CARDWELL, RAMSEY LEE, Ph.D. Classification Consistency and Results Reporting of A Digital-First Computer-Adaptive Language Proficiency Test. (2022) Directed by Dr. Micheline Chalhoub-Deville and Dr. Richard Luecht. 117 pp.

The emergence of digital-first assessments is prompting reconsideration of, and innovation in, aspects of psychometrics, test validation, and test use. Using the Duolingo English Test (DET) as an example, this three-paper series seeks to address issues concerning the estimation of classification consistency and the reporting of results for such assessments. The first paper presents a simulation study investigating the use of CTT–based classification consistency methods in a computer-adaptive testing context. The second paper further investigates CTT–based classification consistency estimates by applying the methods from the first paper, as well as a bootstrapping-inspired approach, to operational test data from the DET. The third paper investigates the reporting of test-related information and test-taker results to results users through a focus group with North American postsecondary admissions professionals. Collectively, the studies address challenges in constructing validity arguments for digital-first high-stakes assessments.

CLASSIFICATION CONSISTENCY AND RESULTS REPORTING OF A DIGITAL-FIRST COMPUTER-ADAPTIVE LANGUAGE PROFICIENCY TEST

by

Ramsey Lee Cardwell

A Dissertation Submitted to
the Faculty of The Graduate School at
The University of North Carolina at Greensboro
in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Greensboro
2022

Approved by

Micheline Chalhoub-Deville
Committee Co-Chair

Richard Luecht
Committee Co-Chair

APPROVAL PAGE

This dissertation written by Ramsey Lee Cardwell has been approved by the following committee of the Faculty of The Graduate School at The University of North Carolina at Greensboro.

Committee Co-Chair _____
Micheline Chalhoub-Deville

Committee Co-Chair _____
Richard Luecht

Committee Members _____
Robert Henson

_____
Geoffrey LaFlair

_____
Date of Acceptance by Committee

_____
Date of Final Oral Examination

Contents

List of Tables

List of Figures

Chapter 1

Introduction

The measurement field, at least in the United States, has long been concerned with the reliability of measurement procedures. There are numerous methods for quantifying reliability through a standard error of measurement and a reliability coefficient within the frameworks of classical test theory (CTT), generalizability theory (GT), and item response theory (IRT) (Haertel, 2006). Additionally, the 1970s saw a proliferation of research on so-called criterion-referenced tests, which, given that the aim of such tests is to classify examinees (e.g., as masters and non-masters of a particular domain of knowledge and abilities), necessitated different methods for quantifying reliability (Sawaki, 2016). A common approach to quantifying the reliability (also called *dependability* or *precision*) of criterion-referenced assessments is to estimate probabilities of accurate and consistent classification into score-based performance categories (Sawaki, 2016); these are often referred to as classification accuracy and classification consistency, respectively. Reporting estimates of classification consistency is also required by the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014) for assessments that make such score-based classifications.

A research literature has developed around the topic of classification consistency, with numerous proposed methods under both the CTT and IRT frameworks. The earliest such methods date back to the 1970s (Haertel, 2006), predating the widespread use of IRT and computer-based testing. These early methods were thus within the CTT framework and were developed for fixed-form tests. Some of the later IRT–based methods were also developed for fixed-form tests. But with the advent and expanded application of computer adaptive testing (CAT), which traditionally operates in an IRT framework in order to permit the real-time selection of items and calculation of intermediate proficiency estimates, there have also been IRT–based methods developed

for such tests, in which each examinee may see a different set and even different number of items. However, to date there has been no research to systematically investigate the performance of fixed-form CTT–based methods when applied to CAT scores[1]. Such research is needed if for no other reason than to substantiate the deterrence of applying such methods in a CAT context.

Besides the expanded use of CAT in recent decades, there are other more recent developments that further complicate the calculation of classification consistency and motivate the interest in applying CTT–based methods in a CAT context. Only a decade ago, Lee (2010) used the term "complex assessment" to refer to mixed-format tests, such as one with both multiple-choice and constructed-response items. Today it is arguably outdated to call such assessments "complex," as the mantle of complexity has been taken up by emerging paradigms such as game-based assessment (Lay, Patton, & Chalhoub-Deville, 2017) and AI–driven computational psychometrics (von Davier, Deonovic, Yudelson, Polyak, & Woo, 2019). These new assessment paradigms are defined by high-dimensional data and computationally intensive score estimation, which pose substantial challenges to modeling and quantifying individual and conditional measurement error, a prerequisite to the IRT–based classification consistency methods. Under such circumstances, it is potentially desirable to use a more straightforward heuristic approach to estimating classification consistency.

An element that is largely missing from the literature on classification consistency is how this concept and the estimates are communicated to test results[2] users, and what results users should do with this information. This is in line with the more general observation by Zenisky & Hambleton (2016) that results reporting has traditionally been an afterthought of test development, meaning that issues of stakeholder communication are often not explicitly considered during the design and validation of assessments. The next frontier of classification consistency research is thus to investigate whether/how results users consider classification consistency estimates in both selecting tests for their particular use case and in interpreting individual examinee test results.

All of the aforementioned issues are connected fundamentally by the validity argument of the Duolingo English Test. In the argument-based framework of Kane (2006,

[1]In this dissertation, the terms "CAT score" and "score" are used to refer to examinee ability estimates on the θ scale or transformations thereof onto a reported score scale, unless otherwise specified.

[2]In this paper I primarily use the terms "results" and "results report," as suggested by O'Donnell & Zenisky (2020), to encompass all information about an examinee's performance that is communicated to stakeholders. This includes scores, but also more broadly verbal descriptions of examinee performance and ability, as well as samples of examinee performance. Any use of the term "score report" should be understood to potentially include results beyond numerical scores. I will use the term "score" to refer specifically to numerical scores, although quotes from other sources may use this term in a more general sense.

2013), validity is a series of inferences based on appropriate evidence leading from an examinee's raw responses to the interpretations and decisions made based on the test results. One of these inferences is the generalization inference, which is that an examinee's performance on the test allows for "claims about expected performance in a universe of possible observations" on tasks like those included on the test "or to an estimated trait value that can be used to draw conclusions about the future performances" (Kane, 2013, p. 10). The first two papers are squarely about methods for supporting the generalization inference. The third paper is less clearly connected to a particular inference, but rather is about the communication of test results and supporting information to results users in a way that directs them to act in accordance with the test's validity evidence. Communication with results users is generally lacking from the literature on classification consistency and on validity theory more broadly. Kane (2013) essentially mentions test users only to assign them primary responsibility for evaluating the consequences of test use (p. 57). But in reality, it is questionable to assume that every higher education institution will have the resources to thoroughly evaluate the use of each test they accept for admissions purposes. It is doubtful that all institutions even have the capacity to critically analyze the validity evidence provided by the test developer. It is therefore crucial that results reporting be considered integral to valid test use, and that more research be conducted on the optimal reporting of results to facilitate valid interpretation and use of test results.

