IELTS Research Reports
Online Series

Improving IELTS reading test score interpretations and utilisation through cognitive diagnosis model-based skill profiling

Eunice Eunhee Jang, Hyunah Kim, Megan Vincett, Christine Barron and Bruce Russell

IELTS™

BRITISH COUNCIL   idp   Cambridge Assessment English

Improving IELTS reading test score interpretations and utilisation through cognitive diagnosis model-based skill profiling

This study sought to investigate validity arguments related to IELTS reading score interpretations and use, exploring issues of consequential validity, the intersection of contextual validity and cognitive validity, as well as scoring validity.

Publishing details

Published by the IELTS Partners: British Council, Cambridge Assessment English and IDP: IELTS Australia © 2019.

How to cite this report

Jang, E. E., Kim, H., Vincett, M., Barron, C., and Russell, B. 2019. Improving IELTS reading test score interpretations and utilisation through cognitive diagnosis model-based skill profiling. *IELTS Research Reports Online Series, No. 2*. British Council, Cambridge Assessment English and IDP: IELTS Australia.
Available at https://www.ielts.org/teaching-and-research/research-reports

Introduction

This study by Eunice Eunhee Jang and her colleagues of the Ontario Institute for Studies in Education (University of Toronto) was conducted with support from the IELTS partners (British Council, IDP: IELTS Australia and Cambridge English Language Assessment), as part of the IELTS joint-funded research program. Research funded by the British Council and IDP: IELTS Australia under this program complement those conducted or commissioned by Cambridge English Language Assessment, and together inform the ongoing validation and improvement of IELTS.

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

In this study Jang et al. used mixed methods to analyse stakeholder perceptions of the IELTS reading component. They focus in particular on student and faculty interpretations of scores and score uses for admissions to higher education. The authors found both students and faculty had limited understanding of how the tasks and scores were relevant to academic studies, and that this, at times, created in the students a negative attitude towards the tests. The authors also found low levels of inferential ability among test-takers scoring 6.5 in IELTS, and suggest this may be linked to their findings that critical thinking skills are under-represented in the reading test.

Negative perceptions held by an individual test-taker or faculty member are understandable and unsurprising. However, there a number of factors to consider in test design. From a purely practical perspective, the numbers of aspiring students, institutions and disciplinary traditions make it unfeasible to customise tests for a huge population of test-takers with different academic destinations. Providing discipline-specific reading tasks would introduce variation and compromise reliability; longer reading texts would make the tests impractical and more expensive; and critical thinking is often embedded within the logic of the discipline itself.

Ultimately, the first priority of the testing organisation is to provide secure, valid and reliable tests and scores which can be used for decision-making in academic and professional contexts. In other words, fitness for purpose takes priority over contextual authenticity. To mitigate this, IELTS reading tasks are designed to engage similar cognitive and critical thinking processes that are involved in academic reading. These include higher order skills such as expeditious and careful reading to locate information, as well as to understand main ideas, analytical reading, evaluation and inferencing.

What this report highlights is the importance of managing the expectations of all test users, from the recognising institutions who set admissions requirements and the test preparation centres to the individual test-taker who needs to set a realistic timeframe to reach the desired level. Assessment literacy is necessary for all stakeholders who need to understand the principles underpinning fair assessment and prepare for this in an informed and timely manner.

Siân Morgan
Senior Research Manager
Cambridge Assessment English

Improving IELTS reading test score interpretations and utilisation through cognitive diagnosis model-based skill profiling

Abstract

This study sought to investigate validity arguments related to IELTS reading score interpretations and use, exploring issues of consequential validity, the intersection of contextual validity and cognitive validity, as well as scoring validity.

Through four phases, the present study sought to explore and better understand international students' perceptions regarding their language proficiency and preparedness for academic demands. We developed reading skill mastery profiles to investigate the possibility of enhancing test score users' interpretations of scores through the use of descriptors developed through cognitive diagnostic modeling and through scale anchoring.

The study results show that both students and faculty/staff have limited knowledge regarding what the IELTS test scores mean. Differences between the IELTS reading texts and the texts that students encounter in their first year of undergraduate study contribute negatively towards students' interpretations of test scores and sense of preparedness and further, to their attitude towards the test. Three reading attributes used to develop reader profiles and proficiency-level skill descriptors include basic comprehension, summarising main ideas, and inferential reasoning. Students who met the local cut-off score (6.5) lack mastery of inferential reasoning at the text level. Our analyses show that there are relatively fewer items measuring text-level critical thinking skills, which may explain why students who met the cut-off score lack inferential reasoning. This finding needs to be further investigated to determine if it is due to a lack of items or if the given local cut-off score is not appropriate for expecting successful academic performance requiring higher-order thinking skills such as inferential reasoning at the text level.

Test score users found the IELTS reading skill descriptors informative and useful for planning future actions to improve reading proficiency and support.

Authors' biodata

Eunice Eunhee Jang

Eunice Eunhee Jang is Professor in the Department of Applied Psychology and Human Development at the Ontario Institute for Studies in Education, University of Toronto. With specialisations in educational assessment and measurement and program evaluation, Dr Jang has led high-impact provincial, national, and international assessment and evaluation studies with various stakeholders. Her research centers on the pedagogical potential of cognitive diagnostic assessment, technology-rich assessment design and validation, and validity and fairness issues for diverse language learners.

Hyunah Kim

Hyunah Kim is a PhD student in the Developmental Psychology and Education program at the Ontario Institute for Studies in Education at the University of Toronto. Her research interests lie in educational measurement and assessment for language learners, as well as bilingual and heritage language education. Hyunah's language teaching experience covers from K–12 to older adults, and both Korean and English as a second, foreign or heritage language.

Megan Vincett

Megan Vincett is a doctoral student in the Developmental Psychology and Education program at the University of Toronto. Her research interests include the topics of fairness and validity in language assessment, particularly as it relates to students with learning difficulties. Prior to joining the lab, she worked as a policy analyst at the Ministry of Health, taught English to young language learners, and worked as an instructor/therapist for the Toronto Partnership for Autism Services.

Christine Barron

Christine Barron is a PhD student in the Developmental Psychology and Education program at the Ontario Institute for Studies in Education within the University of Toronto. Her research interests include language assessment and educational measurement, with a focus on measurement invariance and literacy development among students from diverse language backgrounds. She is also currently collaborating with the Toronto District School board to investigate the achievement, attrition, and demographic characteristics of students in French as a Second Language programs.

Bruce Russell

Bruce Russell is a PhD student in the Department of Applied Psychology and Human Development at the Ontario Institute for Studies in Education at the University of Toronto. His research interests are focused on EAP (English for Academic Purpose) curriculum and assessment in higher education. He is the Director (Academic) of International Programs at the University of Toronto, New College, which provides pre-sessional and in-sessional English language programs for multilingual speakers.

Table of contents

Appendices

List of tables

List of figures

1   Study background

Student mobility across borders has changed the landscape of Canadian higher education as a result of institutional efforts to build global competencies in research and scholarship, to contribute to strengthening the national labour force, and to bring in revenues to offset declining domestic enrolment. As of 2017, the survey of the Association of Universities and Colleges of Canada estimates that 89,000 full-time visa students enrolled in undergraduate programs and 44,000 full-time visa students enrolled in graduate programs in Canadian universities. Since 2000, both figures have tripled from estimates of 22,300 and 18,000 respectively. In addition to increased international student mobility, a large number of English language learning students enter higher education institutions from Canadian secondary schools as immigrant students. Furthermore, these 1.5 generation immigration students make up important student demographics in Canadian institutions.

Whether these students are admitted to universities through meeting English language test or residency requirements, what all these students have in common is that their academic English language competency may need to improve in order to handle the demands of their programs of study. While the internationalisation movement has enriched diversity across Canadian campuses, many institutions recognise that integrating intercultural elements into teaching, providing necessary support for English language learners (ELLs), and ensuring their academic success, require a concerted effort that includes evidence-based practice through systematic research.

Nevertheless, most admission policies across Canadian universities determine ELL students' language proficiency based on test scores from standardised language proficiency tests (such as IELTS, TOEFL, or MELAB) or alternative requirements such as language course credits or language residency requirements. Having multiple options (including the non-testing option) available for students reflects the Canadian educational culture to some extent, in the sense that standardised testing is not the only source of evidence used for determining university applicants' language proficiency. This may result in a lack of transparency and justification in the decision-making process for admission, and more importantly, in determining appropriate language support programs for students with diverse language needs. Furthermore, test score users need more in-depth information regarding the scores used for admission in terms of what students can typically do at different proficiency levels, and how they are expected to perform on discipline-specific academic tasks after admission (Hyatt & Brooks, 2009; McDowall & Merrylees, 1998; Milanovic & Weir, 2010).

Such test score users' concerns about meaningful test score use both for decision-making and resource development for student support after admission are integral to contemporary validity arguments (Bachman, 2005; Chapelle, Enright & Jamieson, 2008; Kane, 2006; Taylor & Weir, 2012). Yet the reporting of scores from large-scale assessments often take the form of a single aggregate score, or at best a series of numerical sub-scores, providing little descriptive information regarding what test-takers within different score ranges can typically do. For example, over 60% of international visa students admitted to the University of Toronto have taken IELTS; as such, IELTS scores and sub-scores are widely used for admission and language course placement for international ELL students. Furthermore, because most large-scale testing programs such as IELTS are designed to assess overall English language proficiency required for success in higher education, test score users rely on aggregated total scores (some programs also consider particular subtest scores) for admission and language course placement for international ELL students. While IELTS speaking and writing band-score descriptors provide additional descriptive information for test score users, no descriptive information is provided for interpreting IELTS reading subtest scores.

2   Conceptual framework

The present study is grounded in a socio-cognitive validity framework (Khalifa & Weir, 2009; Weir & O'Sullivan, 2011) to guide evidence-based inquiry into validity arguments related to IELTS reading score interpretations and use. This validity framework emphasises the development of a priori validity claims, built upon the cognitive processes that are elicited by test items. These processes are also situated within solid theoretical foundations to respond to cognitive validity claims. The context validity facet of Weir and O'Sullivan's (2011) framework addresses the relevance and authenticity to the target language. A posteriori validity within this framework encompasses evidence generated to respond to scoring validity claims. The scoring validity facet concerns the stability of test results over time, which should be both free from bias and consistently sampled. Additionally, a posteriori validity claims address consequential validity claims to respond to the appropriateness of test score interpretations and use. The subsequent sections of this review elaborate upon each of these facets of the socio-cognitive validity framework as it pertains to this proposed research investigation.

2.1   Consequential validity: Test score interpretation and use

The relationship between test score interpretation and use is interdependent and reciprocal. Although an assessment may be considered a valid indicator of what it intends to measure, it is possible that the results can be used inappropriately (Bachman, 2005). Research on IELTS test score use in higher education institutions shows variations among different test score users (including students, administrative staff and academic faculty) regarding attitudes towards IELTS and perceptions about adequacy of institutions' entry level cut-scores (Banerjee, 2003; Deakin, 1997; Coleman, Starfield & Hagan, 2003; McDowell & Merrylees, 1998; O'Loughlin, 2008; Rea-Dickins, Kiely & Yu, 2007). Research generally converges on admission staff's insufficient knowledge about the meaning of IELTS test scores. Identifying the admission cut scores is often left to administrators who may have little knowledge regarding what the test actually measures, and what the test scores reflect about students' language ability (Coley, 1999).

Coleman et al. (2003) reported similar observations about the inadequate knowledge base of IELTS test score users in different institutional contexts including Australia, the UK and China. These can be attributed to a lack of systematic training among university administrative and academic staff (Rea-Dickins, Kiely & Yu, 2007). More importantly, test score users need more in-depth information about how to interpret IELTS test scores in terms of what students typically know and can do at different proficiency levels (Hyatt & Brooks, 2009; McDowall & Merrylees, 1998). There is a dearth of research on IELTS test score interpretations in the Canadian institutional context, with the exception of Golder, Reeder and Fleming's (2009) investigation examining appropriate IELTS test scores at entry level.

Furthermore, test score users' underdeveloped knowledge base can have a profound impact on resource allocation to support students (O'Loughlin, 2008). For example, Ingram and Bayliss (2007) suggest that IELTS has the potential to be used for language support placement, calling for the increased use of sub-scores for such decision-making. Hyatt and Brooks (2009) found that 74% of surveyed stakeholders in UK universities felt that admitted ELL students require post-entry English language support, yet 64% of that group believed that IELTS results did not supply sufficient diagnostic information for this purpose. The present study is motivated to address such concerns regarding consequential validity in terms of the adequacy and appropriateness of score interpretations, and subsequent actions taken based on scores.

2.2 Intersections between contextual validity and cognitive validity

Previous research clearly indicates a need to support test score users' assessment literacy for meaningful score interpretations and use (Hyatt & Brooks, 2009). Understanding links between IELTS band score levels and academic language demands in real-life learning contexts demands systematic domain analysis that provides "insight into the conceptual and organisational structure of the target domain" (McNamara & Roever, 2006, p. 21), as well as careful cognitive analysis of test items. Analyses of domain and cognitive processes underlying test performance can inform both cognitive (e.g., the extent to which test tasks elicit the cognitive processes critical for successful performance on a given test) and context validity-related (e.g., the cognitive processes underlying test performance should be observable in a real-life language domain) arguments (Weir & Khalifa, 2008). Research on IELTS context and cognitive validity supports its alignment with target language domain needs in general. For example, Weir, Hawkey, Green, and Devi (2012) report that similar tasks and processes were employed for use in the target language domain as in the IELTS reading subsections. However, it has been subsequently reported that the strongest comparability with the IELTS reading subtest was in literal, basic comprehension (Moore, Morton & Price, 2012). In this latter research, critical evaluation of texts and tasks that required reference to multiple sources demonstrated weak comparability (Moore et al., 2012).

Another finding of the investigation was that there are substantial variations of language demands among different disciplines of study. The researchers identified these areas as a potential gap in the IELTS reading subtest and recommended that more studies be replicated at other institutions to add to these findings. Bax (2015) recommended that the IELTS reading test specifications include more items related to global reading and expeditious reading, as these skills may be markers of successful readers. Concerns about the relevance and authenticity of tested skills to the target language use domain are pivotal for establishing validity arguments.

The present study was intended to enhance test score interpretation and use by bringing cognitive and context validity concerns together. This was done to better support students through the generation of IELTS reading can-do proficiency descriptors that tap into different skill profiles across band score levels and through the identification of discipline-specific language and literacy demands used in real-life academic work.

2.3 Scoring validity: Blending MIRT CDM with scale anchoring for enhanced test score interpretations

Issues with scoring validity usually concern the extent to which test results are consistent and free from bias. More importantly, scoring models should allow for the comprehensive representation and interpretation of multivariate knowledge and skills which can be expressed in the form of can-do descriptors. Cognitive diagnostic modeling (CDM) is one such model that brings together two advanced modeling approaches: multidimensional item response theory (MIRT) and confirmatory latent class modeling. It typically assigns individual test-takers into distinct skill mastery classes, α^2, based on whether or not they have mastered user-specified skills required for successful test performance (Jang, 2009; Lee & Sawaki, 2009; Rupp, Templin & Henson, 2010). For example, if a test measures a total of three skills, eight (2^3) different skill mastery profiles can be used to classify individual test-takers. Each skill mastery class represents a distinct multidimensional skill profile and is differentiated in terms of the different combinations of mastered and non-mastered skills. Alternatively, posterior probabilities of skill mastery can be used instead of discrete classifications of skill mastery in order to provide more detailed information about the degree of skill mastery towards user-specified thresholds.

RQ1:   What are test score users' perceptions about test scores used for admission in terms of how these test scores translate into real-life academic tasks?

RQ2:   To what extent do academic language and literacy demands differ across programs?

RQ3:   To what extent do IELTS test scores predict academic outcomes as measured by students' self-reported cumulative GPA and competence/importance regarding their academic language and literacy skills??

RQ4:   What are the characteristics of IELTS reading skill profiles?

RQ5:   What proficiency descriptors characterise IELTS band score levels based on blended CDM profiling with scale anchoring?

RQ6:   How do test score users respond to can-do proficiency descriptors across IELTS band scores and to recommendations regarding university disciplinary language and literacy demands?

Resulting diagnostic profiles can provide fine-grained information about what students can do and what they require to further improve. The quality of diagnostic information from CDM depends on the comprehensive specifications of linguistic knowledge and cognitive skills elicited by test items. In turn, its usefulness for informing test score interpretation and utilisation is subject to the extent to which elicited skills represent real-life language demands that make a difference in students' academic performance. Further, both compensatory (additive) and non-compensatory (multiplicative) inter-skill relationships need to be carefully examined by developing item-by-skill specifications, called a Q matrix. For example, for the non-compensatory model, low mastery on one trait dimension cannot be compensated for by high mastery on another trait dimension (Jang, 2009).

Although skill mastery profiles from CDM can provide detailed information about the specific skills that students have mastered and not mastered, MIRT models are infrequently used for item calibrations. The reporting of test scores is frequently done on a unidimensional continuous trait scale, similar to the current IELTS band scores. Scale anchoring methods are commonly used to provide a set of proficiency descriptors for each test score band on the unidimensional ability scale. Scale anchoring is performed by identifying items that were answered correctly within each proficiency level (i.e., conditional item p-values) and creating a set of anchor items for each score level. Salient skills required for a correct response are then derived through content experts' judgment (Beaton & Allen, 1992; Gomez, Noah, Schedl, Wright, & Yolkut, 2007; Sinharay, Haberman & Lee, 2011).

Considering current IELTS reporting practices, the present study sought ways to bring together unidimensional scale anchoring and CDM's MIRT-based skill mastery profiling, along with domain experts' judgment, in order to identify key descriptors associated with different score ranges. We further examined the extent to which resulting can-do proficiency descriptors have the potential to facilitate test score users' test score interpretations, and to guide prospective students' planning of language support programs and resources.

3. Method

3.1 Overview of research design

The purpose of the current project was guided by the overarching goal to facilitate test score users' meaningful IELTS reading score interpretation and use to make decisions about admission to undergraduate programs at the university. The study utilised a mixed methods research design that includes a series of focus groups, a large-scale survey, and a blended psychometric modeling approach involving CDM and scale anchoring. Specifically, the project sought empirical evidence to answer the following research questions listed on the following page.

The project took place over four developmental phases. Table 1 provides the overview of the research design and specific data collection and analysis activities completed to date.

Table 1: *Overview of the research design*

Phase	Purpose	Method & Participants	Analysis
Phase 1	To examine international students' perceptions about their language proficiency and preparedness for academic language demands To examine university faculty's perspectives about international students' language proficiency and their academic performance	• Focus groups (39 students and 16 faculty members) • Domain analysis of course materials	• Grounded theory approach of focus group data analysis • Content analysis of course materials
Phase 2	To develop IELTS reading skill mastery profiles based on CDM application	Form 153 (Form A; N = 5222) and Form 173 (Form B; N = 8251)	• Q matrix developed • CDM model fit comparison based on goodness-of-fit and parsimony • Diagnostic profiling
Phase 3	To examine the perceived importance of, and self-rated competence in, different language skills To examine the extent to which IELTS test scores predict self-reported academic performance and competence in academic language demands	Campus survey (N=917)	• Confirmatory factor analysis; Latent class modeling; ANOVA
Phase 4	To generate skill proficiency descriptors by integrating CDM-based profiles with scale anchoring To examine the extent to which descriptive skill profiles can enhance test score users' score interpretations and facilitate their discussions about student language support	Focus groups (6 students and 6 faculty members and administrators)	• Scale anchoring application • Thematic analyses of field notes from focus groups

3.2   Participants, data collection and analysis

3.2.1   Phase 1

A total of 10 focus groups (N=55) were conducted to examine test score users' interpretations of admissions test scores and the degree of preparedness felt among international students. Focus groups were intended to capture test score users' beliefs and perceptions about international students' language proficiency, meaning of test scores used for admission, and their academic performance (Krueger & Casey, 2000). The focus groups ranged in size from 2 to 10 participants, with a total of 55 participants, including instructors (n=16) and students (n=39) across three programs of study: engineering, commerce, and economics. For the remainder of the report, these programs are referred to as Engineering, Commerce and Economics. These three programs were targeted due to their high representation of international students at the University of Toronto, who have been found to be largely concentrated in two broad fields of study: business, management and public administration (27%) and architecture, engineering and technologies (19%). All students in the focus groups had presented IELTS scores for admission to the university.

The study participants were recruited by contacting program administrators. If a program agreed to participate, a joint plan was developed to recruit students. The research team and program administrators wrote a student recruitment letter, which each program sent to their respective students. The electronic letter consisted of a description of the research project and an email address for interested students to contact the research team. Once interested, students contacted the research team, and more detailed information about the research project and focus group was provided, as well as a link to an electronic consent form. At this point, the research team collected participant names and managed communication with participants regarding the date, time, and location of the focus groups.

To facilitate focus group dialogue of shared experiences, homogenous groups were created across programs by position and year of study (Morgan, 2008). The result was three focus groups per program: first-year students, upper-year students (2nd year and higher) and instructors. Due to a high response rate from upper-year Commerce students, two focus groups took place for this subgroup, making 10 focus groups in total. The majority of students were from mainland China, with Commerce students demonstrating the most diversity: 4 students from India, 2 from Pakistan, and 2 from Russia. Additionally, one economics student was from Taiwan. Ages ranged from 19 to 24 years, and the group was made up of 29 females and 16 males. See Table 2 for more detailed information about focus group participants.

