. *International Journal of Testing, 14*, 313-338.

Liu, Y., Kim, C., Wu, A. D., Gustafson, P., Kroc, E., & Zumbo, B. D. (2019). Investigating the performance of propensity score approaches for differential item functioning analysis. *Journal of Modern Applied Statistical Methods, 18(1)*, 1-27.

Liu, Y., Zumbo, B.D., Gustafson, P., Huang, Y., Kroc, E., Wu, A. (2016). Investigating Causal DIF via Propensity Score Methods. *Practical Assessment, Research & Evaluation, 21(13)*, 1-26.

Livingston, S.A., Lewis, C. (1995). Estimating the Consistency and Accuracy of Classifications Based on Test Scores. *Journal of Educational Measurement, 32(2)*, 179–197.

Loève, M. (1963). *Probability theory* (3rd ed.). New York: Van Nostrand.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading MA, Addison-Wesley.

Lubke, G., Dolan, C., & Neale, M. (2004). Implications of Absence of Measurement Invariance for Detecting Sex Limitation and Genotype by Environment Interaction. *Twin Research, 7(3)*, 292-298.

Mazza, A., Punzo, A., & McGuire, B. (2014). KernSmoothIRT: An R package for kernel smoothing in item response theory. *Journal of Statistical Software, 58(6)*, 1-34.

McCullagh, P. (1980). Regression models for ordinal data (with Discussion). *Journal of the Royal Statistical Society, Series B, 42*, 109–142.

Mellenbergh, G. J. (1989). Item bias and item response theory. *International Journal of Educational Research, 13*, 127–143.

Mellenbergh, G. J. (1994). Generalized linear item response theory. *Psychological Bulletin, 115*, 300–307.

Messick, S. (1989). Validity. In *Educational Measurement* (pp. 13–103). 3rd ed. Edited by R.L. Linn. American Council on Education and Macmillan, New York, N.Y., USA.

Millsap, R.E. (2011). *Statistical Approaches to Measurement Invariance*. New York: Routledge.

Molenaar, I. W. (2001). Thirty Years of Nonparametric Item Response Theory. *Applied Psychological Measurement, 25(3)*, 295–299.

Novick, M. R. (1966). The axioms and principal results of classical test theory. *Journal of Mathematical Psychology, 3*, 1–18.

Pereltsvaig, A. (2012). *Languages of the world: An introduction*. Cambridge University Press. https://doi.org/10.1017/CBO9781139026178.015

Porta, M. S., Greenland, S., Hernán, M., Silva, I. dos S., Last, J. M., & International Epidemiological Association (Eds.). (2014). *A dictionary of epidemiology* (6th ed). Oxford University Press.

Powers, B. A., & Knapp, T. R. (2011). *Dictionary of nursing theory and research* (4th ed). Springer Pub. Co.

R Core Team (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Ramsay, J.O. (1991). Kernel Smoothing Approaches to Nonparametric Item Characteristic Curve Estimation. *Psychometrika, 56*, 611-630.

Ramsay, J. O. (2000). *TESTGRAF: A Program for the Graphical Analysis of Multiple Choice Test and Questionnaire Data*. Unpublished manuscript, McGill University.

Ramsay, J.O., Silverman, B.W. (2005). *Functional data analysis. 2nd edition*. New York: Springer.

Ramsay, J.O., Silverman, B.W. (2002). *Applied functional data analysis*. New York: Springer.

Reise, S. P.; Widaman, K. F.; Pugh, R. H. (1993). "Confirmatory Factor Analysis and Item Response Theory: Two Approaches for Exploring Measurement Invariance. *Psychological Bulletin, Vol. 114 (3)*, 552-566.

Rényi, A. (1970). *Foundations of probability*. San Francisco: Holden-Day.

Rhemtulla, M., Brosseau-Liard, P. É., & Savalei, V. (2012). When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. *Psychological Methods, 17(3)*, 354–373.

Rosenbaum, P., & Rubin, D. (1983). The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika, 70*, 41-55.

Rossi, N., Wang, X., Ramsay, J.O. (2002). Nonparametric item response function estimates with the EM algorithm. *Journal of Educational and Behavioral Statistics, 27*, 291–317.

Rudner, L.M. (2001). Computing the expected proportions of misclassified examinees. *Practical Assessment, Research & Evaluation, 7(14),* 1-5.

Rubin, D. (1974). Estimating Causal Effects of Treatments in Randomized and Non-Randomized Studies. *Journal of Educational Psychology, 66*, 688-701.

Rubin, D. (1977). Assignment to a Treatment Group on the Basis of a Covariate. *Journal of Educational Statistics, 2*, 1-26.

Rubin, D. (1978). Bayesian Inference for Causal Effects: The Role of Randomization. *TheAnnals of Statistics, 6*, 34-58.

Schmitt, T. A. (2011). Current Methodological Considerations in Exploratory and Confirmatory Factor Analysis. *Journal of Psychoeducational Assessment, 29(4)*, 304–321.

Scott, N. W., Fayers, P. M., Aaronson, N. K., Bottomley, A., de Graeff, A., Groenvold, M., Gundy, C., Koller, M., Petersen, M. A., Sprangers, M. A., EORTC Quality of Life Group, & Quality of Life Cross-Cultural Meta-Analysis Group (2009). A simulation study provided sample size guidance for differential item functioning (DIF) studies using short scales. *Journal of clinical epidemiology, 62(3)*, 288–295.