1.1 Overview of the Three Papers

In this dissertation, I investigate issues related to classification consistency, starting with a systematic comparison of existing methods in a novel context, then moving to an application of these methods to real data from an operational assessment, and concluding with a consideration of how test results in general are communicated to and used by test results users. However, while all three papers in this dissertation are connected by this common theme, each is meant to stand alone as a publishable scholarly article. Each paper therefore has its own unique research questions as well as introduction, methods, and discussion sections.

The first paper utilizes a simulation study framework to investigate the performance of CTT–based classification consistency indices in a CAT context. The simulation study includes manipulated factors such as test length and the position of the cut score relative to the examinee ability distribution. The research questions of the first paper are:

1. How accurate are CTT–based classification consistency estimates (test–retest, split-half, and Livingston–Lewis) for CAT scores?

2. Are the results from a CAT context consistent with prior studies in fixed-form contexts?

3. Does each method exhibit a consistent bias?

The second paper builds on the first by applying the methods investigated in the simulation study to data from an operational, innovative CAT: the Duolingo English Test. The Duolingo English Test is innovative in that it was developed using machine learning and natural language processing techniques to estimate item difficulties, obviating the need for item piloting (Settles, LaFlair, & Hagiwara, 2020). However, this means that the items lack the empirical item parameters and individual standard errors necessary for applying existing IRT–based indices of classification consistency. Therefore, the second paper's research questions are:

1. How do results of applying CTT–based classification consistency estimates (test–retest, split-half, Livingston–Lewis) to real CAT scores compare to simulation results?

2. To what extent is it possible to apply one or more classification consistency methods of the IRT paradigm to data from the Duolingo English Test? How do the results compare to those of the CTT–based methods?

Finally, the third paper investigate how the primary population of Duolingo English Test results users—higher education admissions officers—interpret and use the test results, including psychometric information about the test. The high-level research questions of the third paper are as follows, with additional sub-questions provided in the paper:

1. What are the known or supposed informational needs of North American admissions officers as a results report audience of high-stakes English language proficiency (ELP) tests used for admissions purposes?

2. To what extent are current ELP tests' results reporting practices meeting their informational needs?

3. What steps can be taken to improve DET results reporting?

1.2 Assumptions and Delimitations

In order to clearly define the scope of this dissertation, listed below are some key assumptions and delimitations of the research.

1.2.1 Assumptions

1. A simulation study approach provides evidence of how methods perform under hypothetical circumstances, but cannot speak directly to the processes involved in a real-world situation.
2. The Duolingo English Test is primarily measuring a relatively stable latent trait related to English language proficiency.
3. Certified[3] results from different administrations of the same version[4] of the Duolingo English Test are directly comparable.
4. Two individuals with the same reported-scale overall score on the Duolingo English Test have similar abilities on the construct measured by the test.
5. While CTT–based classification consistency methods might not be optimal for a CAT, the direct comparability of results from different CAT sessions is functionally similar enough to parallel forms of fixed-form tests to make CTT–based methods permissible.
6. Classification consistency indices are useful to some test results users for some purpose.
7. A sample of currently practicing admissions officers would be sufficiently representative of admissions officers at similar institutions in the United States in terms of assessment understanding and use to draw conclusions about the presence (but not necessarily prevalence) of English proficiency test–related beliefs and behaviors in the higher-ed admissions population.

1.2.2 Delimitations

1. The simulation study is designed to somewhat resemble the Duolingo English Test in order for the results to be applicable to this context, but some differences are maintained in order to achieve broader generalizability to other CAT assessments. For example, the Duolingo English Test is variable-length with compound stopping criteria. But the simulation study will only consider fixed-length CATs with multiple length conditions.
2. Although the Duolingo English Test reports four subscores, the first two papers in this dissertation are only considering the overall score. The overall score is used for CEFR classifications and is also the sole basis for an admissions cut score at the majority of institutions.

[3]In Duolingo parlance, results are termed "certified" once the test session has gone through the proctoring review process and determined to be free from rule violations or technical errors that could compromise the validity of score interpretations.

[4]The Duolingo English Test has been periodically updated since its launch, including the addition of constructed response items and rescaling from a 100-point to a 160-point reported score scale. Comparability of results across test versions is not claimed.

3. The study on results reporting intends to focus on admissions officers, as this is the population primarily using Duolingo English Test results to make high-stakes decisions. Examinees are also an important audience of test results, but not within the scope of this research due to the practical limitations of direct access to a sample of examinees.
4. This dissertation focuses almost exclusively on classification consistency, as opposed to classification accuracy. Classification accuracy will be reported in the simulation study of the first paper because the "true" examinee abilities are known. However, assessing the classification accuracy of the Duolingo English Test is beyond the scope of this dissertation, as it would require additional measures to establish examinees' "true" classifications.

1.3 Significance

As discussed earlier in the introduction, classification consistency is not a new concept in the measurement field. Rather, what has changed is the advent of more complex assessment paradigms such as GBA and AI–driven assessment that are disrupting the previously exhaustive dichotomy of CTT and IRT. These developments necessitate a revisiting of existing classification consistency methods to systematically investigate their applicability to adaptive assessments that do not fit neatly within a traditional IRT framework. This dissertation seeks to help fill this gap in the literature. Furthermore, the lack of attention to the communication of test results and validity evidence to test stakeholders is a longstanding issue in the field. The third paper thus seeks to help the field continue to advance beyond idealized test results users and prototypical paper-based score reports by critically examining results-user needs and abilities and exploring results reporting possibilities in a digital context.

Chapter 2

Paper 1: CTT-Based Classification Consistency Indices in a CAT Context

2.1 Introduction

Although the field of psychometrics places great emphasis on the reliability of numerical test scores, when scores are used primarily to categorize examinees (e.g., as "low proficiency," "average proficiency," or "high proficiency") or to make categorical decisions about examinees (e.g., "pass" or "fail"), the consistency of such categorizations is arguably of greater importance (Deng & Hambleton, 2013). Indices of classification accuracy (i.e., decision accuracy) and classification consistency (i.e., decision consistency) are used to communicate the stability of such categorical decisions to stakeholders, such as test score users. Furthermore, it is considered best practice to report estimates of classification accuracy and consistency of tests used to make categorical decisions (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014). However, since the classification consistency of computer adaptive tests (CAT) is typically estimated within an IRT framework, the appropriateness of CTT–based classification consistency methods for CAT contexts remains under researched.

The definitions and operationalizations of both classification accuracy and consistency are based on the concepts of true scores and estimated scores. Whenever a test is administered, the test taker receives an estimated score—an imperfect estimate of their true score, or how they would perform in the absence of any measurement error (i.e., non-systematic score fluctuations due to factors other than the ability of interest, such

as fatigue, motivation, the particular sample of items presented, etc.). In addition, the test taker is often classified on the basis of the estimated score into categories such as "passing," "proficient," or "meets university admission criteria." Since their estimated score is influenced by measurement error, there is a possibility that a test taker will be misclassified. This possibility is quantified by classification accuracy, the rate at which test takers are categorized as they would be based on their true scores (Wheadon, 2014). Classification consistency is similar, but refers to the rate at which test takers fall into the same category on multiple independent test administrations, regardless of the accuracy of said classifications. The remainder of this paper focuses on classification consistency, starting with a closer look at its definition.