Table 2: *Composition of focus group participants*

Program	First-year student	Upper-year student	Instructor	Total
Commerce	7	18	5	30
Engineering	2	4	6	12
Economics	5	3	5	13
Total	14	25	16	55

Focus group protocols were developed for each group type (see Appendices A and B). The protocol questions were designed to cover a series of major themes of interest. As first-year students had only been in the university for two months, it was inappropriate to discuss their final exams or transitions between years and programs. Therefore, questions for first-year students focused more on admissions and language test preparation, first weeks in the university, university language demands, IELTS test-taking experiences, and experiences with university language support.

Upper-year students were asked similar questions, with additional questions pertaining to how they've grown over time, how they overcame challenges, and changes in experiences from year to year. Each focus group was video recorded, had an interviewer, note-taker and technical person, and lasted between 60 and 90 minutes. In fact, at times focus groups ran over the allotted time because participants had so much to discuss and wanted to continue. Students were quite candid with the interviewer and one another, which lead the interviews to take on a rather informal, natural tone. Note-takers were instructed to note key observations from the interviews, as well as any initial insights they had. The interviewer, note-taker and camera person routinely debriefed after the focus group sessions to compare their notes and observations for similarities and differences.

Focus group data were analysed by applying a grounded theory approach (Charmaz & Bryant, 2008) and through constant comparison among three groups: first-year students, upper-year students, and instructors. Open coding for initial categories took place until saturation was reached and no new concepts emerged (Coleman & O'Connor, 2007). The data was then axially coded, connecting themes to larger categories within a hierarchical structure of categories and sub-categories (Strauss & Corbin, 1998). Researchers working within each group engaged in constant comparison across groups throughout the process (Glaser & Strauss, 2017), culminating in selective coding of key themes. Findings were re-interpreted and evaluated, examining relationships across all three groups and developing an emergent theoretical framework (Coleman & O'Connor, 2007).

In order to better understand the discipline-specific nature of language demands experienced by students at the university, a content analysis was performed on course materials collected from three participating programs. More specifically, instructors from each program were contacted to provide samples of syllabi and readings from courses they were teaching. To ensure a more representative sample, this original pool was expanded through a purposive sampling approach to include a total of eight courses per program (n=24), with two courses representing each year.

We developed an analytical scheme that included the following categories: overall reading requirements, total number and modality of evaluations, volume of reading required to complete largest evaluation type, text types (e.g. textbook, manual, journal), and text genre and style. From this more detailed analysis, key information that highlighted differences within and between programs were pulled out to create the final comparative matrix which included: overall reading required, text type, total number of evaluations, and evaluation modality/type (see Appendix C).

3.2.2   Phase 2

The purpose of Phase 2 was to examine core knowledge and skills required for correct responses to IELTS reading test items and develop reading skill mastery profiles through the application of CDM. Various CDMs were fit to response data from IELTS Forms 153 (Form A; N = 5222) and 173 (Form B; N = 8251) using the "CDM" package in R (Robitzsch, Kiefer, George, Uenlue & Robitzsch, 2019). The two basic data sources for CDM include response data and a weight matrix, called a Q matrix, that specifies the relationship between items and user-identified attributes. Constructing a defensible Q matrix requires substantive domain knowledge about the target construct in terms of required attributes. The Q matrix specification involves whether or not skill k is necessary for a correct response to item i. Based on the item-by-attribute specification, CDM infers individual test-takers' level of mastery for each skill, given their responses to items associated with each attribute.

A rigorous Q matrix development procedure was utilised to ensure that the attributes selected were "theoretically compelling, empirically sound, and relevant to test use" (Jang, 2009, p. 210). Development was initiated through a review of relevant reading skill taxonomies and strategies prevalent in the literature. Seven content experts conducted item content analyses on Form A and Form B to specify the relationship between the attributes and individual test items. Initial development identified 11 attributes composed of the skills and knowledge hypothesised to be required to successfully complete the IELTS reading test. In the few cases where inter-rater discrepancies emerged, the seven raters collectively reviewed the item, attribute definitions and coding scheme until a unanimous consensus was reached.

The correct classification of attribute mastery profiles requires a sufficient number of items per attribute. Content analyses revealed that the majority of items required explicit comprehension, inferencing and summarising; however, few targeted the other eight attributes. As a result, the initial Q matrix included several attributes measured by two or less items, necessitating further revision due to a lack of sufficient number of items for accurate classifications (Hartz, Roussos & Stout, 2002; Jang, 2005). These attributes measured by insufficient number of items were combined with the most relevant attribute among explicit comprehension, inferencing and summarising. For example, the attribute associated with linking textual information to background knowledge was merged into the skill of inferencing, while understanding text purpose was combined into the summarising skill. Graphic interpretation was removed from the final Q matrix as this attribute was included only in Form A.

CDMs differ from each other in terms of the assumptive nature of inter-skill relationships. For example, some CDMs assume that test-takers have to master all required skills to correctly respond to item *i*. Other CDMs hold a compensatory assumption that the mastery of at least one required skill is sufficient for a correct item response. The compensatory models such as DINO and C-RUM assume that mastery of one or more attributes required for a correct item response can be compensated for by the other attributes. The non-compensatory models such as NC-RUM and DINA assume that a correct item response requires the mastery of all required attributes. The variants of CDM models further differ from each other in terms of parameterisation and estimation algorithm. For example, the DINA model (Junker & Sijtsma, 2001) estimates slipping and guessing parameters for each item with equality applied to all skills. On the other hand, the reduced RUM (Hartz, Roussos, & Stout, 2002) does not constrain equality across skills. Several models were fit to the IELTS response data for IELTS Forms A and B. Appendix D summarises the five CDMs applied to IELTS response data.

Five indices were used to examine the model fit of each model. Regarding Form A, as presented in Table 3, Akaike's Information Criterion (AIC) and the Bayesian Information Criterion (BIC) both indicated that G-DINA demonstrated a good model fit to the given response data, along with C-RRUM. The BIC tends to penalise models with higher parametrisation (Li, Hunter & Lei, 2016), making it particularly informative in comparing model fit. Thus, the BIC finding provides sufficient justification to utilise G-DINA for the CDM analyses for Form A.

Table 3: *Model fit comparison – Form A*

Model	No. of parameters	AIC	BIC	AIC3	AICc	CAIC
DINA	91	233618.7	234215.7	233709.7	233622.0	234306.7
DINO	91	234591.0	235188.0	234682.0	234594.3	235279.0
G-DINA	121	232609.3	233403.1	232730.3	232615.1	233524.1
NC-RRUM	106	235378.2	236073.7	235484.2	235382.7	236179.7
C-RRUM	106	232716.5	233411.9	232822.5	232720.9	233517.9

The root mean squared errors (RMSE) of item correlations for the G-DINA model further support its utilisation in the current study. Measuring the difference between predicted and observed correlations for all pairs of items, the RMSE is a useful model-fit index regarding item and model fit (Yi, 2017). In particular, the average RMSE provides a synthesis of the overall discrepancy between model predictions and observed data, with lower values indicating better fit. The average RMSE for G-DINA was .047, lower than the recommended cut-off score of .05 (Henson & Templin, 2007), indicating adequate fit.

Regarding Form B, G-DINA had the lowest AIC, while C-RRUM had the second-lowest (Table 4). In contrast, C-RRUM had the lowest BIC, followed by G-DINA. This finding corresponds to the statistical characteristic of the BIC scores penalising overly-complex models. Despite the contrasting fit indices across Forms, the G-DINA model was chosen for Form B as well as Form A to ensure comparisons could be made across test forms. Supporting this decision, the RMSE of G-DINA for Form B was adequate at .045.

Table 4: *Model Fit Comparison – Form B*

Model	No. of parameters	AIC	BIC	AIC3	AICc	CAIC
DINA	91	402721.3	403359.9	402812.3	402723.3	403450.9
DINO	91	403173.1	403811.7	403264.1	403175.1	403902.7
G-DINA	109	402046.5	402811.5	402155.5	402049.5	402920.5
NC-RRUM	100	402115.5	402817.3	402215.5	402118.0	402917.3
C-RRUM	100	402071.5	402773.3	402171.5	402074.0	402873.3

3.2.3   Phase 3

The Undergraduate Language Demands Survey was developed to capture several aspects of undergraduate student language experiences. Not only students with an English test score required as part of admission but also those without were invited to complete the survey so that international students' self-reported language competence and perception about language demands in school could be compared to those of domestic students. A total of 917 students responded to the survey between April 2018 and October 2018. Student ages ranged from 17 to 35 years (M = 20.38, SD = 1.74), and approximately 67% of the students were female (n = 615). Among all, 148 (16%) were first year students, with 261 (28%) in second year, 239 (26%) in third, 237 (26%) in fourth, and 32 (3%) students in fifth year and above. Approximately 58% of participants were born outside of Canada (n = 531). Of the 917 participants, 259 (28%) took a language test or program as part of admission into the University of Toronto. The majority of these participants took IELTS (n = 170), followed by TOEFL iBT (n = 53), followed by the University of Toronto's International Foundation Program (n = 12), and TOEFL CBT (n = 8). The remaining students (n = 16) took a variety of other tests.

The link to participate in the online survey developed by the research team was distributed to under-graduate students in various academic units (faculties or programs) on two of the university's three campuses (St George and Scarborough). We approached each of these units with a project description and, for those interested in participating, arranged a meeting to explain the research project in further detail. Upon agreeing to participate, academic units sent out a letter of invitation with a link to the survey to all undergraduate students in their program.

Appendix E provides a copy of the survey instrument. Section A of the survey instrument asked for informed consent and Section B contained a comprehensive set of demographic questions including age, gender, university major and minor, country of origin, year of arrival for those born outside of Canada, Canadian citizenship status, type of high school curriculum completed (within or outside Canada), program year, and a list of students' known languages, including self-reported fluency levels. Sections C–F identified which of 12 prominent standardised tests or programs students had completed in order to be admitted (if applicable). Students provided self-reports of their performance on admissions tests (test score by subsection) and GPA.

Students were also asked whether or not they felt that their test score accurately reflected their English proficiency, and whether or not they felt prepared for academic language demands at the time of entry to university. These sections assessed students' self-reported competence in, and importance of, various language skills and communication activities required in academic life at the university. For each of 32 language skills, students rated how important the activity was in their school work and how well they could do this activity on a five-point Likert scale. In addition to reading skills, which are indeed the focus of this research project, we included some activities related to productive skills, such as speaking and writing, due to the integrated nature of university language demands. Section D was related to students' academic performance and career-related skills. Specifically, self-reported CGPA in university and preparedness for communication and teamwork required for future careers were included in this section.

Survey development consisted of four phases: item generation, pilot testing and revision, full questionnaire administration, and questionnaire validation. First, the demographic variables and latent constructs required to address the research questions were identified. Relevant items were deductively generated based on the theoretical foundation and a thorough literature review of each latent construct (Hinkin, 1998). Subsequently, the language demands section of the survey was shared with student focus group participants, forming a pilot group who were prompted to discuss the ease of interpretability and face validity of the items. Relevant student feedback was integrated into the questionnaire, which was then developed into an online format. The full questionnaire was administered to students in April 2018 (n = 138) and in October 2018 (n = 779).

The survey data were analysed through reliability analyses and exploratory factor analyses (EFA) to reduce 32 observable variables into a parsimonious set of factors underlying the target constructs: self-rated language competence and importance. Confirmatory factor analyses (CFA) were performed to test hypothesised theoretical factor structures based on 32 language competence and language importance items. The strengths, weaknesses, and applicability of EFA and CFA have been the subject of intense debate within the measurement literature (Schmitt, 2011). There tends to be consensus that CFA is optimal when measurement models have a well-defined factor structure, as determined by prior empirical investigations. In contrast, EFA is likely more appropriate for initial scale development, or instances where initial CFA models produce poor fit with a large number of substantial modification indices. Rather than exclusively using one technique, the current study incorporated both EFA and CFA as "a heuristic strategy that builds on the comparative strengths of the two techniques" (Gerbing & Hamilton, 1996, p. 63), in which EFA promoted model specifications that were subsequently cross-validated using CFA (Hurley et al., 1997).

EFA was run separately on language competence and language importance items to determine their respective measurement models. Analyses were conducted in Mplus version 7.4, using the default Geomin (oblique) rotation, as factors were expected to be correlated (Costello & Osborne, 2005). Scree plots and, to a lesser extent, model fit indices were used to determine the optimal number of factors. Scree plots indicated that four-factor models were optimal for both language competence and language importance, paralleling the four-factor theoretical framework utilised to generate items. EFA results were then utilised to identify problematic items, as defined by having prominent cross-loadings, factor loadings below .30, or items with loadings that were not theoretically substantive (e.g., an item targeting reading skill loading onto the productive skill factor). The language competence EFA output suggested five problematic items, which were dropped in the following CFA cross-validation phase. The final four factors in the language competence EFA model results included productive skills, disciplinary literacy skills, general reading skills, and higher-order reading skills.

The language importance EFA results suggested eight problematic items, which were also dropped in the CFA cross-validation. In the language importance EFA, many items from the productive skills factor proved problematic, to the extent that only writing-related items were kept, thus altering the latent construct from productive skills to writing skills. Thus, the final four factors of the language importance model included writing skills, general reading skills, higher-order reading skills, and disciplinary literacy skills.

With the removal of problematic items, two CFA models – one for language competence, the other for language importance – were then conducted to cross-validate EFA results and generate factor scores for subsequent analyses. Model specification followed the default protocol for CFA (i.e., no cross-loadings, no correlated residuals, correlations among factors were freely estimated), with the exception that factor variances were constrained to one in order to freely estimate the factor loadings of the first item in each factor. The language competence CFA (Appendix F) had adequate model fit χ^2 (344) =1761.98, p < .001, RMSEA = .08, CFI = .86, SRMR = .06. Although the CFI was below the recommended .90 cut-off value, RMSEA and SRMR demonstrated adequate model fit. Factor loadings all met the .3 recommended cut-off. The language importance CFA (Appendix G) had borderline adequate model fit χ^2 (246) =1310.00, p < .001, RMSEA = .08, CFI = .84, SRMR = .08, with all items demonstrating factor loadings of above .3. Given adequate measurement models regarding both language competence and language importance, factor scores were generated and used in subsequent analyses.

In further analyses, a series of one-way analyses of variance (ANOVA) and chi-square difference tests, followed by post-hoc tests and effect size estimation where applicable, were conducted to compare self-rated language competence and importance, academic outcome, preparedness for university language demands before entering university, and preparedness for career skills (i.e., communication, teamwork) across different subgroups of students. Stata version 15.1 was used for all of these statistical analyses.

3.2.4 Phase 4

The purpose of Phase 4 was twofold: to develop skill proficiency descriptors across IELTS reading band score levels based on the application of a scale anchoring method to CDM-based profiles, and to examine the extent to which descriptive skill profiles can enhance test score users' score interpretations and facilitate their discussions about language support for students. The scale anchoring method (Beaton & Allen, 1992) is used to develop proficiency descriptors to enhance users' score interpretations. It has been widely applied to large-scale testing programs such as the National Assessment of Educational Progress (NAEP), the Trends in International Mathematics and Science Study (TIMSS), and the TOEFL iBT test.

The development of skill proficiency descriptors for the IELTS reading band score levels included the following steps:

1. specify the score proficiency levels

2. identify anchor items by calculating conditional proportions of correct responses for all items at each proficiency level and select items whose conditional p-values are .65 (.7 after rounding) and conditional *p*-values at the given proficiency level that are different from the adjacent lower level by at least .2

3. integrate diagnostic discrimination indices from CDM

4. create proficiency descriptors from the pool of anchor items at each score level

5. create a list of recommendations that highlight academic language and literacy demands required for successful academic work after admission.

As a final part of the sequential design of the study, two focus groups were conducted that were confirmatory in nature (Teddlie & Tashakkori, 2006). The two focus groups – one with students and the other with faculty members and administrators – sought input on two documents developed through the scale anchoring process: Proficiency Descriptors for IELTS Reading Band Score Levels 5.5 and Recommendations for Incoming Students.

Students were recruited whereby those who had taken the Undergraduate Language Demands survey were prompted to indicate their interest in participating in a follow-up focus group. Those who expressed interest and had taken the IELTS test for admission into the university were contacted. The purpose was for students to discuss their language experiences at the university, and to provide feedback on two documents: 1) Proficiency Descriptors for IELTS Reading Band Score Levels 5.5 and Above and 2) Recommendations for Incoming Students. The focus group protocol, found in Appendix H, covered the following topics: student preparedness for university reading; student responses to IELTS reading scores/report; key reading skills required within different programs; most challenging reading skills within different programs; and feedback on the documents circulated. A moderator led the focus group, two researchers took notes, and another audio-recorded the session. The focus group was scheduled to be 60 minutes, but participants stayed approximately 90 minutes.

The second focus group included faculty members from two programs, as well as administrative representatives from the Office of the Vice President International, the Faculty of Arts and Science Registrar's Office, and a representative from Enrolment Services, a division under the Office of the Vice-Provost Students. This number balanced the need for sufficient diversity across participants, while maintaining a smaller group size in order to encourage sharing and participation (Onwuegbuzie, Dickinson, Leech & Zoran, 2009). The goal was to share findings from the project, solicit feedback, and discuss potential opportunities for collaboration and next steps. The focus group therefore followed a somewhat different format, with three presentations, each followed by opportunity for discussion, as outlined in Appendix I. The first presentation provided a summary of initial focus group findings (student and instructor) from Phase 1 of the project, the second presentation provided information regarding the survey findings and subsequent analyses, and the third presentation summarised the scale anchoring findings and proposed enriched communication to incoming students. One moderator led the focus group, several researchers presented the findings and multiple team members took notes.

Data from the focus groups were again analysed for key themes through open coding (Coleman & O'Connor, 2007) and constant comparison (Glaser & Strauss, 2017). Additionally, due to the more targeted nature of this focus group, feedback regarding the documents was prioritised and discussed among the research team and used to revise the skill proficiency descriptors. In the following section, research results are presented according to inquiry questions rather than by phase.

4. Results

4.1 RQ1: What are test score users' perceptions about test scores used for admission in terms of how these translate to real-life academic tasks?

Based on the grounded theory analysis of 10 focus groups involving international students and faculty members, we developed a framework of international students' academic language proficiency and experiences (see Figure 1). The framework was used to capture the phenomena by first contextualising phenomena through both external and internal factors, then describing teacher and student dispositions, then elucidating tensions and areas of discord resulting from comparisons between teacher and student experiences, and finally discussing consequences. The following findings highlight themes of students' experiences with navigating the language demands within the university context.

4.1.1 Contextual mandate

In order to better understand the phenomena described, it is essential to situate international students' experiences within the broader context. In 2014, the Government of Canada launched Canada's International Education Strategy, pushing Canada to be a leader in international student recruitment (Government of Canada, 2014). In 2017, the number of international students in Canada was 494,525, a growth of 88% from the 2004–05 school year, representing 11% of Canadian university students at large. At the university within the study context, 20.5% of the undergraduate population are identified as international students (University of Toronto, 2016–17). According to the report, students attend this university from as many as 165 countries and regions, with the top five locations of origin among undergraduates being: China (63%), India (4%), South Korea (3%), the United States (2%), and Hong Kong (2%).

Consequently, the majority of international undergraduate students originate from countries where English is not the societal language, making commercial language testing, such as TOEFL or IELTS, a common requirement for entry into the university. Additionally, with English (and French) as the official languages in Canada, international students are not only navigating schools in a foreign language, but are also immersed in an English-dominant context more broadly, within which they must meet all other needs. Although some students are comfortable navigating this context, and, for example, find successful employment which may anchor their experiences and allow them to interact comfortably within the local context, others face tremendous difficulties; for example, some students described challenges ordering food within this different language context:

> I still remember, I took my IELTS test in early 2017. But, during summer holidays [right before entering the university], I came to Canada for campus tour for the first time. And, I cannot even order pizza in PizzaPizza. I don't know it's called pepperoni. I know it's kinda shameful, but…that's me after taking IELTS, I passed it, but…I still cannot English. (First year Commerce student)

CAUSAL FACTORS

DISPOSITIONS

TENSIONS

CONSEQUENCES/OUTCOMES

External Causal Factors

- *Globalization and increased movement of students desiring degrees abroad*

- *Canadian Government's push for increased international student presence at the university*

- *Commercial language tests advertising accurate measurement of academic English proficiency*

- *English speaking context*

INTERNAL

University Admission requirements

Student preparedness (Test scores, language competency)

Program demands (may vary across courses)

-Program language Requirements for entry

Supports available

STUDENT

Student Characteristics and ability
Student preparedness (Test scores, language competency etc.)
Student outcome pressures (grades, employment etc.)