Sekhon, J. S. (2011). Multivariate and propensity score matching software with automated balance optimization: The matching package for R. *Journal of Statistical Software, 42(7)*. https://doi.org/10.18637/jss.v042.i07

Shear, B. R., & Zumbo, B.D. (2013). False Positives in Multiple Regression: Unanticipated Consequences of Measurement Error in the Predictor Variables. *Educational and Psychological Measurement, 73*, 733-756.

Simonoff, J.S. (1996). *Smoothing Methods in Statistics*. Springer Series in Statistics. Springer-Verlag, New York.

Sireci, S.G. (2013), Agreeing on Validity Arguments. *Journal of Educational Measurement, 50,* 99-104.

Stacey, V. (2020, December 10). ETS adds TOEFL Home Edition to product line. *The PIE News (News and business analysis for Professionals in International Education)*. Retrieved from https://thepienews.com/news/testing/ets-adds-toefl-home-edition-to-product-line/[2021-01-27]

Stout, W. (2002). Psychometrics: From practice to theory and back. *Psychometrika, 67*, 485-518.

Tucker, L.R., Koopman, R.F., & Linn, R.L. (1969). Evaluation of factor analytic research procedures by means of simulated correlation matrices. *Psychometrika, 34*, 421-459.

VandenBos, G. R. (Ed.). (2015). *APA dictionary of psychology* (2nd ed). American Psychological Association.

Wu. A. D., Li, Z., & Zumbo, B. D. (2007). Decoding the Meaning of Factorial Invariance and Updating the Practice of Multi-group Confirmatory Factor Analysis: A Demonstration With TIMSS Data. *Practical Assessment, Research and Evaluation, 12(3)*, 1-26.

Zimmerman, D. W., & Zumbo, B. D. (2001). The Geometry of Probability, Statistics, and Test Theory. *International Journal of Testing, 1*, 283-303.

Zumbo, B. D. (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.D. (2007). Three generations of differential item functioning (DIF) analyses: Considering where it has been, where it is now, and where it is going. *Language Assessment Quarterly, 4*, 223-233.

Zumbo, B.D. (2001). Methodology and Measurement Matters In Establishing A Bona Fide Occupational Requirement For Physically Demanding Occupations. In Norman Gledhill, Jean Bonneau, & Art Salmon (Eds.). *Proceedings of the Consensus Forum on Establishing BONA FIDE Requirements for Physically Demanding Occupations*, (pp. 37-52). Toronto, ON.

Zumbo, B. D. (2002). Contemporary Uses of Computer- and Internet-Based Testing, a Particularly Appealing Context for the Application of Bayesian Statistical Methods. *Testing International, 12(1)*, 8-10.

Zumbo, B. D. (2007a). Validity: Foundational issues and statistical methodology. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics (Vol. 26: Psychometrics, pp. 45-79).* Amsterdam: Elsevier Science.

Zumbo, B.D. (2007b). Three generations of differential item functioning (DIF) analyses: Considering where it has been, where it is now, and where it is going. Language Assessment Quarterly, 4, 223-233.

Zumbo, B.D. (2008). *Statistical Methods for Investigating Item Bias in Self-Report Measures, [The University of Florence Lectures on Differential Item Functioning]*. Universita degli Studi di Firenze, Florence, Italy.

Zumbo, B. D. (2009). Validity as Contextualized and Pragmatic Explanation, and Its Implications for Validation Practice. In Robert W. Lissitz (Ed.) *The Concept of Validity: Revisions, New Directions and Applications*, (pp. 65-82). IAP - Information Age Publishing, Inc.: Charlotte, NC.

Zumbo, B.D. (2013). On Matters of Invariance in Latent Variable Models: Reflections on the Concept, and Its Relations in Classical and Item Response Theory. In Paolo Giudici, Salvatore Ingrassia, and Maurizio Vichi (Eds.), *Statistical Models for Data Analysis*, (pp. 399-408). New York: Springer.

Zumbo, B.D. (2016). Standard-setting methodology: Establishing performance standards and setting cut scores to assist score interpretation. *Applied Physiology, Nutrition, and Metabolism, vol. 41, (Number, 6; Suppl. 2)*, S74-S82.

Zumbo, B.D. (2017). Trending Away From Routine Procedures, Towards an Ecologically Informed 'In Vivo' View of Validation Practices. *Measurement: Interdisciplinary Research and Perspectives, 15:3-4*, 137-139.

Zumbo, B.D., & Chan, E.K.H, (Eds.) (2014). *Validity and Validation in Social, Behavioral, and Health Sciences*. New York: Springer.

Zumbo, B.D., Gelin, M.N., and Hubley, A.M. (2002). The construction and use of psychological tests and measures. In *The Psychology Theme of the Encyclopedia of Life Support Systems (EOLSS)*. EOLSS Publishers, Oxford, UK.

Zumbo, B. D., & Hubley, A. M. (2003). Item Bias. In In Rocio Fernandez-Ballesteros (Ed.). *Encyclopedia of Psychological Assessment*, pp. 505-509. Sage Press, Thousand Oaks, CA.

Zumbo, B.D., & Hubley, A.M. (2016) Bringing consequences and side effects of testing and assessment to the foreground. *Assessment in Education: Principles, Policy & Practice, 23*(*2)*, 299-303.

Zumbo, B.D., Liu, Y., Wu, A.D., Shear, B.R., Astivia, O.L.O. & Ark, T.K. (2015). A Methodology for Zumbo's Third Generation DIF Analyses and the Ecology of Item Responding. *Language Assessment Quarterly, 12*, 136-151.