2.1.1 Definitional Issues

Classification consistency is superficially straightforward, but conceptual ambiguity arises when it is operationalized. Classification consistency has been variably defined as:

- "the agreement between the classifications based on two non-overlapping, equally difficult forms of the test" (Livingston & Lewis, 1995, p. 180);
- "the degree to which examinees are classified into the same categories over replications (typically two) of the same measurement procedure" (Lee, 2010, p. 1);
- "the percentage of candidates who are classified into the same proficiency category across two independent administrations (or parallel forms) of the same test" (Deng & Hambleton, 2013, p. 236);
- "the probability that [an examinee] would be classified with the same grade over successive administrations of a test" (Wheadon, 2014, p. 2);
- the "extent [to which] test-takers are consistently reported as having passed or failed" (Alger, 2016, p. 138);
- and "the degree to which examinees are classified into the same performance levels between independent, parallel forms of a test" (Diao & Sireci, 2018, p. 20).

The above definitions seem to agree that classification consistency is:

1. a probability or rate and
2. about performance/classification on multiple measurement occasions (sometimes specified as two and other times unspecified).

However, these definitions are less clear and consistent regarding the following aspects of classification consistency:

1. whether the probability is joint or conditional

2. which source(s) of error are presumed to affect classification

These latter two points are discussed in more detail below.

2.1.1.1 Which probability

The above definitions of classification consistency are not entirely clear about whether classification consistency should be quantified as a joint or conditional probability. Joint probability refers here to the probability of consistent classification (regardless of classification accuracy) and without prior knowledge of examinee performance on either measurement occasion. It is represented mathematically as $P(C_1 = C_2)$ where C_1 represents the categorization or decision made on the basis of the first measurement occasion and C_2 represents that of the second measurement occasion. A conditional probability of consistent classification, on the other hand, would be represented as $P(C_1 = C_2|X_1)$ where X1 represents the estimated score from the first measurement occasion. This seemingly minor distinction can have a larger implication for how classification consistency is interpreted by different stakeholders, a topic largely ignored in the literature on classification consistency.

Based on the definitions of classification consistency quoted above, the literature seems to prefer the joint probability interpretation, as none of the six definitions use language that would indicate a conditional probability interpretation (e.g., "given/conditional on an examinee's first test score"). From this perspective, an examinee with a true score of 85 on an exam with a passing cut score of 90 and a standard error of measurement (SEM) of 5 has a 0.16 probability of being misclassified on any given measurement occasion (corresponding to the shaded area in Figure 2.1) assuming normally distributed measurement error. This individual's joint probability of classification consistency on two independent measurement occasions would then be $(0.84)^2 + (0.16)^2 = 0.7312$. A classification consistency index for the test could then be obtained by averaging the values for each examinee in the test-taker population (essentially the approach of Rudner, 2000). Such an index would give a sense of the overall stability of classifications when the test is administered to the target population, which might be useful for evaluating the general appropriateness of a test for an intended use.

While the classification consistency index just described could be useful for evaluating a test, this joint probability could be misinterpreted by score users if presented in conjunction with individual examinee scores. It is known that non-statisticians have difficulty distinguishing and interpreting joint and conditional probabilities (Welsh, 2018). Thus, a classification consistency estimate of 0.73 might be misinterpreted as meaning that an examinee with an estimated score of 85 has a 73% chance of scoring below the cut score ($X_C = 90$) if they were to retake the test, which would be ascribing a conditional interpretation to a joint probability. Assuming that 85 is an

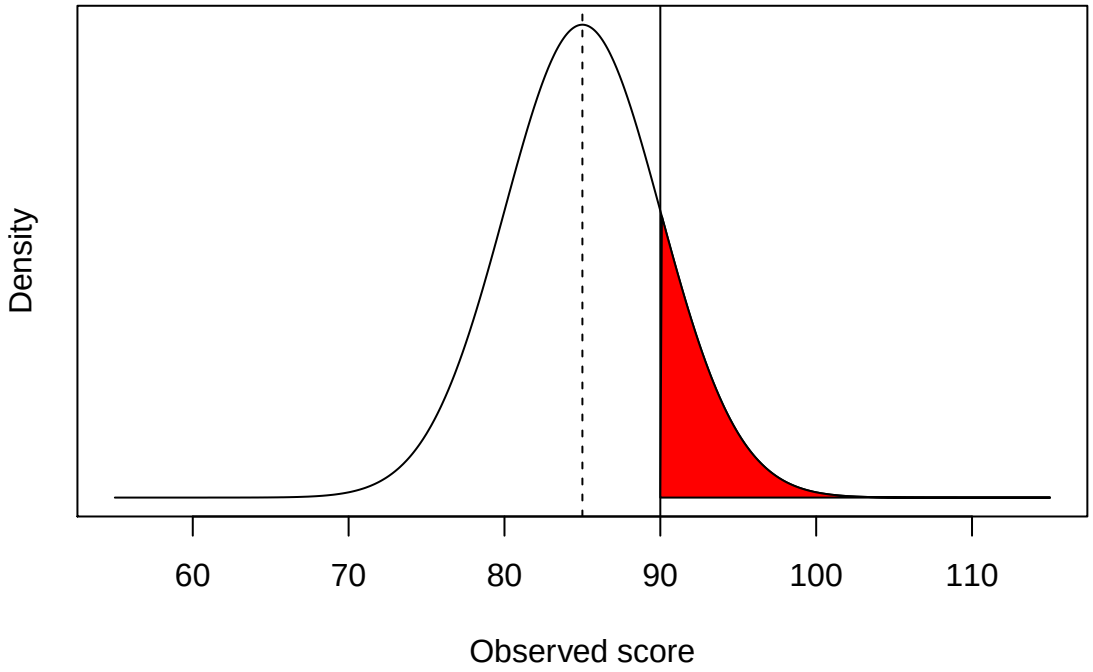

Figure 2.1. Hypothetical distribution of estimated scores for a given true score. *Note:* The dashed vertical line indicates the true score (85) while the solid vertical line represents the passing cut score (90). The shaded red area represents the probability of misclassification on a single measurement occasion (i.e., the complement of classification accuracy).

accurate estimate of the examinee's true score, the examinee would actually have an 84% chance of achieving the same classification on a second test attempt. The results user would therefore be underestimating the stability of the examinee's score estimate.

For the scenario depicted in Figure 2.1, the conditional probability of consistent classification on the second measurement occasion given, the examinee's score on the first measurement occasion, would be 0.84 (assuming $X_1 = T = 85$), corresponding to the probability of correct classification represented by the unshaded area under the curve. In other words, the conditional classification consistency is equivalent to the classification accuracy of a single measurement occasion, or $P(C_1 = C_2|X_1) = P(C_2 = C_T)$ where C_T represents the examinee's true classification, assuming $X_1 = T$ and the measurement occasions are truly parallel and independent (i.e., only subject to random measurement error).