TEACHER

Teacher characteristics
Teacher expectations

TENSIONS

1) Mismatch between required (IELTS) and necessary (program-specific) skills

2) Teacher vs. student views regarding motivation for course selection

3) Teacher vs. student interpretations regarding preparedness

4) Supports available vs. needs met

5) Divergent teacher perspectives: all students struggle in first year vs. diverse variety of student struggles

6) First year vs. Second year +

7) International peer group vs. domestic peer group interactions

8) Teacher views of assignments vs. student views of assignments (e.g. time interpreting assignment)

9) Teacher views re. student struggle (speaking, vocab, comprehension and critical analysis) vs. Student view

10) View that students read for the wrong purpose vs. students struggling to meet reading demands

MITIGATING FACTORS

English background experiences
Utilization of supports
Interactions with same-language peers/other language peers
Outcome pressures
Number of years in program

STUDENT

- Students feel unprepared/lost, with different expectations than those required to enter program
- Students experience psychological side-effects of social isolation, lack of integration, lack of self-confidence, low motivation, shame
- Students feel overwhelmed by workload and ability to meet course demands
- -Students develop strategies targeting assignment completion and grades vs. more comprehensive knowledge development in their field
- Students engage in course selection based on ability to understand professor
- Despite attending local high schools, some students still struggle with academic and non-academic language
- Students in non-Co-op programs feel less prepared to compete in the workplace
- Students feel frustrated that their knowledge is not being recognized due to language barriers
- Many student difficulties remain unaddressed

TEACHER

- Teachers misinterpret student motivations
- Teachers do not know how best to support heterogeneous International student body
- Lack of identification of those requiring additional supports

PROGRAM

- Program graduates who have not performed up to full potential
- Lack of cultural knowledge creates equity issues in curriculum fairness
- Students able to push through challenges have greatly improved experiences in later years

SUPPORTS

- Students are often unaware of supports and those supports available don't meet student needs
- Perceived lack of adequate language supports

4.1.2 Language proficiency required for university admission

A critical question raised by both students and faculty concerned whether or not international students enter university with the language proficiency necessary for successful academic language tasks across disciplines. A number of faculty members expressed that they were not aware of the cut-off scores used for admission of international students. Faculty members across all three programs expressed their concerns about international students' academic performance and social interactions with others in, and outside of, classrooms. In addition, students had little knowledge about what the IELTS test score, 6.5 (the current cut-off score set by most programs at the University of Toronto), meant other than recognising that it was the required score for admission. Interestingly, some students noted that they benefited from language test preparation and tests, and that these were related to test materials similar to their fields of study. For example, a couple of Engineering students agreed that IELTS helped them prepare for writing lab reports, and a Commerce student explained that these helped her understand how to interpret and write about graphs. However, a mismatch between general language demand requirements upon entry and disciplinary-specific program demands was frequently mentioned as one of the most critical challenges students faced upon entry.

This issue appears to be an inevitable consequence of the current admission policy, that adopts external standardised general language proficiency tests for admission requirements. A first-year Engineering student explained:

> For reading questions in IELTS, we know the specific questions we need to answer after reading it and there are some key words we can search in the paragraphs. But, umm for the reading in like in university, umm, maybe the articles are not so structured like the readings in IELTS. So we don't know where to find the information all.

A few students also pointed out that the lectures in IELTS listening were not authentic to their university experience. One Economics student explained how professors at U of T use idiomatic and culturally specific language, which makes university lectures far more challenging than the IELTS listening.

> Even I got satisfied marking on IELTS test, I still feel the knowledge is not enough for here. In the lecture, the professor usually use something really familiar with you because you are local people. Use some words or some examples, but I'm so "what's that? what's that?" So I search google for that maybe a local team, local brand. The name of coffee and I think some, the nouns, they are difficult to remember.

Another Economics student pointed out that IELTS listening has speakers with different British accents, which are not as common in Canada.

> The language…IELTS is like a British language, right? It's a British accent and the basically listening is British. And here we're in Canada. It's like kind of North American. And I kinda feel like that's different, like not connected.

Two Economics students pointed out that IELTS writing demands are quite dissimilar from university demands, particularly with regards to writing, by stating that "before I came here the longest passage is 300 words. But for here, 30 page! Oh! How can I do that?" (Economics student). The interviewer then asked for clarification, specifically regarding if they would prefer that IELTS had a longer writing section. The student responded that they believed it would help if the IELTS writing section were longer.

4.1.3   Meaning of test scores

Regarding the question of how well test scores reflect students' English language ability, most students who participated in the focus groups responded to the question negatively, partly because of language-related challenges that they had experienced after entering the university. Overall, students agreed that despite their high IELTS scores, they encountered various language-related challenges, especially with speaking and writing. One Engineering student in second year reflected upon his first school year as follows:

> After I came to U of T, personally I found the most difficult part is writing, so even though I got the highest on writing, that's umm, that's the part I find the hardest… During the university, I participated in some presentation competition and I got some awards, so I think I don't really have a speaking problem, but the writing part, it's still really tough part.

Another Engineering student had a similar experience with her listening and reading abilities. She explained:

> I did really well in listening and reading. It's almost close to a full mark. So, I was confident in reading. But, when I came here, umm so praxis course requires us to do some research to read a lot of documents. I feel like, for example, in two hours, my partner can complete a research, but I can only do like half of them. So, when I read a website or something I cannot know where is the focus, yeah.

Most students described repeating the IELTS test to receive admissible scores, indicating inflated test scores due to practice effects. For example, several Engineering and Commerce students took IELTS at least twice, with two Engineering students having taken the test four times and two Commerce students having taken the test three times. Many of these students explained that they struggled to get an appropriate score on writing, with one Commerce student complaining that one hour was simply not a long enough time for them to write. Another Commerce student explained that it was harder to improve writing in a non-English speaking country. Several Commerce, two Engineering and one Economics students felt that IELTS speaking was the hardest part. One Commerce student explained that their English lessons had always been in Russian, so they did not have much experience in speaking.

On the other hand, listening was considered quite easy by comparison, according to several Commerce students and an Economics student. Like listening, a few Commerce students also felt that reading was easier than writing and speaking because they could practice for it. An upper-year Engineering student had similar opinions regarding IELTS reading:

> And especially for IELTS reading, if you use the technique you can get a high mark but it doesn't really mean you really like can read a passage and understand it fully. It just means you know how to do questions.

Students identified a lack of feedback on their IELTS scores as problematic. One first-year Engineering student explained her struggles with taking IELTS multiple times.

> The test did not indicate how we need to learn English. Like I said I get 5.5 three times. But, I still don't know how to improve it. And I just need help from courses, from teachers, and follow what they said and I got improve.

International student participants in the study were quite candid, and many expressed their appreciation in being provided the opportunity to share their experiences. Students talked both about academic and non-academic challenges that they had experienced, although it was evident that at times there was significant overlap between these two areas. Contextual factors also influenced students' classroom experiences, particularly as they related to background knowledge, where a number of students noted the impact of not having a Canadian or North American background placed them at a disadvantage to understand and participate in class and complete assignments.

> *I think we lack Canadian background knowledge, making it harder to related to some textbooks and instructors' examples.* (Commerce)

> *When you talk about different companies, native peers already know them, but I didn't know this company...Case-based courses based on Canadian and North American cases…* (Commerce)

> *I was not prepared particularly for local examples or jargons…* (Economics)

Students also described challenges related not only to the volume of reading, but also being able to comprehend the volume of text within the time required. One student in particular described their experience trying to understand instructions for a class assignment.

> *I found the instructions for an assignment incredibly difficult to understand. I ended up spending 24 hours just reading instructions and trying to understand the instructions, partially because some of the words in them had similar meanings. I felt that maybe the words would be better understood by native speakers.* (Engineering)

Difficulties students experienced with speaking in English were pervasive across programs. Students described difficulty speaking in class both with teachers and with domestic peers. Speaking with peers had particularly damaging outcomes for students, as they described decreased confidence and feelings of isolation as a result of poor or little interaction with peers.

> *It is awful. We didn't have enough practice. People don't understand me. It's much easier to talk to international students than native speakers.* (Commerce)

> *Conversations were very hard to continue. I could only make very basic small talk. I didn't understand jokes or slang.* (Engineering)

> *Understanding peers is more difficult than understanding instructors. Peers tend to be more vague when they talk.* (Commerce)

Several students explained that the speaking section of the IELTS test did not prepare them for social language requirements outside of the classroom. Two upper-year Economics students explained that IELTS speaking was comparable to formal university settings, but not informal ones. One student explained that "in formal ways, [IELTS] it's really good. But in informal settings, not at all. Not at all". The student further clarified what he meant by informal setting by stating:

> *Yeah, and slang, which you have to learn from the very beginning. And also in tutorials, which requires a lot of communications. (…) In lecture…the language the professor used was very academic which was good.*

4.1.5 Faculty perspectives about international students' challenges with academic language

Faculty focus group participants varied widely in their perceptions of academic challenges for international students. However, one common thread, especially in the Engineering focus group, was the feeling that academic challenges were not specific to international students, but universal to all students because, compared to high school, different skill sets are required at the university level. They explained that, regardless of their visa status, all students struggled with learning the language of the discipline, reading critically, interpreting academic articles, raising questions, summarising what they had read, and writing coherent academic papers. One instructor in Economics emphasised that all first-year students seemed bewildered as they were not used to these kinds of academic demands.

> *Now to the reading questions, I think that this is a general problem. I think that even our uh, like uh, even domestic students have problems. In my course, uh, I have lots of readings, many readings and uh, and students are surprised when I tell them that readings are required.* (Economics)

> *But on that kind of preparatory uh reading which I think is especially crucial in the third and fourth year courses, I don't identify that as an international student issue at all um, plenty of domestic students or native speakers are not understanding how to read, uh and are not doing it or trying it sufficiently. So for that kind of reading, I think we have a challenge for all of our students and I don't, I wouldn't single out international students there.* (Economics)

However, a couple of professors in Economics acknowledged that the situation could be slightly worse for international students. However, most of the instructors across disciplines admitted that international students had more issues with oral communication and writing, rather than reading; some instructors even believed that no students had problems with reading. Although instructors felt that both oral and writing skills were problematic, most of the anecdotal stories were centred on oral communication. One Engineering instructor argued that students needed to learn to interact with others to see where the field is going. Commerce instructors felt that international students underperformed in projects with real clients because they misunderstood what they were contracted for, where high sensitivity and communication skills in unfamiliar contexts are required. Another Commerce instructor perceived international students as struggling with the vernacular, resulting in not (or mis-) understanding the instructor's humour. They also complained that many Chinese students, who represent over 50% of international students in their courses, were less comfortable with oral communication and, therefore, did not speak up in class. One instructor in Economics described a moment of shock for them when they realised how low some students' oral English proficiency was after a year, or even four years, in the university.

Another theme that emerged amongst Engineering and Economics instructors was that international students take a different approach to their reading and writing, as well as have a different general learning style. They believed international students often read for the wrong purposes and have trouble with critical reading and writing skills. Engineering instructors found that their international students tended to rely on modeling and emulating, rather than work that required more critical thinking skills, and were not able to interpret what they had read and explain how different concepts were linked. One instructor in the Economics program said that their international students struggled to understand implicit information and had difficulty actively comparing articles or commentaries with different points of view. Most instructors across all programs felt that international students were not good at inferring, developing their own opinions, and presenting coherent arguments based on problem solving in writing.

However, a couple of instructors pointed out that their struggles with critical and inferential reasoning may be attributable to their cultural and educational backgrounds rather than their language competencies. For example, one instructor reasoned that some students from Asia had not been encouraged to question, but rather, had been trained to listen, absorb, and memorise discrete knowledge.

> *I think that there is a lot of, uh, relying on modeling for those students, um, so when we give them models or they have models to work with they are more successful but also less capable of actually doing it on their own, so we are kind of caught in a bit of a dilemma. So what does that look like, so is it the modeling encourages a lack of critical thought or is it let them flounder?* (Engineering)

Another common issue amongst instructors was the negative perception of international students' motivational characteristics, although some hard-working students were being recognised. Many international students in Engineering were viewed as not putting much effort towards improving their English because they planned to receive their degree and practice in their own language. Instructors described these students as unmotivated to read; instead, relying on lecture, video, and friends for their assignments and exams. Most of the Economics instructors said students did not invest enough time in their school-work and, as a result, some students' English never improved even after a few years at the university. A few instructors also noted that their international students tended to not participate much in class, did not come to ask for help during office hours, and did not put much effort towards interacting with domestic students. Overall, these findings indicate that while instructors argue that international students have the same academic demands as domestic students, they do see a multitude of issues and struggles that are specific to international students.

> *I see a lot of variation…I'm at least as astonished in the improvement in English language skills amongst at least some students that I see, um, it can be pretty astonishing how, how much they improve…but by the same token, I still, you know, at the beginning of the year it's pretty surprising and then even at the end of the year, I still come across students and I wonder how, how they're coping.* (Economics)

Instructors were aware that international students would choose courses strategically by avoiding courses that require critical thinking skills and discussions. For example, some business courses only have a small number of international students in them. One instructor commented:

> *I think it's consistent with what I see in the preliminary level course that requires a lot of reading. Students are selecting.* (Commerce)

4.1.6   Consequences

Students shared strategies that they used to cope with academic language demands in their courses. Some strategies that were mentioned included: "smart reading" or skimming, reading every morning, switching texts when struggling, taking naps regularly, watching YouTube first and then reading text after, reading and taking notes, and using reading strategies learned from IELTS. Another strategy commonly used was translating texts into their first language. Students in Economics all felt that it was challenging to translate text into Chinese because the translation was often inaccurate. Two students discussed the tension between translating texts in order to receive better grades vs. the need to practice their English to get better. Some students described using Google translate and Grammarly. Several Commerce students used a drop-in communication support program targeting ELL students (e.g., Rotman's communication café), for information sessions and resume development/networking. Students also described participating in groups that provided summaries of readings, both online (e.g. Facebook), as well as in-person student study groups. Several Economics students described benefiting from working with international student teaching assistants (TAs).

A number of students described the limited resources and support available to them at the university. In particular, students described concerns regarding booking appointments, which can take up to two weeks, as well as too many students attending TA office hours, and the amount of time during appointments spent explaining problems vs. getting help. One student described a physical barrier related to living off-campus and distance to access support. Another student described a different kind of barrier, describing feeling disconnected as a reason not to use support. Finally, many students were not aware of the support available and expressed interest in learning more about them. Students also discussed the types of support they wanted, including improvements to what is currently offered. For example, making help on a drop-in basis (vs. appointments), as well as giving students the opportunity to send work in advance for more efficient feedback during booked sessions, and having TAs that spoke their language. Students who were unable to access the intensive, pre-university, English program provided in August expressed interest in such a program occurring later in the year. One student wanted more guidance around reading, including information regarding what is and isn't important, as well as summaries of reading materials. Another student described the potential benefits of a language exchange or mentorship between international and domestic students.

Upper-year students, in general, stated that language becomes easier for them as they progress through their respective programs. While they found that upper-year courses were still difficult, they did not encounter the same language difficulties that they had in their first and second years of the program. Some of these students, however, identified language difficulties in the third and fourth years of their programs. For example, an Engineering student stated that the volume and difficulty of reading increased greatly in his third and fourth years and he did not feel that the reading demands of his first and second-year courses adequately prepared him for subsequent demands. Some students also mentioned that they could apply some high school knowledge during the first years of their program, which somewhat mitigated their language challenges. Commerce and Economic students felt that their ability to handle coursework improved during their program; for example, classes become smaller in the upper years and group work, presentations and more interactive classes (e.g. participation) made classes easier. They felt that they had acquired vocabulary, providing them with confidence. Other students found that they had learned how to use non-verbal language and found it easier to make friends. Some students felt that they did not need to worry so much about GPA and spent more time socialising.

Upper-year students said that they made course choices based on how difficult it was for them to comprehend the professor. Students had different opinions about native and non-native speakers. An Economics student said that "I try to avoid professors with thick accents" while another Economics student said that non-native speaking professors are easier to understand than native speaking professors because they use simpler language. One Commerce student said that she could not understand professors while another Commerce student said that professors speak too fast. Commerce students stated that the academic skills they are required to have include critical thinking when analysing texts, which in the Commerce program might include business case studies, legal documents and consulting documents. Some Commerce students were concerned that their knowledge was not being recognised due to language barriers; for example, they said they sometimes had difficulty demonstrating their thoughts in writing.

Upper-year students raised issues with their lack of cultural knowledge. Commerce students found that they were less familiar with Canadian and American business culture or practices than domestic students, which at times provided an additional challenge in completing coursework. They stated that when American or Canadian business case studies were dated, they did not have the Canadian background knowledge to understand the content. One student recommended that programs should be created with consideration of international students' language difficulties.

4.2 Q2: To what extent do academic language and literacy demands differ across programs?

To examine differences in language and literacy demands across the three programs (Commerce, Engineering, Economics) that participated in Phase 1 of the study, two different data sources were used: focus group discussions and content analysis of course materials. The goal of this analysis was to determine both within (across the years), and between, program variation. As part of student focus group discussions, we used a short self-assessment questionnaire to stimulate participants' thinking regarding different reading skills and strategies and levels of relevance and difficulty of these skills and strategies (Appendix J).

With regards to course materials, students described a range of volume and types of materials. Materials mentioned included: textbooks (all programs), annual reports and financial statements (Commerce), video (Economics) and articles (Economics). Three Commerce students described finding the volume of text overwhelming and challenging. One Engineering student found the textbook too hard, while two other Engineering students described not having much reading assigned, indicating a great deal of variation across courses within the same program.

Students also discussed discipline-specific reading demands. Amongst Economics students, one found discipline-specific vocabulary challenging, while another found the writing format within the discipline to be problematic. For Engineering students, the math content came up as challenging, as well as the instructions provided on certain assignments, although another Engineering student did not find the instructions to be problematic. Several Economics students also described challenges related to lacking the appropriate background knowledge to comprehend texts and finish assignments.

In discussing specific skills, those mentioned as most important included: explicit comprehension, implicit comprehension, interpreting graphs, discipline-specific vocabulary, combining texts from different fields (e.g. financial statements, economics theories etc.), understanding context especially for case studies, performing critical evaluation, understanding the audience for different readings, and summarising, especially for writing reports/convincing customers. There were some differences amongst students both across and within different fields. For example, one Commerce student found implicit comprehension to be important, but another Commerce student found this to be the least important skill. However, there was consensus among Economics students that this was least important. Although one Commerce student felt it was important to understand the audience for different readings, another found the purpose of text to be least important and there was consensus among Economics students finding the purpose to be least important. In addition, one Commerce student found text organisation to be least important.

The content analysis of course materials collected from three participating programs showed differences both within, and across, programs (see Appendix C). The major trends that emerged for Commerce regarding reading type and volume suggest that among first-year students, they largely read textbooks, where in second year this expands to include journals and in third year, to include case studies, thus requiring a shift in reading skills from the ability to tackle more expository text, to utilising more critical and analytical thinking skills. Economics follows a similar trajectory, from first year, which is almost all textbook reading, to second year, which includes some journals but is still predominantly textbook driven, to third and fourth year, where journals become the more dominant reading requirement. In Engineering, however, across all four years, the majority of all readings are textbooks, with some additional online manuals, thus all reading follows a more expository structure within this program.

Although all programs relied heavily on textbook readings, these texts varied significantly in content. Examining differences between texts more closely revealed interesting disciplinary-specific differences between programs. Texts in Commerce were largely descriptive, providing details and instructions, and were predominantly text-based. In addition to text, these readings included some formulas, figures, graphs and tables. Economics text samples included texts that were descriptive, and those that followed a cause and effect format. These texts included less textual content than that noticed in Commerce texts, with a significant number of formulas, figures, graphs, tables and maps. Engineering texts were also descriptive and included many problems. These readings were less text-heavy, and consisted mostly of flowcharts and formulas, with some including figures, tables, and diagrams.

Distinct differences regarding how students are evaluated were evident both as students progress through the years within a program, as well as between programs. First-year Commerce students engage in independent work, the majority of their grade coming from more traditional exams and quizzes. In second year, however, although the majority of the evaluation still comes from exams and quizzes, group assignments are introduced, suggesting a shift in emphasis towards productive language skills that emphasise communication. In third year, students engage in a more diverse variety of tasks, still including exams, but also both individual and group assignments, pushing them to engage in analytical thinking, writing and presenting. This continues into fourth year, where some courses no longer provide traditional exams, but instead might include more flexible formats, for example, take-home reflections. In the Economics courses sampled, assessments followed a somewhat similar trajectory to Commerce, where exams dominate in first year, and become one of a multitude of smaller and more diverse assessment types in second year (homework, tutorials) and third year (problem sets, individual assignments, quizzes), yet all of these requirements are assessed individually, with no group work evident until the fourth year.

Similar to Commerce, not all fourth year Economics courses have exams, for example, one course evaluated students largely based on writing assignments and reflections. This again signals a shift in focus from reading for memorisation and to answer multiple choice questions, to more productive language requirements through increased writing and presentations. Across all four years in the Engineering, courses followed a similar make-up, with the majority of the evaluation stemming from midterms and exams, at times 100%, and a smaller percentage of the mark made up of quizzes, assignments, tutorials and projects. These assignments appeared to be more hands-on tasks, suggesting an emphasis on both memorisation of expository knowledge, but also knowledge demonstration and application. Based on the sample of courses reviewed, the majority of the work in Engineering programs across all years was independent, with only one course providing a group project in the fourth year.

4.3 Q3: To what extent do IELTS test scores predict academic outcomes as measured by students' self-reported cumulative GPA and competence/ importance regarding their academic language and literacy skills?

4.3.1 Self-rated language competence and importance

Self-reported language importance and competence were compared across students with different language background status and academic programs using factor scores generated from CFA models. Because the disciplinary literacy factor incorporates a variety of program-specific skills, results based on the disciplinary literacy factor scores would be difficult to interpret; thus, this variable was excluded in the subsequent analyses.

In determining students' language background status, the legal status of residence (i.e., international students with a study permit, permanent resident or citizen) may not be the best categorisation method to distinguish students who might require language support, as many international students come from English-speaking countries and many permanent residents or citizens may have learned English as an additional language. Therefore, students were categorised into three different language status groups according to their country of birth and whether or not they took one of the English language tests or programs as part of the admissions process: (1) students who were born outside of Canada and took a language test or program (Born outside with test); (2) students who were born outside of Canada but were not required to take a test or program (Born outside without test); (3) students who were born in Canada and were not required to take a test or program (Domestic). The Born outside with test group (n = 241, 26%) represents most of the international visa students from non-English speaking countries, except for those who completed Canadian high school curriculum outside of Canada. The Born outside without test group (n = 290, 32%) includes both students who immigrated and graduated from high school in Canada, those who completed Canadian high school curriculum outside of Canada, and those who came from an English-speaking country. The Domestic group (n = 365, 40%) is mostly comprised of those who were born and raised in Canada; however, this group also includes some second-generation immigrants who may have learned English as an additional language. There were also a relatively small number of students who were born in Canada, but were required to prove their English proficiency through a language test or program (n = 21, 2%). These Canadian-born students who had to take an English language test might be those who had left Canada when they were younger and came back in the middle of high school or in the beginning of university. For ease of interpretation, these students were removed from the analyses in this section.

A series of one-way ANOVAs were conducted to examine group differences regarding self-reported competence and importance in different skills of academic language demands (Figure 2). Regarding self-reported competence in academic language skills (see the upper part of Figure 2), results showed that the effects of language background status on self-reported competence in general reading skills and productive skills were significant, $F(2, 576) = 7.48$, $p = .001$, $F(2, 576) = 6.95$, $p = .001$. The strength of these relationships was relatively small, as indexed by η^2 of .025 and .023, respectively. Post-hoc Bonferroni tests indicate that, for both general reading and higher-order reading skills, the Born Outside With Test group and Domestic group differed significantly from each other. The Born Outside With Test group's self-rated competence in these two skills were significantly lower than the Domestic group. On the other hand, effects on competence in higher-order reading skills were not statistically significant, $F(2, 576) = 2.73$, $p = .07$.

The lower part of Figure 2 illustrates the results of the ANOVA of perceived importance of language skills across language background status groups. Results indicate that only the perceived importance of higher-order reading skills was different among the three language status groups, $F(2, 632) = 7.96$, $p < .000$, with a small effect size, $\eta^2 = .025$. A post-hoc Bonferroni test showed that the Domestic group rated the importance of higher-order reading skills significantly higher than the other two groups.

Figure 2: *Comparison of self-rated language competence and importance by language group*
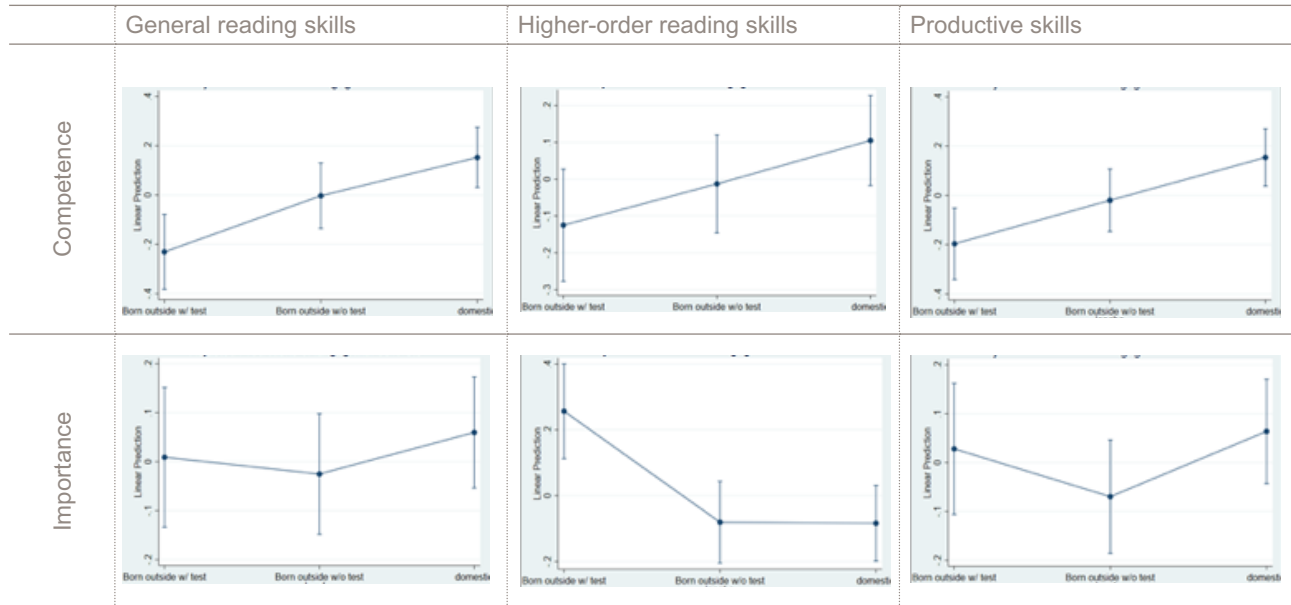
Subsequently, the differences in self-rated language competence and importance across academic programs were investigated. The participating students had a variety of majors and minors, with the majority of students majoring in Engineering (n = 278, 30%) and Life Sciences (n = 243, 27%), followed by Commerce and Social Sciences (n = 122, 13%), and Computer, Physical, and Mathematical Sciences (n = 97, 11%). Students who reported majoring exclusively in the Humanities (n = 6, 0.7%), in Architecture (n = 4, 0.4%), and in two or more distinctive disciplines (n = 132, 14%) were excluded from the analyses in this section, along with those whose program response was missing (n = 35, 4%).

A series of one-way ANOVAs compared self-reported competence in, and importance of, language skills required in university settings (Figure 3). Findings regarding language competence, illustrated in the upper part of Figure 3, suggest that students in Commerce and Social Science reported the highest competence across all three language skills, followed by Life Science and Engineering, while students in Computer Science and Physical and Mathematical Science had the lowest self-reported competence. However, among the three skills, general reading was the only skill that significantly differed across academic programs, which only had a small effect size, $F(3, 478) = 3.96$, $p = .008$, $\eta^2 = .024$. Students in Computer Science and Physical and Mathematical Science indicated significantly lower competence than those in Life Science.

Students demonstrated more variation in the perceived importance of different language skills across programs (the lower part of Figure 3). There was a significant difference in perceived importance across programs for all three skills: general reading skills, $F(3, 522) = 8.40$, $p < .000$, $\eta^2 = .046$; higher-order reading skills, $F(3, 522) = 10.44$, $p < .000$, $\eta^2 = .057$; and productive skills, $F(3, 522) = 2.84$, $p = .037$, $\eta^2 = .016$.

The subsequent post-hoc tests revealed that students in Life Science rated general reading skills as significantly more important for academic success than students in Engineering, Computer and Physical and Mathematical Science. For Commerce students, higher-order reading skills were perceived to be more relevant for their academic work, specifically when compared to Engineering program students.

Figure 3: *Comparison of self-rated language competence and importance by academic program*
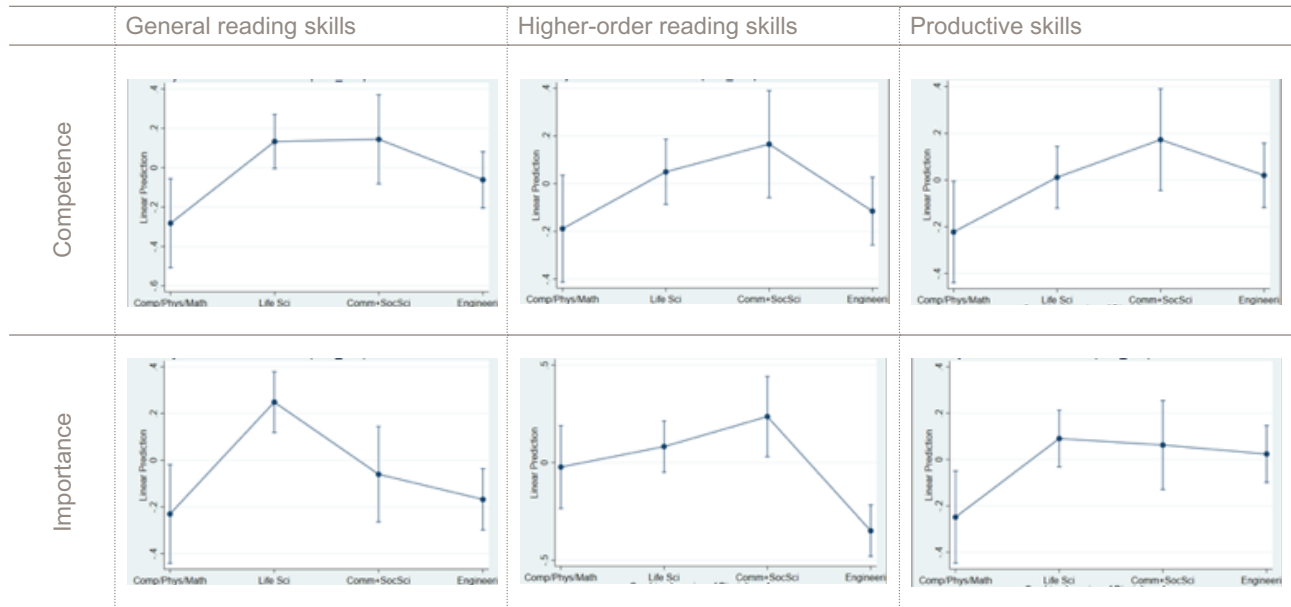
4.3.2   Predictive validity of IELTS test scores

The predictive validity of IELTS test scores was evaluated by comparing how categories of performance on IELTS reading band scores and IELTS average band scores differ on three separate criterion variables: 1) academic outcomes (i.e., CGPA); 2) self-reported language competence; and 3) a degree of preparedness for university language demands. The analyses in this section included only a sub-group of the survey respondents – those who took IELTS to meet the English language requirement as part of the admission to the university – resulting in the smaller sample size of 170.

Students' post-secondary academic performance, as measured by self-reported CGPA was positively correlated with their IELTS reading band scores ($r = .202*$) and their overall band scores ($r = .183*$)[1]. Students' perceived preparedness for language demands after having met the admission requirements was statistically significantly correlated with their IELTS reading ($r = .34**$) and overall test scores ($r = .34**$). This perceived preparedness was not predictive of academic achievements as measured by CGPA. However, students who felt prepared for school work showed a higher level of disciplinary literacy competence ($r = .17*$). The IELTS reading and overall test scores were not statistically significantly associated with self-reported language competence and importance factors. Interestingly, self-reported language competence in productive and disciplinary literacy skills was statistically yet weakly significantly correlated with CGPA ($r = .10*$ for both).

1. Due to small subsample sizes, we do not report on the relationship between CGPA and IELTS across years of program

4.4   Q4: What are the characteristics of IELTS reading skill profiles?

CDM was utilised in order to generate multidimensional reading skill profiles based on IELTS reading scores. The attributes included in the first Q matrices consisted of: (1) explicit textual comprehension; (2) making inferences beyond the text; (3) summarising the text; and (4) processing vocabulary knowledge. Explicit textual comprehension was defined as having basic comprehension of information in the text. This involves processing information explicitly stated in one or two sentences at the local level (Jang, 2009). Inferential processing requires readers to reason beyond the text, using background knowledge (i.e., personal experience and world knowledge) or information from the text to generate hypotheses (Wang & Gierl, 2011), predict future events following what is presented in the text, or infer the author's purpose or intent. Often with the goal of determining causal relationships or outcomes, inferencing items require readers to comprehend implicit information that goes beyond the direct statements in the text. Summarising was defined as the attribute required to comprehend key ideas at a global level by connecting, integrating or summarising information across sentences or paragraphs. Recognising the organisational structure of a text and the ability to identify main ideas from supporting details are required for effective summarising. Processing vocabulary knowledge involved deducing the meaning of difficult (i.e., infrequent) words using phonological, orthographic, syntactic or textual knowledge, or prior background knowledge. Items were coded as requiring vocabulary if understanding the meaning of a specific word was necessary for comprehension, and the word was infrequent enough to prove difficult for examinees (Li, 2011). The overall process of identifying attributes used in the analysis is summarised in Table 5.

Table 5: *Process of attribute identification*

Initial attributes set (11)	Granularity adjustment (7)	Final attribute set (4)
Explicit comprehension at the local level	Explicit comprehension	Explicit comprehension
Explicit comprehension at the global level		
Summarising main ideas	Summarising main ideas	Summarising main ideas
Understanding text organisation		
Understanding text purpose	Understanding text purpose	
Inferencing at the local level	Inferential reasoning	Inferential reasoning
Inferencing at the global level		
Linking to background knowledge	Linking to background knowledge	
Technical, discipline-specific vocabulary	Processing vocabulary knowledge	Processing vocabulary knowledge
Non-technical, general vocabulary		
Graphic interpretation	Graphic interpretation	N/A

The attribute distributions for Form A and Form B are presented in Table 6. As shown in the table, most items out of 40 required explicit textual comprehension (65.0% and 57.5% respectively). The next most prevalent attribute was inferencing (30.0% and 25.0%) followed by summarising (25.0% and 27.5%). Processing vocabulary knowledge was necessary for 17.5% and 12.5% of the items for Form A and Form B, respectively. Overall, over 50% of the total items for both forms measured explicit textual comprehension, which may pose a potential threat to the construct representation at the reading attribute level. A similar observation was made by Li (2011), who found 50% of the items in MELAB required extracting textually explicit information.

As noted by Jang (2009), this can be a pertinent issue in CDM when it is retro-fitted to existing tests developed with a specific Q matrix. Issues are attributable to a lack of balanced item distribution across attributes, as well as construct under-representation due to omission of core skills underlying the target construct.

Table 6: *Item distribution by attributes in the Q matrix*

Attributes	Form A		Form B	
	N	%	N	%
Explicit textual comprehension	26	65.0	23	57.5
Inferential reasoning	12	30.0	10	25.0
Summarising main ideas	10	25.0	11	27.5
Processing vocabulary knowledge	7	17.5	5	12.5

Once we selected the GDINA model for estimating skill mastery profiles from the IELTS response data, we subsequently examined diagnostic discrimination power at the item level before creating reading skill mastery profiles. Three values were utilised to evaluate the items' diagnostic discrimination power: *phat (m), phat (nm)*, and *pdiff (m-nm)*. *Phat (m)* represents the probability of correctly answering an item given mastery of the attributes required, while *Phat (nm)* refers to the probability of correctly answering an item without the mastery of all the relevant attributes. The value of *pdiff (m-nm)* is the difference between *phat (m)* and *phat (nm)* (Kim, 2015). These three statistics for Forms A and B are presented in Appendix J. The average *phat (m)* were .794 and .705 for Form A and B respectively. Thus, the average probability for getting an answer correct given the mastery of required attributes was moderately high, at 79.4% and 70.5% for both forms. The average *phat (nm)* scores, in contrast, were low, at .405 and .374 for Form A and Form B respectively, indicating that test-takers without mastery of required attributes had an average chance of 40.5% and 37.4% of providing the correct answer. On average, masters outperformed non-masters by an average of 38.8% on Form A and 33.1% on Form B. Overall, skill masters were well differentiated from non-skill masters.

In addition, the model estimated item statistics were not significantly different from observed item statistics, which was corroborated by adequate RMSE values. Some items failed to differentiate skill masters from non-masters. Specifically, items 18 (.041) and 30 (.051) on Form A and items 5 (-.006), 8 (-.006), and 30 (.040) on Form B which had poor *pdiff (m-nm)* scores. These findings are corroborated by the graphical representations in Figures 4 and 5. These items may require additional analysis of both attributes specified and item characteristics.

Figure 4: *Item p-values between masters and non-masters for Form A*
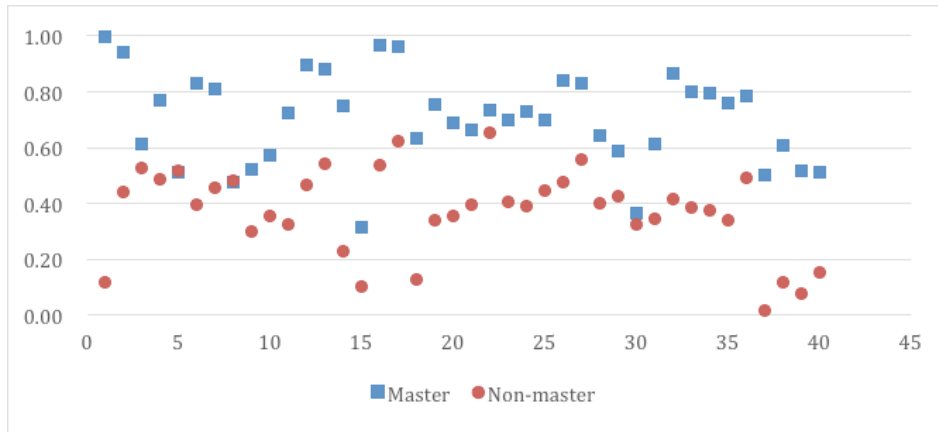
From the final application of the G-DINA model to the response data, skill mastery probabilities were estimated for all individual test-takers and for all skills. As shown in Table 7, the summarising attribute showed the highest proportion of mastery, with 56.5% and 50.3% of test-takers having likely mastered the given skill for Form A and Form B respectively. The vocabulary skill showed 49% and 34% of test-takers having likely mastered this skill for Form A and Form B, while the explicit comprehension skill showed 35% and 40% mastery respectively. Overall, the inferential reasoning skill showed the lowest proportion of mastery among the four skills, with 28% and 31% respectively. Of note, the vocabulary skill showed variations in its proportions of mastery between forms. Along with some poor performing items reported above, this variation may require additional attention to Q matrix specifications and comparability of forms in terms of difficulty and discrimination. Note that in our final step where skill proficiency descriptors were developed, we removed the vocabulary skill from the Q matrix and reapplied the G-DINA model to update the skill profiles.

Table 7: *Proportions of mastery across attributes*

Reading attribute	Proportions of mastery	
	Form A	Form B
Explicit comprehension	.34	.40
Inferential reasoning	.28	.31
Summarising main ideas	.57	.50
Vocabulary	.49	.35

In developing individual test-takers' skill mastery profiles, we applied a cut-off point of .5 (e.g., Kim, 2015; Lee & Sawaki, 2009b; Li, 2011; Ravand, 2016; Ravand & Robitzsch, 2015; Yi, 2017). As shown in Table 8 and Figure 6, out of 16 total possible mastery classes, 27% of test-takers who took Form A did not master any of the four skills. The next most frequent profiles included 0011 (masters of summarising and vocabulary) and 1111 (masters of all skills), representing 22% and 15% respectively.

Form B showed a slightly different distribution of mastery classes from Form A. About 41% of Form B test-takers did not master any skill, whereas 23% mastered all. About 10% showed mastery of summarising. Form B's class distribution is common in CDM applications, as two flat classes (0000 and 1111) tend to be the most frequent ones. However, interestingly, Form A deviates from this observation. Differences in mastery patterns between forms require additional attention.