Figure 2.2 depicts the relationship between the joint and conditional approaches to classification consistency over a range of classification accuracy. At relatively low (< 0.55) or high (> 0.95) levels of classification accuracy, the conditional and joint

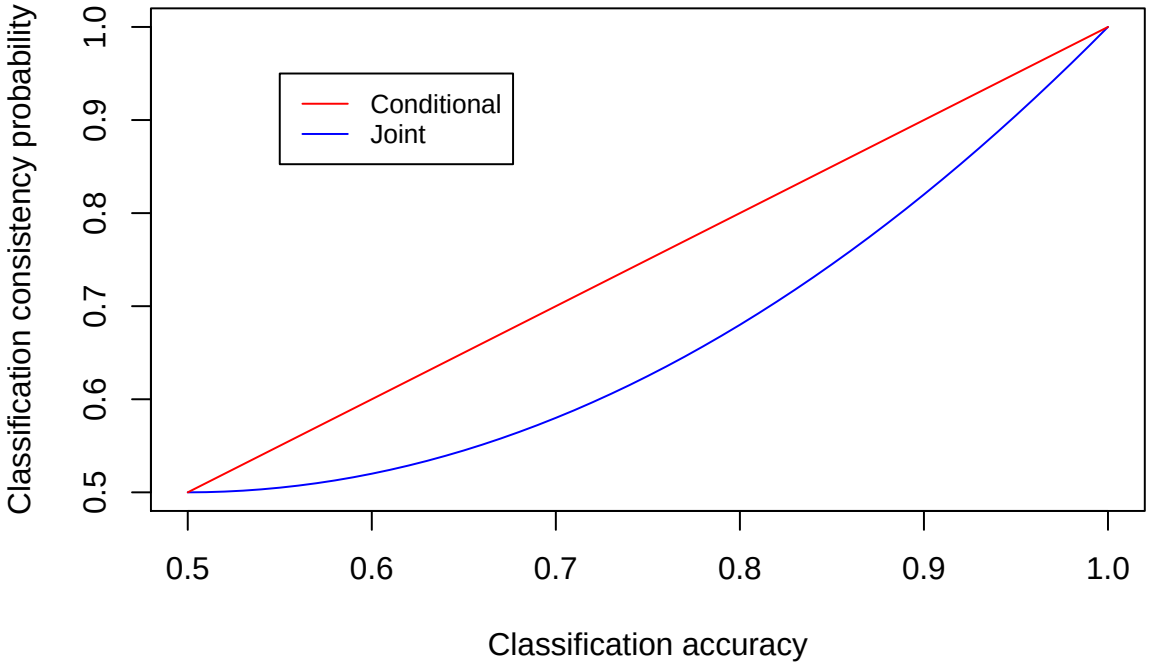

Figure 2.2. Relationship between conditional and joint classification consistency over a range of classification accuracy.

probabilities are similar, so confusing the two would be of little consequence. However, at all intermediate levels of classification accuracy, the difference between the two is greater, reaching a maximum of 0.125 (or 12.5 percentage points) when classification accuracy is 0.75. This discrepancy could cause score users to underestimate the stability of individual examinee classifications if the correct interpretation is not clearly explained.

2.1.1.2 Which sources of error?

In addition to the probability issue, another way in which definitions and operationalizations of classification consistency differ is the sources of error accounted for. In the above conceptualization of classification consistency, only random measurement error (assumed to be normally distributed) is considered as a potential cause of inconsistent classification. Such an index is therefore purely theoretical and only speaks to the interchangeability of test forms or measurement occasions, and not necessarily to the actual consistency of classification for repeat test takers. Such an index could nevertheless be useful as a baseline against which to compare other estimates or for tracking changes in classification consistency related to the test's psychometric properties over time. In addition to random measurement error, the scores of repeat test takers are potentially also influenced by practice effects or even true change in

latent ability between measurement occasions. In effect, a test taker is a slightly different person each time they take the test, and the resulting scores are not truly independent in the sense of a "blank slate" test taker (one who retakes the test with no memory or lasting effect of the first test administration). Or as Haertel (2006) explains, single-administration estimates of classification consistency (or reliability more generally) "cannot reflect error arising from examinees' inconsistency over time" (p. 101). Depending on how classification consistency is operationalized for a given testing program, test score users might require additional information and advice in order to accurately interpret the likelihood of classification change over multiple test attempts.

2.1.2 Extant Classification Consistency Indices

Given that the information needed to best estimate classification consistency—multiple independent test scores for each test taker—is not available, there are multiple methods to estimate classification consistency that each make certain assumptions and compromises. These methods date back as early as the 1970s (Haertel, 2006; e.g., Livingston, 1972) corresponding to the advent of criterion-referenced tests (Sawaki, 2016) and can be classified into CTT–based and IRT–based approaches. The vast majority of methods in both categories are so-called "single administration" methods, meaning that only one test score per examinee is required. Discussed below are the classification consistency methods examined in the present study, all but one of which are single administration methods.

2.1.2.1 CTT–based approaches

Classical true-score theory (aka, classical test theory; CTT) is a measurement paradigm that posits examinee estimated scores as the sum of examinee true scores and error scores (aka, error of measurement; Allen & Yen, 1979). The error scores are assumed to be normally distributed with a mean of 0 and uncorrelated with true scores or error scores of another test. Based on these assumptions, the concept of parallel tests is defined as two (or more) tests on which examinees have equivalent true scores and for which the error score distributions are identical. On the basis of these assumptions, there are several methods for estimating classification consistency.

2.1.2.1.1 Test–retest. As Lee (2010) asserts, the test–retest approach to estimating classification consistency is perhaps the most straightforward. When a sufficient number of examinees attempt the exam more than once, scores from repeat test takers can be used to calculate the agreement coefficient (p_o), defined as:

$$p_o \equiv \frac{|\{i|j_i = k_i\}|}{N} \tag{2.1}$$

where $|\{i|j_i = k_i\}|$ signifies the cardinality of the set of examinees i such that the classification on measurement occasion j is equivalent to that resulting from measurement occasion k, and N represents the total number of examinees. This notation is an alternative to that presented in Subkoviak (1988) and, furthermore, constitutes a generalization from the two-category scenario considered by Subkoviak to any number of categories. This agreement coefficient can also be used to calculate the kappa coefficient (Cohen, 1960), which corrects for the possibility of consistent classification by chance using the marginal probabilities of each category (Deng & Hambleton, 2013).