Table 8: *Frequency of skill mastery classes/patterns*

Attribute pattern (Exp-Inf-Sum-Voc)	Form A		Form B	
	Frequency	**Percentage**	**Frequency**	**Percentage**
0000	1,396.61	26.7%	3,385.57	41.0%
1000	151.20	2.9%	148.04	1.8%
0100	0.18	0.0%	7.05	0.1%
1100	129.23	2.5%	87.21	1.1%
0010	304.76	5.8%	793.17	9.6%
1010	188.22	3.6%	430.12	5.2%
0110	0.12	0.0%	3.33	0.0%
1110	509.68	9.8%	511.18	6.2%
0001	567.72	10.9%	368.93	4.5%
1001	0.56	0.0%	15.79	0.2%
0101	3.20	0.1%	6.59	0.1%
1101	21.45	0.4%	79.82	1.0%
0011	1,142.23	21.9%	345.91	4.2%
1011	6.48	0.1%	183.56	2.2%
0111	20.38	0.4%	12.47	0.2%
1111	779.99	14.9%	1,872.25	22.7%

Figure 6: *Comparison of attribute pattern distribution between Forms A and B*
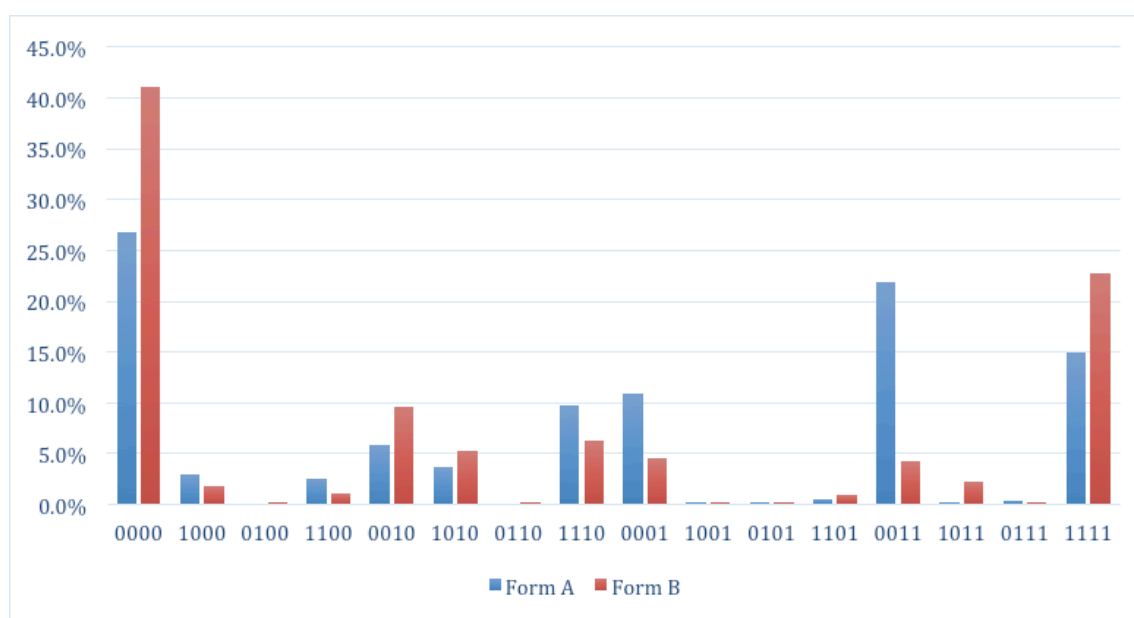
4.5 Q5: What proficiency descriptors characterise IELTS band score levels based on blended CDM profiling with scale anchoring?

4.5.1 Step 1: Determining the proficiency levels

Although the IELTS test is reported on nine band score levels, not all levels are sufficiently distinguishable in terms of distinct skills and knowledge. To determine the number of levels for proficiency descriptions, we examined the relationship between the IELTS reading band scores and CDM-based skill mastery profiles. We paid close attention to 6.5, as it is the most common cut-off score level used at the University of Toronto. Figure 7 shows the overall score distribution based on Form A response data. Given discrepancies between the two reading test forms, we used only Form A for proficiency descriptor development.

Figure 7: *Band score distribution for Form A*
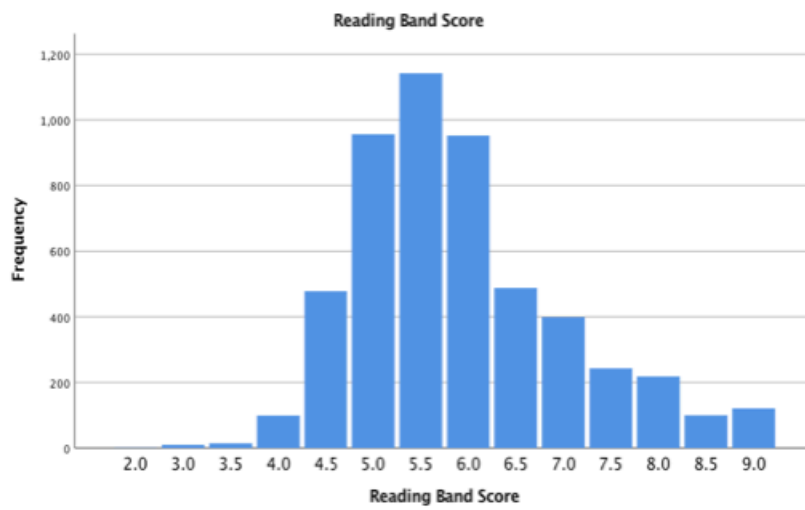
As previously mentioned from the description of the earlier CDM results, the vocabulary attribute did not show a monotonic relationship with the IELTS band scores, partly due to a relatively small number of items associated with it and lack of diagnostic discrimination power among associated items. Thus, in this phase of proficiency descriptor development, we revised the Q matrix for Form A by excluding the vocabulary attribute, chose the final model (G-DINA) through the comparison of model fits among five CDMs, and updated skill mastery profiles based on the three aforementioned attributes. Table 9 compares proportions of mastery by reading attribute without the vocabulary attribute to those resulted with vocabulary from our earlier analysis. Frequency and proportion of each skill mastery class were also presented in Table 10.

Table 9: *Comparisons of proportions of mastery: three- vs. four-attribute models*

Reading attribute	Proportions of mastery	
	Without vocabulary	With vocabulary
Explicit comprehension	.34	.34
Inferential reasoning	.27	.28
Summarising main ideas	.56	.57
Vocabulary	N/A	.49

Attribute pattern (Exp-Inf-Sum)	Frequency	Percentage
000	2023.62	38.8%
100	143.85	2.8%
010	-1.41	0.0%
110	109.78	2.1%
001	1412.37	27.1%
101	212.78	4.1%
011	30.57	0.6%
111	1290.43	24.7%

Figures 8–10 show the distribution of model-estimated skill mastery levels for three attributes across the IELTS band score levels. In Figures 8–10, the red horizontal line marks the level of .5 as an indicator of skill mastery status.

Figure 8: *Skill mastery probability estimates for basic comprehension across the IELTS band scores*
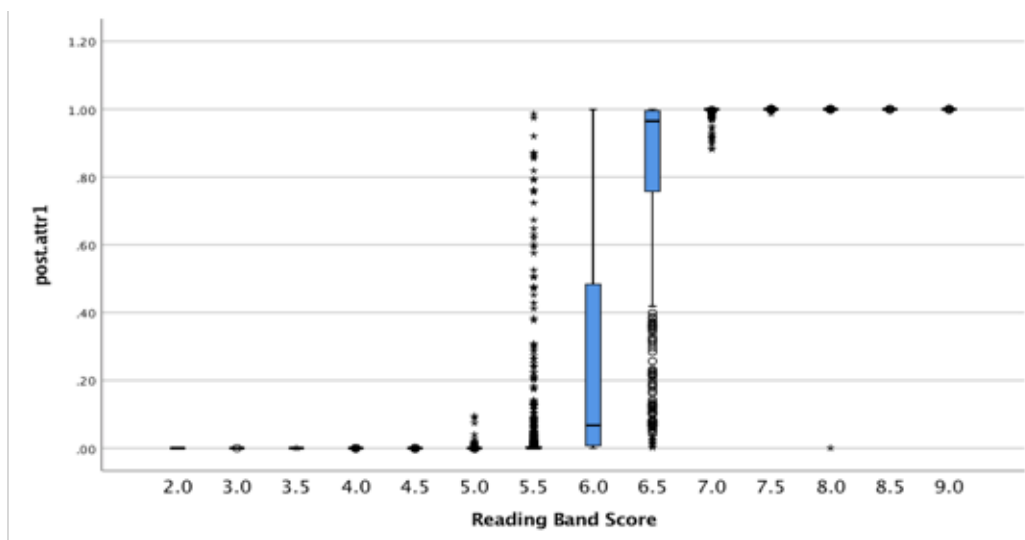
Figure 9: *Skill mastery probability estimates for summarising main idea across IELTS band scores*
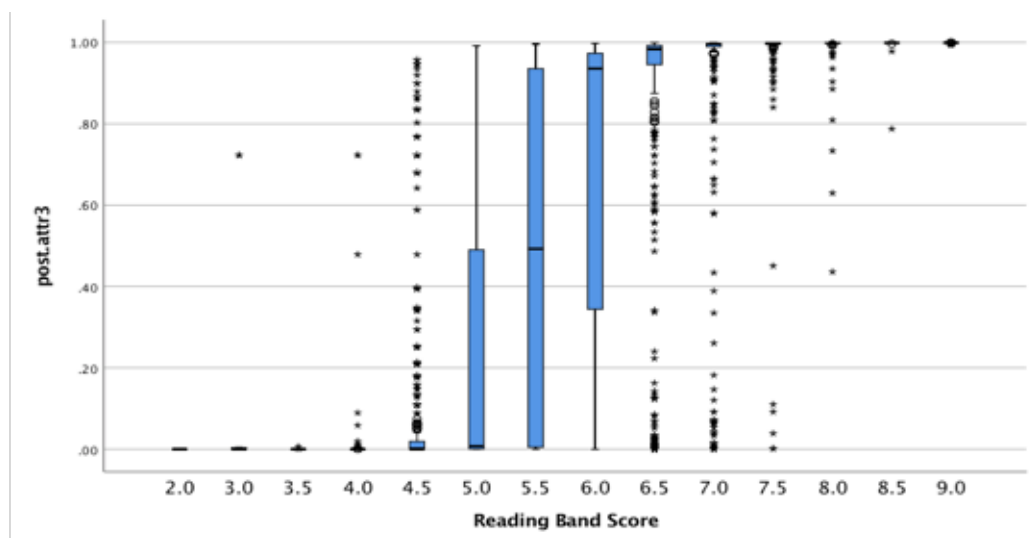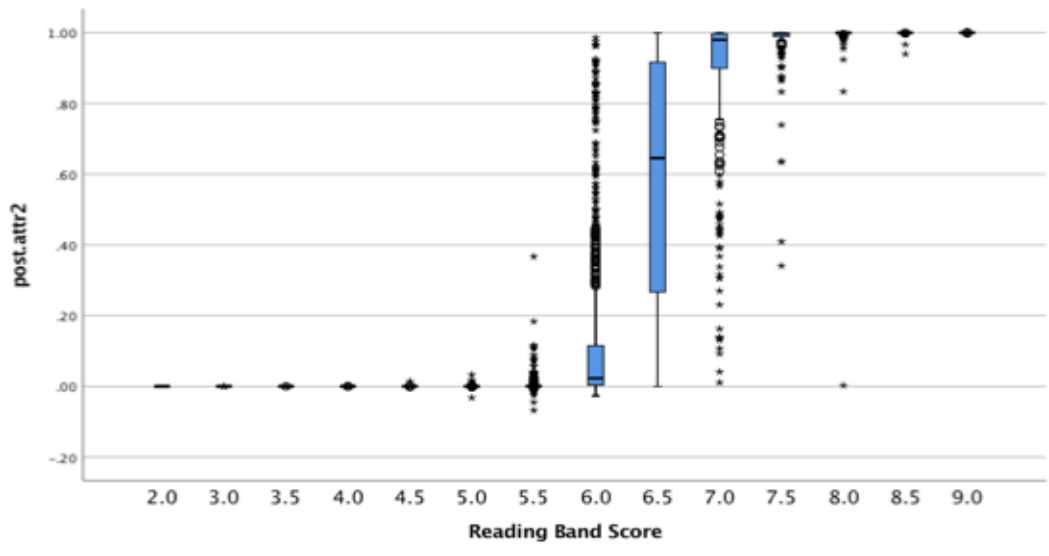
As shown in Table 11, students whose IELTS band score is 6.5, which is the cut-off score set by most programs at the University, clearly demonstrate the mastery of two comprehension skills (explicit comprehension and summary of main ideas). However, their mastery of inferential reasoning was slightly over .5, indicating a lack of sufficient mastery even at the cut-off score level.

Table 11: *Average skill mastery estimates across the IELTS band scores*

Attribute	4.5 M	5.0 M	5.5 M	6.0 M	6.5 M	7.0 M	7.5 M
Basic comprehension	0	.00	.03	.26	**.81**	1.00	1.00
Summarising main ideas	.10	.25	.48	.68	**.85**	.92	.97
Inferential reasoning	0	0	.00	.12	**.58**	.89	.98

4.5.2 Step 2: Identify anchor items

Conditional *p*-values were calculated for each of the IELTS band score levels. We applied two main criteria in order to identify anchor items. As noted, we identified items whose conditional *p*-values are .65 or .7 after rounding and are distinguished from the adjacent lower level by .2. Table 12 shows anchor items in shaded cells along with overall *p*-values and item difficulty and discrimination parameter estimates based on the application of 2PL IRT.

Table 12: *Conditional p-values across the IELTS band scores*

Item	IELTS band score levels							Overall p-value	IRT b parameter	IRT a parameter
	4.5	5	5.5	6	6.5	7	7.5			
1	.34	.49	**.68**	.82	.9	.94	.98	.71	-.68	1.52
2	.29	.43	.62	**.72**	.85	.89	.94	.65	-.66	1.19
3	.51	**.71**	.82	.89	.92	.94	.97	.81	-1.63	1.11
4	.58	**.75**	.88	.94	.97	.98	.97	.86	-1.77	1.34
5	.03	.06	.09	.1	.14	.13	.18	.1	5.02	.45
6	.12	.18	.31	.45	.56	**.7**	.81	.42	.56	1.58
7	.34	.53	.62	**.75**	.83	.88	.95	.68	-.49	1.12
8	.54	**.64**	.65	.73	.76	.82	.89	.71	.97	1.9
9	.08	.18	.32	.42	.49	.58	.6	.37	.71	.87
10	.34	.51	.62	**.77**	.83	.89	.93	.68	-.33	1.28
11	.18	.32	.45	.54	.64	**.72**	.78	.5	.01	.81
12	.01	.08	.18	.29	.5	.59	**.72**	.31	.74	1.56
13	.04	.09	.19	.3	.42	.56	.63	.29	.94	1.22
14	.38	.41	.5	.62	**.69**	.76	.91	.58	.71	1.83
15	.46	.53	**.64**	.75	.81	.9	.93	.69	.21	1.57
16	**.67**	.76	.83	.89	.94	.96	.97	.85	-.14	1.42
17	.51	.63	**.7**	.76	.8	.85	.92	.72	-.36	.77
18	**.81**	.85	.84	.89	.89	.9	.9	.87	-11.12	.17
19	.27	.24	.25	.31	.38	.52	**.64**	.36	1.37	2.54
20	.54	**.64**	.73	.76	.82	.86	.91	.74	.07	.84
21	.49	**.64**	.76	.83	.85	.88	.93	.76	-1.67	.77
22	.3	.44	.5	.59	**.67**	.74	.77	.55	.33	.8
23	.31	.32	.37	.47	.57	**.64**	.73	.47	1.1	2.03
24	.35	**.64**	.75	.84	.89	.92	.97	.75	-1.25	1.09
25	.34	.56	**.72**	.83	.83	.85	.84	.71	-1.14	.9
26	.16	.28	.41	.49	.63	**.69**	.77	.47	.32	1.01
27	.35	.51	**.65**	.78	.84	.88	.88	.68	-.74	.96
28	.26	.34	.44	.54	.59	.61	**.72**	.49	.65	.78
29	.33	.37	.45	.51	.6	.63	**.77**	.5	1.17	1.25
30	.24	.31	.35	.38	.35	.37	.34	.35	2.67	2.09
31	.24	.26	.32	.39	.46	.6	**.79**	.41	1.09	2.37
32	.26	.38	.5	**.71**	.85	.95	.99	.62	.33	5.29
33	.24	.33	.47	.59	**.75**	.88	.93	.56	.56	5.33
34	.25	.32	.4	.55	**.69**	.84	.9	.52	.66	16.15
35	.23	.28	.4	.55	**.68**	.86	.92	.52	.65	15.52
36	.27	.43	.55	**.68**	.8	.84	.87	.61	-.15	1.14
37	0	.01	.04	.13	.3	.57	**.68**	.22	.93	2.85
38	.05	.09	.21	.34	.52	**.73**	.82	.35	.59	2.25
39	.02	.05	.11	.24	.45	.54	**.64**	.27	.89	1.77
40	.06	.11	.19	.29	.49	.56	**.7**	.32	.81	1.67

4.5.3   Step 3: Diagnostic discrimination indices

We examined diagnostic discrimination index values at the attribute and item level estimated from CDM. In Table 13, the higher the discrimination index value an item-attribute pair provides, the more informative it is.

Table 13: *Diagnostic discrimination index at the item and attribute level*

Item	Basic comprehension	Summarising main ideas	Inferential reasoning	Global item discrimination index
1		1.82		1.21
2		1.25		0.83
3		0.67		0.45
4		0.27		0.18
5	0.03			0.02
6	0.44			0.29
7			0.33	0.22
8	0.06		0.05	0.05
9	0.21			0.14
10	0.21	0.04		0.17
11	0.21			0.09
12	0.44	0.06		0.33
13	0.32	0.03		0.23
14			0.30	0.20
15	0.24			0.16
16	0.14			0.09
17			0.08	0.06
18	0.01			0.01
19		0.05	0.17	0.12
20	0.08			0.05
21			0.14	0.10
22	0.15			0.10
23	0.21			0.14
24	0.32		1.31	0.63
25	0.15			0.10
26	0.30			0.20
27		0.16		0.11
28		0.01	0.13	0.09
29			0.16	0.11
30			0.01	0.00
31			0.34	0.23
32	0.81			0.54
33	0.39	0.04		0.28
34			0.63	0.42
35	0.61			0.40
36	0.34			0.22
37		0.02	1.13	0.77
38	0.87			0.58
39	1.10		0.08	0.70
40	0.57			0.38
Global attribute discrimination index	8.09	4.43	4.88	

4.5.4   Step 4: Creating proficiency descriptors from the anchor item pool

In selecting key band score levels for creating proficiency descriptors, we considered the cut-off score set at 6.5 for admission to the University and the fact that few anchor items were below 5 and above 7.5. The final band score levels included 5.5 and above. Table 14 shows a set of proficiency descriptors for each of these selected IELTS band score levels.

Table 14: *Proficiency descriptors for IELTS reading band score levels 5.5 and above*

Band score	At each IELTS reading band score level, students in general can:
5.5+	• locate a keyword or a topic sentence by scanning and skimming a text • comprehend the literal meaning of short phrases or simple sentences • figure out the meaning of high-frequency vocabulary
6.0+	• understand the main idea from a paragraph • distinguish the main idea from supporting details • figure out the meaning of moderately difficult vocabulary
6.5+	• comprehend implicit meaning in a text • paraphrase main ideas • summarise the main idea from a long, grammatically complex text • figure out the meaning of low-frequency vocabulary • **start to infer implicit meaning from text**
7.0+	• synthesise the main idea with supporting details from the text • make inferences about implicit information from the text • understand logical connections between ideas across sentences
7.5+	• infer meaning in the text that is specific to a certain culture • figure out colloquial expressions in the text • comprehend the text with abstract vocabulary and grammatically complex sentence structures (e.g., if-then, although-)

As noted above, we paid close attention to the observation that students at 6.5 demonstrate insufficient mastery of inferential reasoning. This is noted in bolded font above.

4.5.5   Recommendations for preparing students for discipline-specific academic language and literacy demands

One of the key findings from the project was the significant amount of discrepancy between students' beliefs about their preparedness for academic performance based on their IELTS test scores and their actual preparedness. As such, we developed a set of recommendations for test-takers and test score users (Table 15). For example, university admissions offices and international student affairs units can use the recommendations to provide more comprehensive information to incoming students regarding expectations before they start their programs.

Table 15: *Recommendations for incoming students*

Recommendations for incoming students:
Most undergraduate students report that their academic work requires much more advanced skills and knowledge than they expected. More importantly, different fields of study (e.g., engineering, economics) have different skill requirements. As you may be aware, these specific skills may not be assessed by language tests such as IELTS or TOEFL. Here are some key demands and skills identified by U of T international students. Prepare yourself for academic work by developing competencies in these skills.

1	Be prepared that the length and number of texts/materials that students are assigned to read in courses are much longer than typical passages used in IELTS or TOEFL. Most courses require sustained stamina and attention while reading long texts or multiple sources.
2	Basic comprehension of main ideas is insufficient for successful academic work. Most coursework requires students to critically appraise or compare different viewpoints presented in a long text or across multiple texts.
3	Vocabulary becomes technical and abstract. Some vocabulary may have different meanings depending on the field of study. In fact, technical and abstract vocabulary is one of the most significant challenges international students face when reading academic materials or discussing them in speech or writing.
4	Challenges with reading academic materials are not just due to complex grammar or vocabulary. Texts convey meanings and views that are specific to certain cultures. When international students come from a different cultural background, they often struggle due to a lack of cultural or background knowledge.
5	Often reading passages used when studying English are rather general. Academic reading materials are quite different across different fields of study. Some may require interpretations of statistical tables, figures, mathematical formulas, or computational codes. They vary widely depending on the intended audience and publication type (e.g. text book vs. business journal vs. technical manual).
6	International students often have difficulty understanding collocations (e.g., land a deal, make progress) and idiomatic expressions (e.g., under the weather, a piece of cake). Although these are not frequently used in academic reading materials, they are often used in oral conversations and in lectures.
7	As you know, language isn't used in isolation. Most academic work requires reading multiple reading materials, listening to lectures, participating in discussions, and writing short or long reports or essays. Frequently, course assignments require integration of multiple language tasks, more than just reading or listening alone.

4.5 Q6: How do test score users respond to can-do proficiency descriptors across IELTS band scores and to recommendations regarding university disciplinary language and literacy demands?

Between the IELTS report that students received and university admission acceptance communications, students felt that they were being informed of their preparedness for university-level English demands, yet upon arrival, felt vastly under-prepared. From communications received (e.g. band score level, acceptance letter), students felt that they had no real understanding of the types of reading demands required within their programs until they arrived.

Students described being tasked with course readings that differed significantly from those that they had encountered during admissions assessments in a variety of ways including: level of difficulty, volume, length and stamina required, technical and disciplinary-specific vocabulary, and text structure. They felt that they would have benefited from strategies on, for example, how to navigate more complex, disciplinary-specific texts (e.g. business journals, scientific articles), including how to determine the main point and interpret results.

Overall, although students would have appreciated additional and more detailed communication regarding their language mastery and expectations of university-level demands, they also would have liked more targeted advice regarding how they could develop these requisite skills in the form of actionable steps, samples of texts and real-world examples to contextualise what these skills might look like in practice. Specifically, students provided feedback on the proficiency descriptors and recommendations as follows[2]:

- some terms are too general

- simplify some wording to be less academic

- clarity required around what cultural references/culturally-specific language means

- clarity required for some statements, perhaps provide examples/context, e.g. "evaluating consequences to infer meaning"

- make skills actionable, provide context or example of specific tasks to connect descriptors to real-world requirements

- provide resources that can be used to improve skills

- overall, skill descriptors seem accurate based on student's memory.

Students highlighted the following areas where they felt that further preparation was lacking and would have been highly beneficial:

- importance of knowing how to skim/scan, speed read, manage high volume of reading

- summarising and extracting argument from lengthy readings

- understanding text organisation, extracting logical structure and understanding relationship between paragraphs and where to find information in an article

- understanding vocabulary that is complex/technical/colloquial/includes metaphors.

- importance of practising these skills

- writing/speaking and productive language skills, listening in lectures

- professional communication (social pragmatics): asking professional questions, communicating via email.

Faculty/admissions focus group participants confirmed findings from the Phase 1 focus groups and survey based on their own observations. In particular, they expressed concerns regarding students' preparedness to tackle course content upon admission, as well as the increase in silos of same-language peers as students became increasingly ghetto-ized. Participants discussed the inherent tension between continuing to draw students to the university by maintaining comparable admissions requirements with other Canadian institutions (although some noted UT standards are already higher than others), yet the importance of not setting students up for failure where significant discrepancies exist between the language ability level achieved and program language requirements. They felt that significantly more effort is needed in providing more comprehensive communication regarding what language demands students should anticipate at the university, and how these demands are not fixed, but change over the years.

The group engaged in extensive discussion regarding recommendations that included four main areas: 1) sharing findings with relevant collaborators at the university (e.g. Dean's office, committees); 2) improved pre-arrival communication/education of incoming students (through the website, online videos, module development, expanded orientation); 3) improved ongoing university support (increased awareness of support, targeted support for upper-year students, development of ongoing online support and modules to track progress); and 4) professional development (expanding incoming instructor orientation to include education regarding international students' experiences and strategies for providing support).

2. The Skill Proficiency Descriptors across the IELTS Reading Band Score Levels (Table 12) and set of recommendations (Table 13) reflect feedback from the student focus group, discussed in this section.

5.   Discussion

The current study sought to investigate validity arguments related to IELTS reading score interpretations and use, exploring issues of consequential validity, the intersection of contextual validity and cognitive validity, as well as scoring validity (Weir & O'Sullivan, 2011). Specifically, the study conducted a local investigation of IELTS test score interpretations and use in a higher education institution in Canada. Although standardised tests, such as TOEFL and IELTS, have been used for screening and admissions purposes at the University of Toronto, there exists no local investigation of test score users' interpretations and use of such external test scores. Chalhoub-Deville and Turner (2000) called for test users to be responsible for appropriate test use and interpretation by carrying out local investigations to ensure that their admission requirements are in alignment with the academic language demands required for success in specific programs. Test developers are responsible for developing tests that meet professional standards, providing high-quality information about students' language abilities, and providing user guides to support decision-making processes in local contexts (AERA, NCME & APA, 1999).

The present study examined ways in which local IELTS test score users at the University of Toronto interpret test scores and use them for making critical decisions about admission into undergraduate programs. Through four phases, we sought to explore and better understand international students' perceptions regarding their language proficiency and preparedness for academic demands, and to develop reading skill mastery profiles in order to examine the possibility of enhancing test score users' interpretations of scores through the use of descriptors developed through scale anchoring.

The study results show that many international visa students felt inadequately prepared for meeting academic language and literacy demands at entry, even after they had met the language proficiency requirement. These concerns are well aligned with previous research on international students' academic performance in the context of English for academic purposes (Bruce & Hamp-Lyons, 2015; Hamp-Lyons, 2011; Sawir, Marginson, Forbes-Mewett, Nyland & Ramia, 2012) and discipline-specific literacy practice (Shanahan & Shanahan, 2012). Our study findings also concur with previous research on IELTS reading. Weir, Hawkey, Green, Unaldi and Devi (2009) reported that "the reading skills and strategies tapped by the IELTS Reading Module test may need further investigation and possible modification to more closely represent the academic reading constructs of university students through texts and tasks that test more extensively students' expeditious reading skills" (p. 133).

Furthermore, our results raise concerns about the lack of items at the text level. Our CDM analyses showed that a large number of items assess comprehension at the sentence level. Similar to Weir et al. (2009), items that elicit summary completion are associated with word matching and sentence-level comprehension strategies rather than text-level expeditious reading. A lack of sufficient items that measure text-level expeditious reading is a cause for concern given students' struggles with real-life academic reading demands involving a large amount of reading under time constraints.

Although course materials we collected represent a small sample of the diverse array of courses available within these programs, our analysis of course materials highlights the unique and specialised reading-related skills required to navigate different text types. Some common challenges that students face across the programs reviewed (Commerce, Economics and Engineering) were related to the amount of reading, abstract and technical academic vocabulary, and cultural background knowledge.

The majority of students found the volume of reading 'overwhelming', they found the discipline-specific vocabulary challenging, and they found that implicit cultural assumptions and norms interfered with their text comprehension. All of these contribute to deepening discrepancies between students' experiences with IELTS texts and undergraduate texts. Similarly, Weir et al. (2009) found that IELTS texts had much lower estimated grade levels than undergraduate texts, which were further complicated by cultural specificity and discipline-specific vocabulary. Our course material analyses concur that the IELTS text does not adequately reflect the level of challenge with undergraduate texts that students encounter upon entry to programs.

Whether or not high-stake tests should reflect real-life texts and reading conditions is a critical question for consideration because full authenticity is unrealistic for testing conditions constrained by limited time and a need to avoid biases associated with cultural and field specificity. However, valid test score interpretations are subject to the extent to which test features and cognitive strategies elicited by test items reflect reading activities and processes that students encounter during school-work (Green, Unaldi & Weir, 2010).

Taken together, these findings highlight the unique challenges that international students face coming to the university. Upon entry to programs, they realise that their mastery of basic language skills may be insufficient to navigate increasingly discipline-specific texts ranging from more expository textbooks to more analytical journals and case studies. Such findings highlight the importance of program-specific support across the years, in order to tailor language-related support to the varied requirements within programs, but also, to scaffold students as their language demands change and become more discipline-specific.

In addition, international students' struggles with social speaking skills and academic language demands appear to result in decreased confidence regarding their language ability as they immerse themselves in school-work. For example, our survey data shows that international visa students rate their language and literacy skills much lower than domestic peers, whereas their perceptions of the importance of such skills, especially, higher-order inferential reasoning skills, are much higher than their domestic peers. Interestingly, several upper-year students explained that their desire to improve English decreased over time because they felt that they could still get high-enough marks by taking courses that do not require active participation in discussions. This loss of confidence and avoidance mentality is counterproductive and detrimental for international students, as the majority wish to pursue careers in Canada after graduation.

Further analyses of IELTS test-takers show that students who performed better on IELTS overall, and reading specifically, felt significantly more prepared for academic language demands than students who just met the cut-off score. Overall, IELTS test scores, as well as reading band scores, were positively correlated with CGPA, but the strength of the relationship was weak for both overall and reading band scores. However, differences in CGPA across the IELTS band score levels were not statistically significant. Note that we grouped the IELTS test-taker sample into three groups based on the cut-off score (≤6.5, 7.0-7.5, >7.5); therefore, the results indicate that the cut-off score may not be sensitive enough to predict different academic achievement levels.

The study findings are somewhat consistent with previous research that reports a weak or no predictive relationship between IELTS and academic performance (Dooey & Oliver, 2002). Similar to Cotton and Conrow (1998) and Dooey and Oliver (2002), our study showed that the IELTS reading test had a weak positive correlation with CGPA, suggesting a predictive relationship with academic achievement. Students' perceived preparedness for school work based on their language test scores also show a positive correlation with CGPA. Students with higher competence levels in productive language and disciplinary literacy skills tend to achieve higher CGPA.

This finding needs further analysis because the effect may be further moderated by program type and year of program, results which could facilitate a discussion about whether or not different programs use different cut-off scores. For instance, highly competitive programs would likely have students with a higher IELTS score, but lower CGPA due to the intensity of the program in question. In addition, the effect of IELTS test scores may be more significant for first-year students. Due to a small sub-sample size, the effect of these moderating variables could not be tested. Another methodological limitation is that both IELTS performance and CGPA were self-reported, which may have led to a social desirability bias, in which students falsely inflated their test scores or CGPA. Likewise, it is important to highlight that many dependent variables measured self-reported competence and preparedness. While these variables were efficacious in measuring subjective student experiences, concrete conclusions regarding the predictive validity of IELTS test scores should use more objective measures, such as official IELTS performance and CGPA scores.

Building on input from various test score users regarding their score interpretations and use for admission purposes, we sought ways to develop reading skill profiles through the identification of crucial skill descriptors associated with different IELTS reading band scores. We further examined how test score users perceive the skill descriptors along with guidelines for incoming students.

Iterative analyses involving the application of G-DINA to response data from two IELTS test forms showed a discrepancy in the distributions of skill mastery classes and diagnostic quality of the items between the two forms. Of the Form A test-takers, 27% did not master any skills, compared with 41% of Form B test-takers who did not master any skills. Approximately 15% of Form A test-takers and 23% of Form B test-takers mastered all skills. Form B's class distribution is common in CDM application, providing two flat classes (0000 and 1111) (DiBello, Stout & Roussos, 2007; Jang, 2009). However, interestingly, Form A deviates from this observation. Differences in mastery patterns between forms require additional attention. Note that the vocabulary skill showed variations in its proportions of mastery between forms, and some poor performing items were associated with the vocabulary skill. As a result, the final Q matrix used for developing skill descriptors excluded vocabulary skill. We reapplied the G-DINA model to Form A response data in order to update the skill profiles. Form equivalence is critical for ensuring fair and valid score interpretations for test takers. Careful attention needs to be paid to ensure that test forms are comparable in terms of attribute specifications and diagnostic quality of items.

Scale anchoring was used to determine proficiency levels, identify anchor items, examine diagnostic discrimination index values, and create proficiency-level skill descriptors. Instead of developing descriptors for each of the band score levels, we focused our analysis on scores adjacent to the local cut-off score. The study results show that students who met the institutional cut-off score level demonstrate mastery of two comprehension skills: basic textual comprehension and summary of main ideas. However, their inferential reasoning ability requiring global reading beyond the sentence level was only slightly over .5, indicating a lack of sufficient mastery even at the cut-off score level. Students who had the IELTS band score of 6.5 or higher can comprehend implicit meaning in a text, paraphrase main ideas, summarise from grammatically complex text and figure out the meaning of low-frequency vocabulary, but may not demonstrate sufficient evidence for inferential reasoning.

When we sought feedback from test score users about the IELTS reading proficiency-level skill descriptors, students welcomed additional information that could help them better understand IELTS band scores. Faculty and admissions staff focused their discussion on how to utilise the skill descriptors to support students.

Some of their suggestions included:

1. sharing findings with relevant collaborators at the university (e.g., Dean's office, committees)

2. improving pre-arrival communication/education of incoming students (through the website, online videos, module development, expanded orientation)

3. improving university support (increased awareness of support, targeted support for upper-year students, development of ongoing online support and modules to track progress)

4. enhancing professional development (expand incoming instructor orientation to include education regarding international students' experiences and strategies for providing support).

In sum, test score users have limited information about what test scores mean and welcome additional descriptive information about what students can do at different proficiency levels. International visa students enter school without sufficient knowledge about academic language and literacy demands. They appear to believe that once they meet the admission language requirement, they should be prepared for success at academic school-work.

The study findings highlight the importance of local investigations into test score interpretations and use in specific contexts. Together, results emphasise the need to inform international students of academic language demands beyond general language abilities, and to inform faculty and administration staff of incoming students' abilities, as well as additional support required in order to succeed at the university level, paying special consideration to discipline-specific language demands.

6. Conclusions

In brief, we recommend that the IELTS partners should consider the following aspects of the test.

Provide more information about what the IELTS test scores mean to users

IELTS is one of the most frequently used tests among international visa students seeking admission at the University of Toronto. While the test is widely known to university staff and faculty, their knowledge of the test is limited to the range of band scores associated with it. It has gained a "symbolic value" (Smith & Haslett, 2007, p. 36), yet its function as an indicator of language proficiency needs to be further validated and improved. IELTS needs to provide decision-makers, program staff, faculty and students with more information about what test scores mean and what the test does and does not measure.

Consider ways to facilitate students' test score interpretations and use, to prepare them for academic study

Differences between the IELTS reading texts and the texts that students encounter in their first year of undergraduate study contribute significantly to students' interpretations of test scores, sense of preparedness, and attitudes towards the test. While it may not be possible to make the test conditions and materials fully authentic (Green et al., 2010), IELTS can support students' desire to have more information about what their scores mean and what they should expect and be prepared for during undergraduate study. Test developers need to consider ways to facilitate students' test score interpretations and use, to further prepare them for academic study.

Investigate if the reading test has sufficient items that measure higher-order thinking skills

The three reading attributes used to develop reader profiles and proficiency-level skill descriptors include basic comprehension, summarising main ideas, and inferential reasoning. Our analyses show that relatively fewer items elicit text-level, global reading beyond local literal configuration, a finding which is similar to previous research (Moore, Morton & Price, 2012; Weir et al., 2009). Our finding that students at the band score 6.5 (the most common local cut-off score across programs at the University of Toronto) still lack inferential reasoning was a cause for concern among faculty and staff. They expressed that the result would significantly impact student experiences. This result needs to be further investigated to determine if it is due to a lack of sufficient items that measure higher-order thinking skills, such as inferential reasoning at the text level.

Investigate IELTS' predictive relationship with academic success across different programs and year of program

Higher IELTS scores (overall and reading specifically) were associated with students feeling significantly more prepared for academic language demands. Furthermore, the IELTS reading test scores were positively correlated with academic success as measured based on CGPA. IELTS' predictive validity needs to be further investigated by examining its predictive relationship with academic success across different programs and year of program.

Facilitate test score users' score interpretations and use by providing descriptive information about what the test measures and what it doesn't

Test score users welcomed the IELTS reading proficiency skill descriptors. Students found them useful for understanding what the test band scores mean and upon what students need to further improve. In particular, they appreciated a list of recommendations that highlight the characteristics of academic language and literacy demands. Faculty and staff found the skill descriptors informative and were interested in various ways to support students for academic success. Test developers need to facilitate test score users' score interpretations and use by providing descriptive information about what the test measures and what it *doesn't*.

References

Bachman, L. F. (2005). Building and supporting a case for test use. *Language Assessment Quarterly: An International Journal*, 2(1), 1–34.

Banerjee, J. V. (2003). *Interpreting and using proficiency test scores.* Unpublished doctoral thesis, Department of Linguistics and Modern English Language, Lancaster University, Lancaster, United Kingdom.

Bax, S. (2015). Using eye-tracking to research the cognitive processes of multinational readers during an IELTS reading test. *IELTS Research Reports Online Series*, 2015/2.

Beaton, A., & Allen, N. (1992). Interpreting scales through scale anchoring. *Journal of Educational Statistics*, 17(2), 191–204.

Bruce, E., & Hamp-Lyons, L. (2015). Opposing tensions of local and international standards for EAP writing programmes: Who are we assessing for? *Journal of English for Academic Purposes*, 18, 64–77.

Chalhoub-Deville, M., & Turner, C. E. (2000). What to look for in ESL admission tests: Cambridge certificate exams, IELTS, and TOEFL. *System*, 28(4), 523–539.

Chapelle, C. A., Enright, M. K., & Jamieson, J. M. (Eds.) (2008). Test score interpretation and use (p. 1–25). *Building a validity argument for the Test of English as a Foreign Language*. New York: Routledge.

Charmaz, K. & Bryant, A. (2008). Grounded theory. In L. M. Given (Ed.), *The SAGE encyclopedia of qualitative research methods* (pp. 375–376). Thousand Oaks, CA: SAGE Publications Ltd. doi: 10.4135/9781412963909.n189

Coleman, G., & O'Connor, R. (2007). Using grounded theory to understand software process improvement: A study of Irish software product companies. *Information and Software Technology*, 49(6), 654–667.

Coleman, D. S., Starfield, S., Hagan, A. (2003). The attitudes of IELTS stakeholders: Student and staff perceptions of IELTS in Australia, UK and Chinese tertiary institutions. *IELTS Research Reports* Vol. 5. Canberra: IDP Education Australia.

Coley, M. (1999). The English language entry requirements of Australian universities for students of non-English speaking background. *Higher Education Research & Development*, 18(1), 7–17.

Costello, A. B., & Osborne, J. W. (2005). Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Practical Assessment, Research & Evaluation*, 10(7), 1–9.

Cotton, F., & Conrow, F. (1998). An investigation of the predictive validity of IELTS amongst a group of international students studying at the University of Tasmania. *IELTS Research Reports* Vol 1, p. 72.

Deakin, G. (1997). IELTS in context: Issues in EAP for overseas students. *EA Journal*, 15(2), 7–1.

DiBello, L., Roussos, L., & Stout, W. (2007). Cognitive diagnosis part I. In C. R. Rao, & S. Sinharay (Eds.), *Handbook of statistics (Vol. 26): Psychometrics* (pp. 979–1030). Amsterdam: Elsevier.

Dooey, P., & Oliver, R. (2002). An investigation into the predictive validity of the IELTS Test as an indicator of future academic success. *Prospect*, 17(1), 36–54.

Gerbing, D. W., & Hamilton, J. G. (1996). Viability of exploratory factor analysis as a precursor to confirmatory factor analysis. *Structural Equation Modeling: A Multidisciplinary Journal*, 3(1), 62–72.

Glaser, B. G., & Strauss, A. L. (2017). *Discovery of grounded theory: Strategies for qualitative research.* Routledge.

Golder, K., Reeder, K., & Fleming, S. (2009). Determination of appropriate IELTS band score for admission into a program at a Canadian post-secondary polytechnic institution. *IELTS Research Reports* Vol 10, 1. Canberra: IELTS Australia and London: British Council.

Gomez, P. G., Noah, A., Schedl, M., Wright, C., & Yolkut, A. (2007). Proficiency descriptors based on a scale-anchoring study of the new TOEFL iBT reading test. *Language Testing*, 24(3), 417–444.

Government of Canada. (2014). *Canada's international education strategy*. Retrieved from: http://international.gc.ca/global-markets-marches-mondiaux/assets/pdfs/overview-apercu-eng.pdf

Green, A., Ünaldi, A., & Weir, C. (2010). Empiricism versus connoisseurship: Establishing the appropriacy of texts in tests of academic reading. *Language Testing*, 27(2), 191–211.

Hamp-Lyons, L. (2011). *What Is EAP? Handbook of Research in Second Language Teaching and Learning*, 2, 89.

Hartz, S., Roussos, L., & Stout, W. (2002). *Skills diagnosis: Theory and practice. User Manual for Arpeggio software*. Educational Testing Services.

Henson, R., & Templin, J. (2007). *Large-scale language assessment using cognitive diagnosis models*. Paper presented at the annual meeting of the National Council for Measurement in Education, April 2007, Chicago, IL.

Hinkin, T. R. (1998). A brief tutorial on the development of measures for use in survey questionnaires. *Organizational Research Methods*, 1(1), 104-121.

Hurley, A. E., Scandura, T. A., Schriesheim, C. A., Brannick, M. T., Seers, A., Vandenberg, R. J., & Williams, L. J. (1997). Exploratory and confirmatory factor analysis: Guidelines, issues, and alternatives. *The International Journal of Industrial, Occupational and Organizational Psychology and Behavior*, 18(6), 667–683.

Hyatt, D., & Brooks, G. (2009). Investigating stakeholders' perceptions of IELTS as an entry requirement for higher education in the UK. *IELTS Research Reports* Vol 10, 1. Canberra: IELTS Australia and London: British Council.

Ingram, D., & Bayliss, A. (2007). IELTS as a predictor of academic language performance, Part 1. *IELTS Research Reports* Vol 7, 1. Canberra: IELTS Australia and London: British Council.