While the test–retest approach is, as Lee (2010) asserts, conceptually straightforward, a covert issue is the representativeness of the repeat test-taker sample. Classification consistency is, in a sense, the reliability of a test's score-based categorizations, and reliability is by nature sample-dependent (Haertel, 2006). A test–retest classification consistency estimation is therefore only representative of the entire examinee population if the sample of repeat test takers is also representative, a criterion that is likely not met in practice. In the case of certification and licensure testing, for which the binary pass/fail outcome is of sole importance, there is no apparent reason for a passing examinee to retake the test. The repeat test-taker population would therefore constitute only individuals who failed their first test attempts, which would certainly not resemble the general test-taker population. In the case of admissions testing, for which there is not a single universally applicable cut score and higher scores are believed to confer a competitive advantage, it is more plausible that some examinees at all score levels might retake the test. Nevertheless, the rate of repeat testing likely varies along the ability continuum, again resulting in a distinct repeat test-taker population. Any classification consistency method that relies on data from self-selected repeat test takers needs to account for this, either in the calculation of classification consistency or its interpretation.

2.1.2.1.2 Split-half. Just as split-half correlations can be used as a reliability index for continuous test scores, a split-half approach can be used to evaluate classification consistency from scores on a single test administration. In this approach, all test items are split into two equal groups and a score computed for each group of items. Haertel (2006) describes two such methods: Woodruff & Sawyer (1989) and Breyer & Lewis (1994). These methods assume that estimated and true split-half test scores follow a bivariate normal distribution and use the Spearman-Brown formula to "step up" correlations between the split-half test scores, which are then used in conjunction with a bivariate normal table to derive an estimated agreement coefficient on full-length tests. Both approaches also make assumptions about the cut scores on the half-length tests, as these methods were developed for dichotomously or polytomously scored items resulting in a number-correct total score. Crucially, both approaches also

assume the test halves are parallel, an assumption that might be difficult to satisfy in practice, especially if content balancing is desired.

An earlier method by Subkoviak (1976) obviates the aforementioned assumptions about half-test cut scores and parallelism by employing a double-length test and calculating the average classification consistency over all possible split halves. However, the use of a double-length test precludes the possibility of using operational test data, limiting its utility. Marshall & Haertel (1975) proposed a classification consistency index—coefficient beta (β), a counterpart to coefficient alpha—based on all possible split halves, but this method also does not seem to have received much attention in the literature. With the more recent advent of easily implemented item response theory (IRT) modeling, the parallelism and cut scores of test halves is less of a concern, and perhaps it is time to revisit the split-half approach to classification consistency with the added benefit of IRT scoring.

2.1.2.1.3 Livingston–Lewis. The Livingston & Lewis (1995) procedure is arguably at present the most popular CTT–based classification consistency method (Deng & Hambleton, 2013), as evidenced by a large number of citations, even in recent years, as well as several software options available for implementing it—the computer program BB-CLASS (Brennan, 2004) and the R package betafunctions (Haakstad, 2020). This method is applicable to tests comprising any item types using any scoring scheme, not just equally-weighted dichotomously-scored items. It achieves this flexibility by calculating an "effective test length," a number of dichotomously-scored items that would produce a reliability equal to that of the actual test. It then uses the estimated score distribution and assumed true score distribution (approximated by the 2- or 4-parameter beta distribution, depending on software) to generate an "alternate form" estimated score distribution for comparison. Measurement errors are assumed to follow the binomial or beta distribution, again depending on the software. Beta-distributed errors are continuous and do not require rounding the effective test length (Haakstad, 2020). Thus, all that is required for this method are a test reliability estimate, the raw score distribution, cut score(s), and quadrature settings.

2.1.2.2 IRT–based approaches

Item response theory (IRT) is a measurement paradigm that assumes test responses reflect a latent trait of the test taker and derives scores from a model that accounts for characteristics (principally difficulty) of the test items (Embretson & Reise, 2000). Two aspects of IRT crucial for present purposes are the notion of item-free person measurement (a.k.a., person-parameter invariance) and the conceptualization of ability estimate precision. The former implies that a given ability can be estimated with any items that have been calibrated on the same latent ability/difficulty scale, although

the precision of the estimate depends on the number and difficulty of the items used (De Ayala, 2013). The conceptualization of measurement precision within the IRT framework is based on the concept of the item/test information curve, with the test information curve being the sum of the individual item information curves, and an individual item's information curve being a function of the item parameters and ability. The approximate standard error of estimate (SEE) at a given ability value is then the inverse of the test information at that value. An important implication is that the precision of an ability estimate depends on the number of items administered, their item parameters, and the ability of the examinee. While multiple IRT–based methods for quantifying classification consistency exist, only one is presented here in order to serve as a comparison for the CTT–based approaches that are the focus of the present study.

2.1.2.2.1 Rudner (2000, 2005). While other IRT methods were proposed before and have been developed since, Rudner's method is a popular and conceptually straightforward IRT–based approach to classification consistency. It uses IRT ability estimates and SEEs to compute the probability of misclassification for each score point on the test's scale, assuming normally distributed measurement error. Thus, individual misclassification probabilities can be calculated and plotted to show change in misclassification probability across the score scale or aggregated to compute an overall classification consistency. The method is implemented in software by the author and in the R package cacIRT (Lathrop, 2015). While the focus of the present study is on CTT–based methods, this method is included for comparison purposes.

2.1.3 Prior research comparing classification consistency methods

Deng & Hambleton (2013) estimated that, at the time of writing, there had been "limited comparative studies" (p. 235) of classification consistency methods. One of these few studies is Wan, Brennan, & Lee (2007), which used simulated and real data to compare five non-IRT approaches capable of handling data from tests with mixtures of dichotomous and polytomous items. These approaches are a normal approximation procedure, the Breyer-Lewis procedure, the Livingston-Lewis procedure, a bootstrap procedure, and a compound multinomial procedure. The authors found that the normal approximation and Livingston-Lewis methods were the most accurate.

Deng & Hambleton (2013) compared the Livingston–Lewis procedure with a common IRT-based procedure Lee (2010) under different conditions of test length, examinee ability distribution skewness, and local item dependence. The authors found that both classification consistency methods exhibit relatively small negative bias (i.e., underestimation of the true classification consistency) of between –0.01 and –0.04

across test lengths, with the Lee method proving more robust to shorter tests. Similar results were shown for the different examinee ability distribution conditions, with the Lee method slightly but consistently outperforming the Livingston–Lewis method. The bias of the Livingston–Lewis method was exacerbated when a short test was combined with a negatively skewed ability distribution. However, the Livingston–Lewis method proved more robust to changes in the degree of local item dependence, although the bias estimates of both methods again remained relatively small under all simulation conditions. The authors conclude that the Lee method overall performs somewhat better than the Livingston–Lewis method, as long as the assumptions of IRT models are met and the model fit is acceptable.

2.1.4 Duolingo English Test cut scores

The present study is motivated by a gap in the classification consistency literature exemplified by the Duolingo English Test, a computer adaptive test (CAT) of general English proficiency currently accepted by over 3,000 institutions of higher education for admissions purposes (Duolingo English Test, 2020). The test comprises five objectively scored item types and multiple constructed response items in both the speaking and writing modalities; objectively scored items are administered adaptively from a very large item bank based on an algorithm that maximizes precision of ability estimates (Cardwell, LaFlair, & Settles, 2021). An unreported score is calculated for each examinee on each item type, using an IRT–like procedure for the objectively scored item types and machine learning– and natural language processing–based automated scoring algorithms for the constructed response speaking and writing items; the reported total score (range: 10–160, in increments of 5) is then derived as a linear combination of the separate item type scores (Cardwell, LaFlair, & Settles, 2021). Furthermore, the difficulties of the objectively scored items are determined by CEFR–based machine learning and natural language processing models (Settles, LaFlair, & Hagiwara, 2020). These characteristics of the Duolingo English Test deviate from more traditional CATs and complicate calculation of classification consistency estimates.

The nature of CATs presents challenges to applying existing classification consistency indices. All mainstream classification consistency indices have been developed, either explicitly or implicitly, for fixed-form tests and have not been definitively endorsed for use on CATs. The extent to which they are appropriate for a CAT context is therefore an empirical question that has not been sufficiently addressed. In the case of CTT–based approaches, operational data might violate distributional assumptions, and there is also potential concern about the representativeness of repeat test takers whose data would be used for a test–retest approach. Furthermore, existing IRT–based approaches rely on empirical operational item parameters or model-based standard

errors of examinee ability estimates, neither of which is available for the Duolingo English Test due to the innovative nature of its development. And both CTT– and IRT–based approaches assume some degree of parallelism between test forms used in hypothetical replications. Before presenting the research questions and methods, I will first define the cut scores of interest for the analyses. There are two broad categories of cut scores relevant to the Duolingo English Test—CEFR level cut scores and admissions cut scores.

2.1.4.1 CEFR level cut scores

Given that the Duolingo English Test scale is aligned to the CEFR (Settles, LaFlair, & Hagiwara, 2020), the most fundamental classification applied to all scores is that of a CEFR level (e.g., B2). Each CEFR level is defined by a lower and upper cut score on the Duolingo English Test scale (Table 1). Thus we would like to know the classification consistency given these cut scores, i.e., at what rate will test takers fall into the same CEFR level over multiple administrations (assuming their language ability has not actually changed). Classification consistency can be evaluated for a single level, such as the rate of being consistently classified as B2. It is not necessarily the case that classification consistency is equal across levels, as differences between test takers of different abilities in the number of items responded to within the test time limit or differences in item information at different levels of item difficulty could lead to variation in score stability along the score scale. In addition to classification consistency of specific levels, an overall classification consistency can be calculated that indicates the general rate of consistent classification. The present study, however, focuses on single cut scores for admissions purposes.

Table 1: Duolingo English Test reported alignment to the CEFR

Duolingo	CEFR
10–20	A1
25–55	A2
60–85	B1
90–115	B2
120–140	C1
145–160	C2

2.1.4.2 Admissions cut scores

In addition to CEFR levels, Duolingo English Test scores are also used by academic institutions to make admissions decisions. In this case, the cut score represents a categorization of test takers based on whether or not they meet the institution's

English language proficiency criteria. Unlike the CEFR level cut scores, admissions cut scores vary by institution, reflecting various factors such as the linguistic demands of the course of study and available support resources for students at the institution (e.g., remedial English courses, writing centers, etc.). Indeed, many institutions have different cut scores for different internal divisions (e.g., school of journalism), majors, and/or level of study. Additionally, some schools also have multiple cut scores for different admissions categories, such as unconditionally admitted and admitted with a requirement to take remedial courses. Figure 2.3 summarizes admissions cut scores used at 269 accepting institutions for which such information was publicly available as of June 2020.

As depicted in the figure, the vast majority of institutional admissions cut scores fall within the B2 CEFR level, and thus the B2/B1 cut score (90) will be an important one to consider in analyses. Additionally, the mean and median admissions cut score (both 105) is another important one to investigate. However, it should be noted that the Duolingo English Test is also used for admission and placement into intensive English language programs (not depicted in Figure 2.3), which are more concerned with ability levels below B2. The Duolingo English Test is also increasingly used for admissions purposes at the graduate level, which could require language proficiency closer to the C1 level. Thus, for the Duolingo English Test, there is no single cut score of primary importance, and the test has been designed to estimate various ability levels with comparable accuracy to accommodate varied use cases.

2.1.4.2.1 A note on cut scores, estimated scores, and true scores. It should be noted that, by definition, cut scores are set on estimated scores. But the true score is actually of interest. So while stakeholders might not possess the conceptual background to articulate their use of cut scores in terms of true and estimated scores, they could implicitly understand the distinction. Anecdotally, some institutions set "conservative" (i.e., higher) cut scores, presumably motivated at least partially by a desire to minimize misclassifications due to measurement error. So while the CEFR B2 level is widely considered and used as the level of language proficiency necessary for undertaking higher education (Deygers, Zeidler, Vilcu, & Carlsen, 2018)—which suggests a desired DET true score of 90 or greater—institutions may use cut scores of 100 or 105 to "make sure" that admitted applicants really have a B2 level of English (i.e., that their true scores are >= 90). Brown & Hudson (2002) refer to this conservative approach to setting a cut score as "the protecting-the-institution strategy." The implication for classification consistency analyses is that, if we assume that a cut score reflects the institution's desired true score, we may be underestimating the consistency of the classification that institutions actually care about—the B2 level.