Jang, E. E. (2005). *A validity narrative: Effects of reading skills diagnosis on teaching and learning in the context of NG-TOEFL*. Unpublished doctoral dissertation, University of Illinois at Urbana–Champaign, Urbana, IL.

Jang, E. E. (2009). Demystifying a Q-matrix for making diagnostic inferences about L2 reading skills. *Language Assessment Quarterly*, 6(3), 210–238.

Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement*, 25(3), 258–272.

Kane, M. T. (2006). Validation. *Educational Measurement*, 4(2), 17–64.

Khalifa, H., & Weir, C. J. (2009). *Examining reading* (Vol. 29). Ernst Klett Sprachen.

Kim, A. Y. (2015). Exploring ways to provide diagnostic feedback with an ESL placement test: Cognitive diagnostic assessment of L2 reading ability. *Language Testing*, 32(2), 227–258.

Krueger, R. A., & Casey, M. A. (2000). *Focus groups: a practical guide for applied research*, 3-19.

Lee, Y. W., & Sawaki, Y. (2009). Application of three cognitive diagnosis models to ESL reading and listening assessments. *Language Assessment Quarterly*, 6(3), 239–263.

Li, H., Hunter, C. V., & Lei, P. W. (2016). The selection of cognitive diagnostic models for a reading comprehension test. *Language Testing*, 33(3), 391–409.

McDowall, C., & Merrylees, B. (1998). Survey of receiving institutions' use and attitude to IELTS. *IELTS Research Reports* Vol 1, 116.

McNamara, T., & Roever, C. (2006). *Language testing: The social dimension* (Vol. 1). John Wiley & Sons.

Milanovic, M., & Weir, C. J. (2010). Series editors' note. *Relating language examinations to the Common European Framework of Reference for Languages: Case studies and reflections on the use of the Council of Europe's Draft Manual*, viii–xx.

Moore, T., Morton, J., & Price, S. (2012). Construct validity in the IELTS Academic Reading test: A comparison of reading requirements in IELTS test items and in university study. *IELTS Research Reports* Vol 11, 1–89. Canberra: IDP: IELTS Australia and London: British Council.

Morgan, D. (2008). Focus groups. In L. M. Given (Ed.), *The SAGE encyclopedia of qualitative research methods* (pp. 353–354). Thousand Oaks, CA: SAGE Publications Ltd. doi: 10.4135/9781412963909.n178

O'Loughlin, K. (2008). The use of IELTS for university selection in Australia: A case study. *IELTS Research Reports*, Vol 8, 1. Canberra: IELTS Australia.

Onwuegbuzie, A. J., Dickinson, W. B., Leech, N. L., & Zoran, A. G. (2009). A qualitative framework for collecting and analyzing data in focus group research. *International Journal of Qualitative Methods*, 8(3), 1–21.

Ravand, H. (2016). Application of a cognitive diagnostic model to a high-stakes reading comprehension test. *Journal of Psychoeducational Assessment*, 34(8), 782–799.

Ravand, H., & Robitzsch, A. (2015). Cognitive diagnostic modeling using R. *Practical Assessment, Research & Evaluation*, 20, 1–12.

Rea-Dickins, P., Kiely, R., & Yu, G. (2007). Student identity, learning and progression: The affective and academic impact of IELTS on 'successful 'candidates. *IELTS Research Reports*, Volume 7, 1. Canberra: IELTS Australia and London: British Council.

Robitzsch, A., Kiefer, T., George, A. C., Uenlue, A., & Robitzsch, M. A. (2019). Package 'CDM'. *Handbook of diagnostic classification models*. New York: Springer.

Rupp, A. A., Templin, J., & Henson, R. A. (2010). *Diagnostic assessment: Theory, methods, and applications.* New York: Guilford.

Sawir, E., Marginson, S., Forbes-Mewett, H., Nyland, C., & Ramia, G. (2012). International student security and English language proficiency. *Journal of Studies in International Education*, 16(5), 434–454.

Schmitt, T. A. (2011). Current methodological considerations in exploratory and confirmatory factor analysis. *Journal of Psychoeducational Assessment*, 29(4), 304–321.

Shanahan, T., & Shanahan, C. (2012). What is disciplinary literacy and why does it matter? *Topics in Language Disorders*, 32(1), 7–18.

Sinharay, S., Haberman, S. J., & Lee, Y. H. (2011). When does scale anchoring work? A case study. *Journal of Educational Measurement*, 48(1), 61–80.

Smith, H., & Haslett, S. (2007). Smith, H., & Haslett, S. (2007). Attitudes of tertiary key decision-makers towards English language tests in Aotearoa New Zealand: Report on the results of a national provider survey. *IELTS Research Reports*, Vol 7, 1–44. Canberra: IELTS Australia and London: British Council.

Strauss, A., & Corbin, J. (1998). *Basics of qualitative research* (2nd ed.). Thousand Oaks, CA: Sage.

Taylor, L., & Weir, C. J. (Eds.). (2012). *IELTS collected papers 2: Research in reading and listening assessment* (Vol. 2). Cambridge University Press.

Teddlie, C., & Tashakkori, A. (2006). A general typology of research designs featuring mixed methods. *Research in the Schools*, 13(1), 12–28.

University of Toronto. (2016–17). *Enrolment report*. Retrieved at: UT: http://www.provost. utoronto.ca/Assets/Provost+Digital+Assets/Enrolment+Report+2016-17+-+Final.pdf

Wang, C., & Gierl, M. J. (2011). Using the attribute hierarchy method to make diagnostic inferences about examinees' cognitive skills in critical reading. *Journal of Educational Measurement*, 48(2), 165–187.

Weir, C., Hawkey, R., Green, A., & Devi, S. (2012). The cognitive processes underlying the academic reading construct as measured by IELTS. *IELTS Collected Papers 2: Research in Reading and Listening Assessment*, 2, 212.

Weir, C., Hawkey, R., Green, A., Unaldi, A., & Devi, S. (2009). The relationship between the academic reading construct as measured by IELTS and the reading experiences of students in their first year of study at a British university. *IELTS Research Reports* Vol 9, 97. Canberra: IELTS Australia and London: British Council.

Weir, C., & Khalifa, H. (2008). A cognitive processing approach towards defining reading comprehension. *Cambridge ESOL: Research Notes*, 31, 2–10.

Weir, C. J., & O'Sullivan, B. (2011). *Test development and validation*. Palgrave Macmillan.

Yi, Y. S. (2017). Probing the relative importance of different attributes in L2 reading and listening comprehension items: An application of cognitive diagnostic models. *Language Testing*, 34(3), 337–355.

Appendix A: Phase 1 focus group protocol – student

<center>**Before Focus Group**</center>

1. [To be filled out by the team]

Date:	
Place:	
Team members:	
# of participants:	
Field of study	
Participant type	

2. Participants are given the following documents to be collected before the focus group starts:
- Consent form
- Student background questionnaire
- Name tag

<center>**During Focus Group**</center>

3. Greetings and introduction
- Introduce the team
- Explain the purpose of the study and focus group
- Set rules for FG

4. FG discussions

1.   REASON FOR APPLYING TO UOFT

1.1.1.   When did you decide to apply to U of T and why?
Probing:

2. PREPARATION FOR ADMISSION

1.1.2.   How did you prepare your application for admission?
Probing:

1.1.3.   Let's talk more about language requirements. Which test did you take for admission? How was your experience with the test?
Probing:

1.1.4.   What did the test score say about your English language proficiency?
Probing: Was it a fair representation of your ability?

3. LANGUAGE DEMANDS IN THE CLASSROOM

1.1.5.   Let's go back to the first week of the school semester. Can you describe your experience in the first week in terms of language demands?
Probing:

1.1.6.   How about language demands outside of the classroom?
Probing:

4. READING DEMANDS

1.1.7. Since the start of the semester, you may have read a number of course materials, including textbooks, online resources, manuals, lecture notes etc. What reading materials do you usually read?
Probing: Are there any challenges with these materials?
Probing: How do cope with them? Do you use any particular reading strategies?

1.1.8. Now, we'd like to think about specific reading processes and strategies. Here is a list of reading strategies. Take a moment to read and rate your level for each skill. Also, indicate if you think you need to work on it or have improved over the past months.

Distribute the self-assessment form

The form includes:
- Reading strategies
- Relevance (is it an important skill for my field of study)?
- Current ability level (Likert)
- Difficulty level
- Rate of improvement

Upon finishing the form, students keep the form for discussion and return it later.

1.1.9. Is there any strategy that is not relevant to your field of study?

1.1.10. Is there any strategy that is not included in the list but critical for your field of study?

1.1.11. What is the most difficult strategy to master?

5. LANGUAGE SUPPORT

1.1.12. Have you sought any help to deal with English language demands?
Probing: What kind?
Probing: How useful was it?
Probing: Are there any supports you wish the school provided for international students' language needs?

Appendix B: Phase 1 focus group protocol – instructor

Before Focus Group

Instructors will be asked to share online or bring hard copy materials from a specific course based on their experience with international students in that course. They will be asked to share their syllabi to help us understand the full language demands. The course selected should be one where international students have the most difficulty.

1. [To be filled out by the team]

Date:	
Place:	
Team members:	
# of participants:	
Field of study	
Participant type	

2. Participants are given the following documents to be collected before the focus group starts:
- Consent form (Collect before starting the focus group)
- Name tag

During Focus Group

3. Greetings and introduction
- Introduce the team
- Explain the purpose of the study and focus group
- Set rules for FG

Hello and thank you so much for taking the time to join today's focus group. I am _____. Our research project was funded by Cambridge IELTS and its main purpose is to enhance how we interpret test scores and use information for making informed decisions about student admission and further support. One of the main concerns from the field is to do with a mismatch between what the test measures and what students are expected to have for successful academic work in a specific discipline like engineering. So we're seeking input from students and instructors to better understand specific reading demands in a specific discipline (such as Engineering, Commerce and Economics) and facilitate discussions about more effective support strategies across the campus. As you may well know, I am here to facilitate your discussions, and please feel free to interact with other focus group members. Do you have any question? Great. Let's get started.

4. FG discussions

1.   GENERAL LANGUAGE DEMANDS

1.1.1   How do international students do in your courses? Let's start with first year students?
Probing: What is your general observation of their language proficiency?
How well are they prepared when they start the program?
Where do they struggle most? Why?

1.1.2   How about upper year students? What changes do you see as students progress to upper years?
        Probing: Do students improve their language proficiency?
        Where do you see the most improvement?
        Where do they struggle even after spending time here?

1.1.3   Now let's delve into specific academic reading demands in your courses.

        We thought it may be more useful to ground our discussions in specific course materials, so we asked you to share and bring some. Now take a look at the materials that you brought. Can you briefly share the following: Which class is this for? (Field general/specific) and Who is the targeted audience?

1.1.4   Can you identify some key reading activities students are expected to do?
        Probing: Are there any unique skills your program students should have for your course using the material?
        Where do students struggle most?

1.1.5   Did you have any successful experiences with supporting students struggling with specific reading demands?
        Probing: Do you have any suggestions for other faculty teaching international students for the first time?
        What other areas do you think students are in need of?
        Are you aware of any initiatives (e.g. professional development workshops) or supports provided to faculty to improve how they support/teach international students?

Appendix C: Domain analysis of academic language and literacy demands

COMMERCE			
Year	Course	Text type	Evaluation
1	MGT100 Fundamentals of Management	Textbook Chapters (13)	Midterm Exam (39%) Final Exam (50%) Research Requirement (1%) Online homework/Quizzes (10%)
1	RSM100 Introduction to Management	Textbook Chapters (13)	Midterm Exam (39%) Final Exam (50%) Research Requirement (1%) Online homework/Quizzes (10%)
2	RSM230 Financial Markets	Textbook Chapters (10)	Midterm Exam (25%) Final Exam (40%) Online homework/Quizzes (10%) Essays (10%) Group Case Assignment (15%)
3	RSM352 Marketing Research	Textbook Chapters (9) Case Studies (6)	Midterm Exam (30%) Research Requirement (3%) Group Assignment (Paper + Presentation) (35%) Class Participation/Contribution (17%) Individual Case Assignment (15%)
3	RSM361 Human Resource Management	Textbook Chapters (15)	Midterm Exam (30%) Final Exam (32%) Research Requirement (3%) Group Assignment (Presentation) (20%) Class Participation/Contribution (10%) News Briefings (5%)
4	RSM466 Environmental and Social Responsibility for Management	Textbook Chapters (12) Professional Journals (4)	Take-home Exam/Reflection (25%) Group Assignment (35%) Class Participation/Contribution (10%) Individual Case Assignment (x2) (30%)
4	RSM457H Strategic Marketing Communication	Professional Journals (13) Case Studies (9)	Midterm Exam (20%) Final Exam (30%) Research Requirement (3%) Group Assignment (Presentation) (35%) Class Participation/Contribution (12%)

ECONOMICS			
Year	Course	Text type	Evaluation
1	ECO101 Principles of Microeconomics	Textbook Chapters (16) Handout(6)	Textbook Chapters (46%) Final Exam (40%) Tutorial/Class Participation (4%) Portal Quizzes/Online Tests (9%) Warm-up Exercise (1%)
1	ECO102H Principles of Macroeconomics	Textbook Chapters (12)	Midterm Exam (46%) Final Exam (40%) Tutorial/Class Participation (4%) Portal Quizzes/Online Tests (9%) Warm-up Exercise (1%)
2	ECO206 Microeconomic Theory	Textbook Chapters (18)	Midterm Exam (48%) Final Exam (25%) Tutorial/Class Participation (5%) Homework (8%) Assignment (writing)(20%)
2	ECO220 Quantitative	Textbook Chapters (11) Professional Journals	Midterm Exam (49%) Final Exam (25%) Tutorial/Class Participation (5%) Portal Quizzes/Online Tests (21%)
3	ECO365 International Monetary Economics	Textbook Chapters (7) Professional Journals	Midterm Exam (30%) Final Exam (50%) Problem Sets (20%)
3	ECO324 Economic Development	Textbook Chapters (15) Professional Journals (17)	Midterm Exam (30%) Final Exam (45%) Portal Quizzes/Online Tests (10%) Assignment (Writing) (15%)
4	ECO417 Economic Development Policy: Community Engaged Learning	Professional Journals	Reflections (40%) Course Overview (12%) Group Assignment (report/presentation) (48%)
4	ECO410 Mergers and Competition Policy	Textbook Chapters (5) Professional Journals (15)	Final Paper (30%) Tutorial/Class Participation (15%) Assignment (Writing) (30%) Group Assignment (Presentation) (25%)

ENGINEERING			
Year	Course	Text type	Evaluation
1	ECE110H Electrical Fundamentals	Textbook Chapters (16)	Midterm Exam (30%) Final Exam (45%) Lab (15%) Projects/Assignment (10%)
1	MAT188 Linear Algebra	Textbook Chapters (12)	Midterm Exam (35%) Final Exam (40%) Quizzes (2%) Practical (15%) Tutorial (8%)
2	ECE241 Digital Systems	Textbook Chapters (8)	Midterm Exam (30%) Final Exam (50%) Lab (10%) Project/Assignment (10%)
2	ENG259 Electromagnetism	Textbook Chapters (8)	Midterm Exam (32%) Final Exam (40%) Quizzes (23%) Tutorial (5%)
3	CHE353 Engineering Biology	Textbook Chapters	Midterm Exam (20%) Final Exam (60%) Tutorial (20%)
3	ECE302 Probability and Applications (Electrical/Computer Engineering)	Textbook Chapters (5)	Midterm Exam (25%) Final Exam (45%) Project/Assignment (10%) Quizzes (20%)
4	ECE472 Engineering Economic Analysis & Entrepreneurship	Textbook Chapters (7)	Final Exam (50%) Quizzes (50%)
4	CSC444 Software Engineering	Online Manuals	Midterm Exam (35%) Final Exam (40%) Group project (25%)

Appendix D: Characteristics of five CDMs

Model	Type	Link function
DINA (Junker & Sijtsma, 2001)	Non-compensatory	Identity
DINO (Templin & Henson, 2006)	Compensatory	Identity
G-DINA (de la Torre, 2011)	General	Identity
NC-RUM (DiBello et al., 1995; Hartz, 2002)	Non-compensatory	Log
C-RUM (Hartz, 2002)	Compensatory	Logit

Appendix E: Undergraduate language demand survey

Section A: Introduction and Consent

A1. I have read the above information.

☐ Yes

☐ No

A2. I agree to participate in the survey.

☐ Yes

☐ No

A3. I agree to participate in a focus group

☐ Yes

☐ No

A4. Name:

A5. Academic Major(s):

A6. Academic Minor(s):

A7. Current Year of Study (Year 1, Year 2, etc.):

A8. Email Address:

A9. Phone Number:

Section B: Demographic Information

The following section will ask about your demographic background

B1. What is your year of birth?

B2. What is your gender?

☐ Female

☐ Male

☐ Non-binary: _____

B3. Which language do you use most fluently?

B4. Do you use another language(s)?

☐ Yes

☐ No

B5. Indicate each and your proficiency level:

B6. Indicate each and your proficiency level:

B7. Were you born in Canada?

☐ Yes

☐ No

B8. In which country were you born?

B9. What year did you arrive in Canada?

B10. What is your current status?

    ☐   Citizen

    ☐   Permanent resident

    ☐   Study permit (Visa)

    ☐   Other

B11. What high school curriculum did you complete?

    ☐   Canadian curriculum in Canada

    ☐   Canadian curriculum outside of Canada

    ☐   Non-Canadian curriculum in Canada

    ☐   Non-Canadian curriculum outside of Canada

B12. What was your high school Grade Point Average (GPA) in English?

B13. What was your high school Grade Point Average (GPA) in Math?

B14. What was your overall high school Grade Point Average (GPA)?

B15. How much time did you take off?

Section C: Demographic Information, Part II

C1. What year did you start studying at the University of Toronto?

C2. Did you take any time off from University of Toronto?

    ☐   Yes

    ☐   No

C3. Select the language test or program you took as part of your admission to the University of Toronto?

    ☐   Cambridge English Language Assessment

    ☐   Canadian Academic English Language (CAEL) Assessment

    ☐   CanTEST

    ☐   ELDA/COPE

    ☐   ELDA/TOP

    ☐   IELTS

    ☐   International Foundation Program (IFP) course

    ☐   MELAB

    ☐   TOEFL cBT

    ☐   TOEFL iBT

    ☐   TOEFL PBT

    ☐   UT School of Continuing Studies, Academic English (Level 60)

    ☐   Other preparatory course at U of T

    ☐   None

C4. On the Cambridge English Language Assessment what was your score for...

    Reading:

    Writing:

    Speaking:

    Listening:

C5. On the CAEL what was your score for...

    Reading:

    Writing:

    Speaking:

    Listening:

C6. On the CanTEST what was your score for...

    Reading:

    Writing:

    Speaking:

    Listening:

C7. On the ELDA/COPE what was your score for…

    Writing:

    Listening:

    Reading:

C8. On the ELDA/TOP what was your band score?

C9. On the IELTS what was your score for...

    Reading:

    Writing:

    Speaking:

    Listening:

C10. In the International Foundation Program (IFP) course what was your grade?

C11. On the MELAB what was your score on...

    Listening:

    GCVR:

    Writing:

    Speaking:

C12. On the TOEFL cBT what was your score on...

    Listening:

    Structure/Writing:

    Reading:

C13. On the TOEFL iBT what was your score for...

    Reading:

    Writing:

    Speaking:

    Listening:

C14. On the TOEFL PBT what was your score for...

     Reading Comprehension:

C15. In the UT School of Continuing Studies Academic English (Level 60) course
     what was your grade?

C16. In the other preparatory course at U of T what was your grade?

C17. Did you think your language test score reflected your language proficiency
     at that time?

    ☐   Yes

    ☐   No

C18. Please explain why you feel this way:

C19. After you met the language requirements for university admission
     (e.g., achieved IELTS 6.5 or high school completion) and started your university
     program, did you feel you were prepared for the language demands in your
     courses?

    ☐   Yes

    ☐   No

C20. In what faculty are you registered?

    ☐   Arts and Science

    ☐   Applied Science and Engineering

    ☐   Music

    ☐   Architecture

    ☐   Kinesiology

    ☐   Other

C21. For your current program, what year are you in?

    ☐   1

    ☐   2

    ☐   3

    ☐   4

    ☐   Other

C22. What is your current major or specialist?

C23. Have you changed your major?

    ☐   Yes

    ☐   No

C24. Are you considering changing your major?

    ☐   Yes

    ☐   No

C25. What is your career goal after graduation?

C26. Why are you interested in this career?

C27. What kind of skills do you think are necessary for the career in which you are interested? (You can choose more than one.)

☐ Communication

☐ Teamwork

☐ Problem Solving

☐ Initiative and Enterprise

☐ Planning and Organising

☐ Self-management

☐ Learning

☐ Technology

C28. Select all skills in which you feel prepared:

☐ Communication

☐ Teamwork

☐ Problem Solving

☐ Initiative and Enterprise

☐ Planning and Organising

☐ Self-management

☐ Learning

☐ Technology

C29. Please explain why you feel this way.

C30. What is your Cumulative Grade Point Average (CGPA)?

C31. Does your Cumulative Grade Point Average (CGPA) reflect your academic ability?