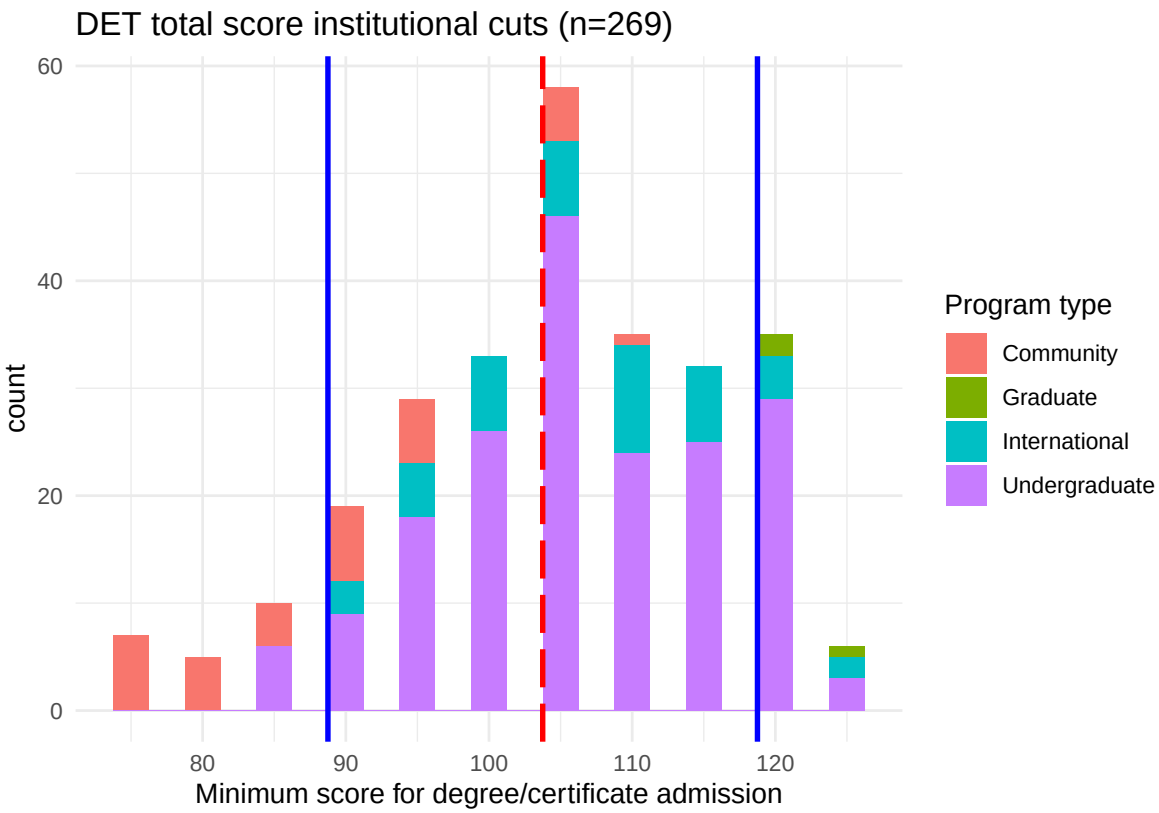

Figure 2.3. Histogram of DET minimum scores (on total score scale) for admissions purposes as of June 2020.
Notes: 1) Blue lines represent B1/B2 cut (90) and B2/C1 cut (120). Red line represents the mean/median institutional cut score (105). 2) The above plot uses a single cut score from each accepting institution for which such information is available. If an institution uses multiple cut scores (e.g., for different programs or levels of study), the minimum value among degree/certificate programs was used.

2.1.5 Research Questions

The present study is motivated by non-traditional CATs, of which the Duolingo English Test serves as an example. Several characteristics of the Duolingo English Test (ML/NLP–based item difficulty estimates, continuous item scores, and automatic essay and speech grading) make the IRT–based classification consistency methods typically applied to CATs impractical, thus necessitating consideration of CTT–based methods. However, the results are generalizable to other CAT contexts, even those based in IRT, with one important caveat: the present study does not consider the distribution of the item bank across the difficulty scale, which could be an important factor in contexts in which item availability at certain difficulty levels is limited.

The study uses a simulation to investigate the following research questions concerning classification consistency around a single pass/fail cut score:

1. How accurate are CTT–based classification consistency estimates (test–retest, split-half, and Livingston–Lewis) for CAT scores? (Accuracy is assessed by comparison to a Rudner-like approach using the known ability values instead of ability estimates)
2. Are the results from a CAT context consistent with prior studies in fixed-form contexts?
3. Does any method exhibit a consistent bias (i.e., over- or underestimating classification consistency across simulation conditions)?

In light of the answers to these questions, I will also discuss the appropriateness of CAT test developers reporting CTT–based classification consistency indices, as well as implications for test score users.

2.2 Methods

The research questions are addressed via a simulation study, described below in four parts: simulation conditions (manipulated factors), data generation, analysis of simulated scores, and analysis of simulation results.

2.2.1 Simulation Conditions

The following five factors are manipulated in the generation of data in order to observe their effects on the performance of the different classification consistency indices considered:

1. **Test length** – two different test lengths are considered, corresponding to short (20 items)and long (40 items) conditions.
2. **Item difficulty parameter precision** – three different levels of item difficulty parameter precision are considered: true, small error, and large error. In the "true" condition, the same item difficulties are used for generating examinee responses, selecting items during the CAT, and estimating $\hat{\theta}$. In the small and large error conditions, the "true" item difficulties are used to generate examinee responses, while a different set of difficulties are used to select items and calculate final $\hat{\theta}$. This second set of difficulties is derived as

$$b_2 = b_1 + e,\ e \sim N(0, \sigma),\ \sigma \in \{0.684, 1.233\} \tag{2.2}$$

   where b_1 represents the true difficulty and b_2 represents the imprecise difficulty estimate. The two values of σ correspond to difficulty parameter estimation reliabilities of approximately 0.95 and 0.80 respectively.

3. **Item selection criterion** – three different criteria are used for selecting the subsequent item during the CAT process: maximum Fisher information (MFI); the proportional method (Barrada, Olea, Ponsoda, & Abad, 2008), which randomly selects the next item with a probability equal to the Fisher information raised to a power that begins at 0 and increases over the duration of the test; and random item selection.
4. **Calibration models** – two different models are used for estimating $\hat{\theta}$: the 1PL and the 2PL.
5. **Location of cut score** – four different binary cut scores are considered when evaluating classification consistency methods: 1 SD below the population mean, 0.5 SD below the population mean, at the population mean, and 0.25 SD above the population mean. These cut scores were chosen to imitate institutional DET cut scores, the majority of which lie within the B2 CEFR band, corresponding to approximately from 1 SD below to 0.5 SD above the population mean test score.

2.2.2 Data Generation

Data are generated for 2,000 examinees per replication, with 250 replications, using the catR package (Magis & Barrada, 2017) in the statistical software R (R Core Team, 2020). Examinee abilities are sampled from a random subset of 4,000 real final ability estimates from the Duolingo English Test after they have been transformed to a [-4,4] scale[1] to more closely resemble the logit scale. Item difficulty parameters are likewise drawn from the transformed item difficulty estimates of a random sample of 4,000 Duolingo English Test items with sufficient response data for estimating item–total correlations. Figure 2.4 displays the distributions of examinee ability and item difficulty used in the simulation. Since the Duolingo English Test does not use item discrimination operationally, discrimination parameters are approximated for the purpose of the simulation based on the item–total correlation of each item, transformed to a range of [.5,1.5]. Figure 2.5 depicts the relationship between the difficulty and discrimination parameters for the items used in the simulation. The response matrix of all sampled examinees to all sampled items is generated twice from a 2PL model and used in all simulation conditions. Two CAT sessions are simulated for each examinee under each condition, with ability assumed to be constant. The starting item of each simulated CAT is chosen at random from the 50 items with the highest item information at $\theta = 0$.

[1]The Duolingo English Test uses a [0,10] scale for both item difficulties and examinee abilities.