☐ Yes

☐ No

C32. Please explain why you feel this way:

Section D: Language Demands, Part I

The following section will ask about your academic language demands

D1. Reflecting on your school-work since the start of the current school year, how often do you do the following...
(Never, Rarely, Sometimes, Often, Most of the time, Always)

- Communicate with course instructors/tutors/TAs to discuss coursework

- Write a short assignment (e.g., lab report, short paper)

- Write a long paper (e.g., final term paper)

- Take notes during lectures

- Give oral presentations

- Participate in online discussions

- Participate in group projects

- Read lecture notes or PowerPoint slides

- Read long academic materials (e.g., articles, book chapters)

- Read professional magazines (e.g., *Accounting Today, American Banker, The Economist*)

- Read digital materials online

- Read text with statistics tables and figures

- Read entrepreneurial, legal and government documents

- Read text in a language other than English

- Read literature and/or fiction (e.g., novels, short stories, poetry)

- Read mathematical equations and computational programming codes

- Understand written instructions for assignments and tests

- Scan and skim for keywords in course readings

- Understand main ideas in course readings

- Summarise main ideas in course readings

- Distinguish the main idea from minor details in course readings

- Distinguish facts from opinions in course readings

- Understand technical vocabulary in course readings

- Understand cultural and idiomatic expressions in course readings

- Understand grammatically complex text in course readings

- Make prediction based on course readings

- Evaluate a writer's viewpoint using additional sources

- Generate questions based on course readings

- Understand implied ("hidden") meanings in course readings

- Solve problems by applying the information from a course reading to real life

- Read a large amount of materials in a limited time

- Have sufficient background knowledge about other cultures in order
  to understand course readings

Section E: Language Demands, Part II

E1. Reflecting on your school-work since the start of the current school year,
    how important is it to do the following...
    (Not well at all, Not well, Somewhat not well, Somewhat well, Well, Very well)

- Communicate with course instructors/tutors/TAs to discuss coursework

- Write a short assignment (e.g., lab report, short paper)

- Write a long paper (e.g., final term paper)

- Take notes during lectures

- Give oral presentations

- Participate in online discussions

- Participate in group projects

- Read lecture notes or PowerPoint slides

- Read long academic materials (e.g., articles, book chapters)

- Read professional magazines (e.g., *Accounting Today, American Banker,
  The Economist*)

- Read digital materials online

- Read text with statistics tables and figures

- Read entrepreneurial, legal and government documents

- Read text in a language other than English

- Read literature and/or fiction (e.g., novels, short stories, poetry)

- Read mathematical equations and computational programming codes

- Understand written instructions for assignments and tests

- Scan and skim for keywords in course readings

- Understand main ideas in course readings

- Summarise main ideas in course readings

- Distinguish the main idea from minor details in course readings

- Distinguish facts from opinions in course readings

- Understand technical vocabulary in course readings

- Understand cultural and idiomatic expressions in course readings

- Understand grammatically complex text in course readings

- Make prediction based on course readings

- Evaluate a writer's viewpoint using additional sources

- Generate questions based on course readings

- Understand implied ("hidden") meanings in course readings

- Solve problems by applying the information from a course reading to real life

- Read a large amount of materials in a limited time

- Have sufficient background knowledge about other cultures in order to understand course readings

Section F: Language Demands, Part III

F1. Reflecting on your school-work since the start of the current school year, how well can you do the following...
(Not well at all, Not well, Somewhat not well, Somewhat well, Well, Very well)

- Communicate with course instructors/tutors/TAs to discuss coursework

- Write a short assignment (e.g., lab report, short paper)

- Write a long paper (e.g., final term paper)

- Take notes during lectures

- Give oral presentations

- Participate in online discussions

- Participate in group projects

- Read lecture notes or PowerPoint slides

- Read long academic materials (e.g., articles, book chapters)

- Read professional magazines (e.g., *Accounting Today, American Banker, The Economist*)

- Read digital materials online

- Read text with statistics tables and figures

- Read entrepreneurial, legal, and government documents

- Read text in a language other than English

- Read literature and/or fiction (e.g., novels, short stories, poetry)

- Read mathematical equations and computational programming codes

- Understand written instructions for assignments and tests

- Scan and skim for keywords in course readings

- Understand main ideas in course readings

- Summarise main ideas in course readings

- Distinguish the main idea from minor details in course readings

- Distinguish facts from opinions in course readings

- Understand technical vocabulary in course readings

- Understand cultural and idiomatic expressions in course readings

- Understand grammatically complex text in course readings

- Make prediction based on course readings

- Evaluate a writer's viewpoint using additional sources

- Generate questions based on course readings

- Understand implied ("hidden") meanings in course readings

- Solve problems by applying the information from a course reading to real life

- Read a large amount of materials in a limited time

- Have sufficient background knowledge about other cultures in order to understand course readings

Thank you for completing the survey. Your answers to the questions in each section of the survey are important to the research study.

Appendix F: Language competence CFA mode
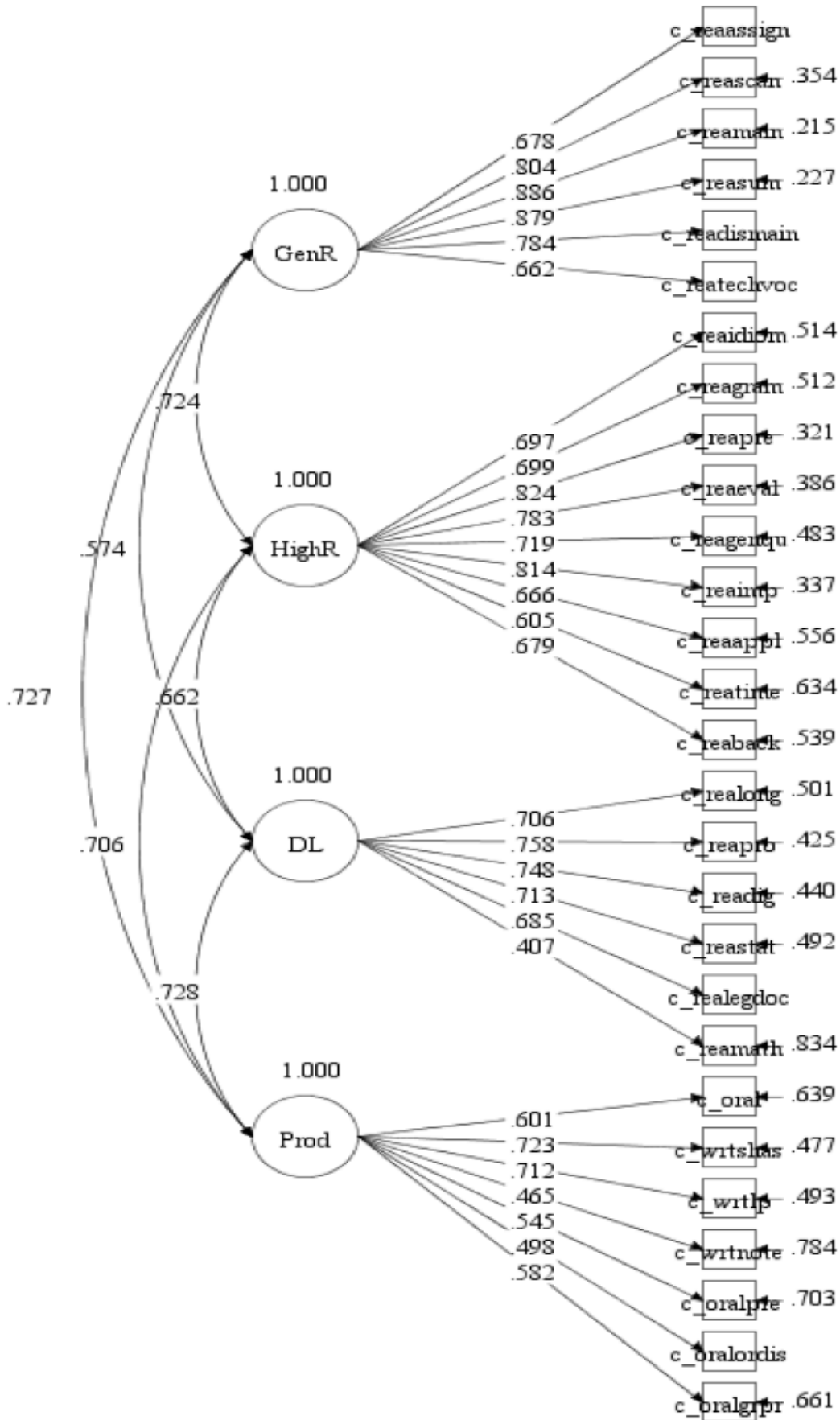
Appendix G: Language importance CFA mode
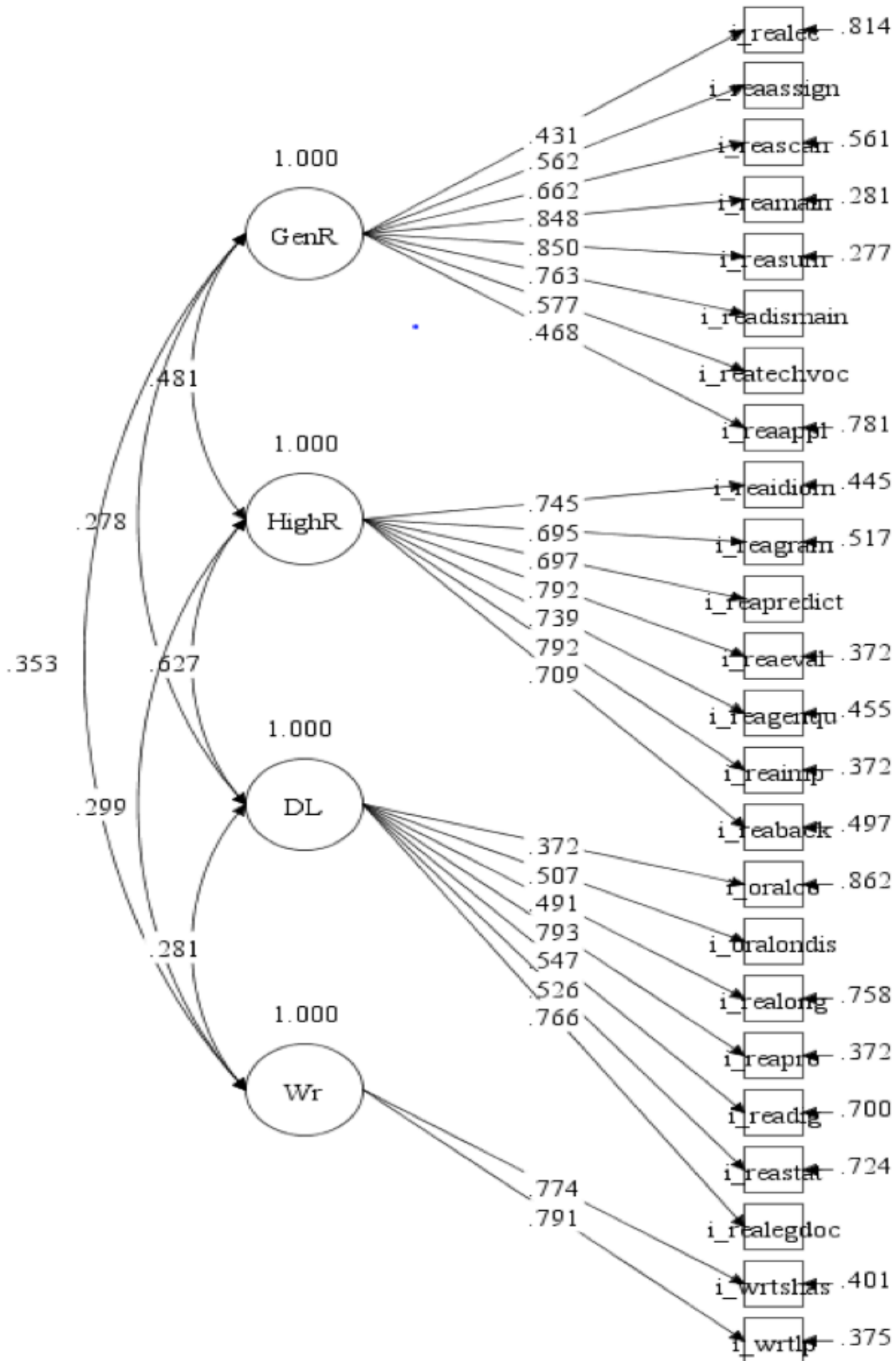
Appendix H: Phase 3 focus group protocol – student

Before Focus Group

Date:	
Place:	
Team members:	
Participants:	

Participants are given the following documents to be collected before the focus group starts:

- Consent form
- Student background questionnaire
- Can-do descriptor document "IELTS Reading Skill Report" (with draft letter)
- Name tag

During Focus Group

1. GREETINGS AND PROJECT INTRODUCTION

- Introduction of researchers and participants
- Project background and description
- Purpose of focus group and general rules
- Questions?

2. PREPARATION FOR ADMISSION

- Once you met the required IELTS reading band score, do you think you were well-prepared for the reading demands at UofT?
    Probing: In terms of reading skills, how well prepared were you when you started at UofT?
    What reading skills do you feel you need to develop to succeed in your program?
    Think about your IELTS reading band score. What did the test score say about your reading skills? What do you think students can do with a ___ score?

3. FEEDBACK ON THE "IELTS Reading Skill Report"

- [Introducing the Can-Do Descriptor Form – IELTS Reading Skill Report]:
    We are now passing around the document we explained earlier. We suggest this document, "IELTS Reading Skill Report," can be attached to the university's admission letter to future international students.

- Although students might have achieved the minimum score to get admitted, they usually do not understand what the reading score they received means in terms of what they can or cannot do. Moreover, through the previous stages of this research project, we have learned that there are still many language-related (or reading-related) challenges that students face once they start at UofT.

- Before finalising this report, we want to get some feedback from current students who have already gone through the experience as an international student. Please take some time to carefully read this "IELTS Reading Skill Report". Then, we will ask you some questions about this report.

- Are there any words or phrases that are confusing? If yes, how can we improve it?
    (Ease of interpretability; understandable)
    Probing: Is the formatting and description easy to understand?
    If not, how can we improve it?

- Think about your IELTS score and your reading ability when you took the IELTS test. Do you think that the descriptions are accurate? **(Accuracy of skills described)**
  - Probing: Are there any skills that are too easy or too difficult, and should be moved to a different band score?

- Thinking about your IELTS score and reading ability, are there any important skills that you had that aren't mentioned in the document? **(Missing skills)**
  - Probing: What are the most important reading skills that you've mastered?
  - What important reading skills do you struggle with the most?
  - Are these skills important for everyone, or just your field of study? (DL)

- How important are the reading skills we described for your field of study? **(DL piece)**
  - Probing: What reading skills do students need to master before they are prepared for your field of study's language demands?

- Look at the areas of improvement table. Do you think those skills are important to improve? **(Improvement table accuracy)**
  - Probing: What other reading skills do all students need to work on?

4. WRAPPING UP

Thank you so much for your time. Hearing your experiences is extremely helpful in understanding the language demands that international students face at UofT.

Do you have any questions or comments before we wrap up?

Appendix I: Phase 3 focus group protocol – instructor

Before Focus Group

Date:	
Place:	
Team members:	
Participants:	

Participants are given the following documents to be collected before the focus group starts:

- Consent form
- Name tag

During Focus Group

1. GREETINGS AND PROJECT INTRODUCTION

2. PRESENTATION OF THEME 1:
Phase 1 project findings (Student and Instructor Focus groups and thematic analyses)
     • Open for discussion/feedback/response

3. PRESENTATION OF THEME 2:
Findings and analyses from the Undergraduate Language Demands Survey
     • Open for discussion/feedback/response

4. PRESENTATION OF THEME 3:
Description of scale anchoring findings and sharing of proposed enriched communication to incoming students
     • Open for discussion/feedback/response

5. RECOMMENDATIONS:
     • Open discussion regarding recommendations, applications within their own programs/faculties

Appendix J: Self-assessment questionnaire to stimulate participant thinking during Phase 1 focus groups

Section A: Reading skills

A. I can do the following in my courses...
(Strongly Agree, Agree, Somewhat Agree, Somewhat Disagree, Disagree, Strongly Disagree)

- Understanding the main points and supporting details in course readings.
- Understanding hidden meanings ("reading between the lines") in course readings.
- Understanding information that is directly stated in a course reading.
- Understanding how parts of a text are organised.
- Connecting what is written in the text to my own knowledge.
- Understanding and identifying different types of texts (e.g. reports, government documents, etc.) on a topic.
- Reading a text and understanding why it was written and who it was written for.
- Using my vocabulary and sentence knowledge to understand technical words in a text.
- Using my vocabulary and sentence knowledge to understand non-technical words in a text.
- Critically evaluating the claims, evidence or data presented in a printed text.
- Evaluating the validity and accuracy of digital texts.
- Drawing on ideas from a range of texts to support my own argument.
- Connecting ideas from a variety of text types and media.
- Interpreting graphs, tables, and other types of data visualisation.
- Using information from texts to make predictions.

Section B: Reading Skills Importance

B. This is a very important skill for students in my program.
(Strongly Agree, Agree, Somewhat Agree, Somewhat Disagree, Disagree, Strongly Disagree)

- Understanding the main points and supporting details in course readings.
- Understanding hidden meanings ("reading between the lines") in course readings.
- Understanding information that is directly stated in a course reading.
- Understanding how parts of a text are organised.
- Connecting what is written in the text to my own knowledge.
- Understanding and identifying different types of texts (e.g. reports, government documents, etc.) on a topic.
- Reading a text and understanding why it was written and who it was written for.
- Using my vocabulary and sentence knowledge to understand technical words in a text.
- Using my vocabulary and sentence knowledge to understand non-technical words in a text.
- Critically evaluating the claims, evidence or data presented in a printed text.
- Evaluating the validity and accuracy of digital texts.
- Drawing on ideas from a range of texts to support my own argument.
- Connecting ideas from a variety of text types and media.
- Interpreting graphs, tables, and other types of data visualisation.
- Using information from texts to make predictions.

Appendix K: *p*-values for items between skill master and non-masters

	FORM A				FORM B		
Item	**Master (m)**	**Non-master (nm)**	**(m – nm)**	**Item**	**Master (m)**	**Non-master (nm)**	**(m – nm)**
1	0.995	0.333	0.662	1	1.000	0.121	0.879
2	0.938	0.272	0.666	2	0.942	0.441	0.500
3	0.981	0.595	0.387	3	0.612	0.526	0.086
4	0.963	0.725	0.239	4	0.773	0.489	0.284
5	0.183	0.056	0.127	5	0.513	0.519	- 0.006
6	0.906	0.142	0.765	6	0.830	0.399	0.432
7	0.909	0.585	0.325	7	0.812	0.457	0.356
8	0.819	0.647	0.172	8	0.478	0.484	- 0.006
9	0.577	0.261	0.316	9	0.522	0.300	0.222
10	0.902	0.493	0.409	10	0.572	0.355	0.217
11	0.703	0.392	0.310	11	0.727	0.327	0.399
12	0.655	0.086	0.570	12	0.896	0.466	0.430
13	0.570	0.105	0.465	13	0.882	0.545	0.337
14	0.942	0.400	0.542	14	0.748	0.232	0.516
15	0.880	0.595	0.285	15	0.316	0.102	0.214
16	0.954	0.792	0.163	16	0.966	0.536	0.430
17	0.865	0.667	0.198	17	0.964	0.623	0.341
18	0.892	0.852	0.041	18	0.634	0.127	0.507
19	0.624	0.283	0.341	19	0.756	0.341	0.415
20	0.848	0.681	0.167	20	0.688	0.354	0.333
21	0.907	0.700	0.207	21	0.666	0.399	0.268
22	0.728	0.464	0.264	22	0.737	0.655	0.082
23	0.833	0.340	0.492	23	0.699	0.405	0.294
24	0.928	0.662	0.266	24	0.729	0.391	0.338
25	0.887	0.446	0.441	25	0.701	0.446	0.255
26	0.826	0.207	0.619	26	0.842	0.479	0.363
27	0.867	0.582	0.285	27	0.830	0.559	0.271
28	0.687	0.408	0.279	28	0.642	0.400	0.242
29	0.759	0.388	0.371	29	0.588	0.428	0.160
30	0.385	0.334	0.051	30	0.367	0.327	0.040
31	0.694	0.306	0.388	31	0.613	0.346	0.268
32	0.953	0.445	0.508	32	0.868	0.416	0.452
33	0.888	0.375	0.513	33	0.800	0.384	0.416
34	0.884	0.383	0.501	34	0.797	0.374	0.423
35	0.855	0.342	0.513	35	0.762	0.341	0.421
36	0.855	0.480	0.375	36	0.787	0.493	0.294
37	0.692	0.023	0.669	37	0.505	0.018	0.487
38	0.746	0.143	0.603	38	0.610	0.117	0.493
39	0.636	0.087	0.549	39	0.516	0.076	0.440
40	0.639	0.144	0.494	40	0.514	0.155	0.359
Average	**0.794**	**0.405**	**0.388**	**Average**	**0.705**	**0.374**	**0.331**

Note. Items with shaded *m-nm* difference values indicate poor discrimination between masters and non-masters.