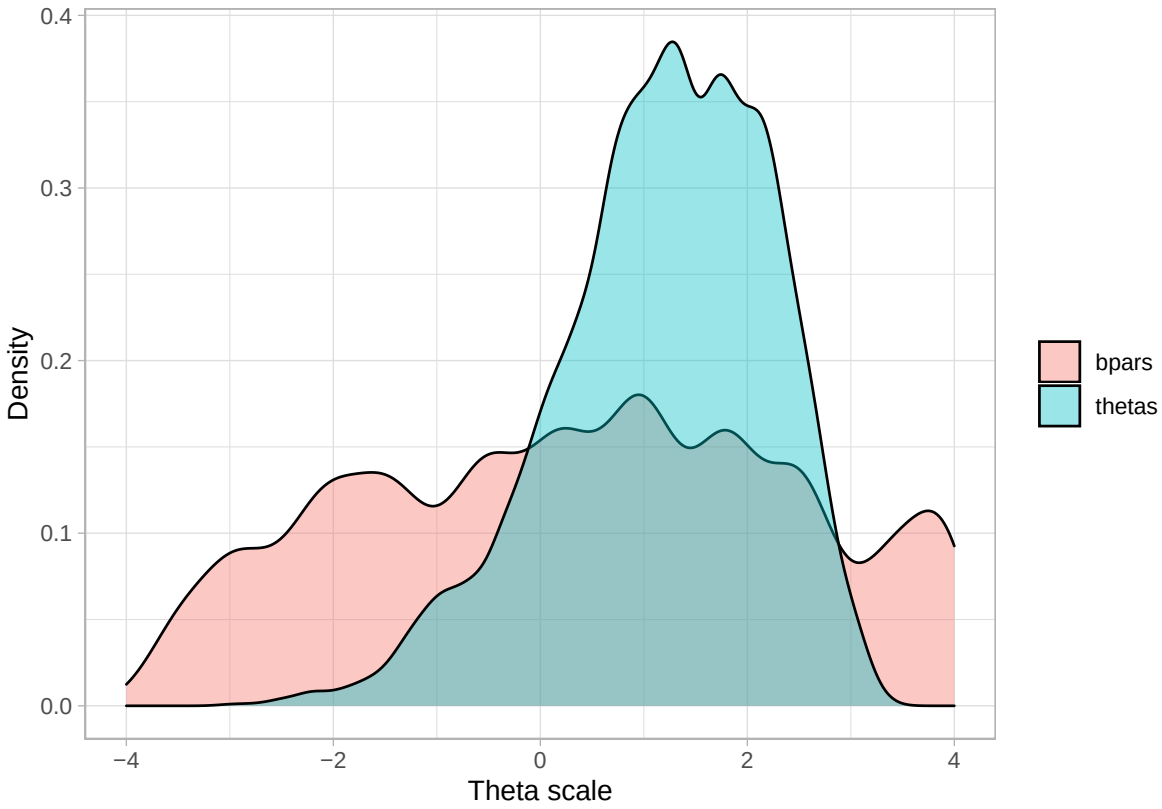

Figure 2.4. Densities of examinee abilities and item difficulty parameters used in the simulation

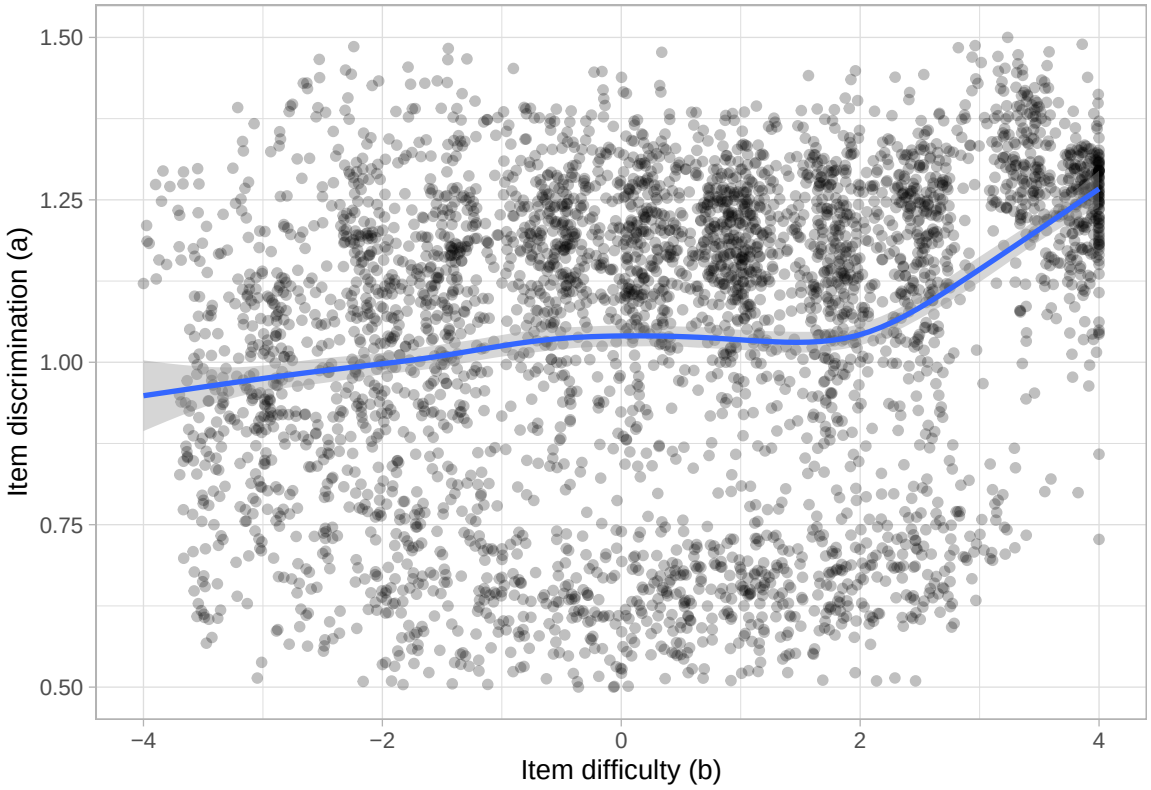

Figure 2.5. Scatterplot depicting the relationship between discrimination and difficulty parameters of the items used in the simulation study

2.2.3 Computing Classification Consistency Indices

The simulated final ability estimates are used to compute the following classification indices under each simulation condition:

2.2.3.1 Rudner (2000, 2005)

This is a popular IRT–based method included in this study for comparison purposes. Using the ability estimates from the first CAT administrations ($\hat{\theta}_{i1}$) and their standard errors ($SE(\hat{\theta}_{i1})$) for all examinees i, the classification consistency percent agreement estimate of the test is given by

$$p_o = \frac{1}{N}\sum_{i=1}^{N} P(\hat{\theta}_i > \theta_c|\hat{\theta}_{i1})^2 + P(\hat{\theta}_i < \theta_c|\hat{\theta}_{i1})^2 \tag{2.3}$$

where θ_c represents the cut score in question and normally distributed errors are assumed, such that the expected distribution of an examinee's score estimate is $N(\hat{\theta}_{i1}, SE(\hat{\theta}_{i1}))$. Additionally, this method could be compared to that of Luecht (2015).

2.2.3.2 Test–Retest

Using estimated scores from both CAT administrations, the test–retest percent agreement estimate is given by

$$p_o = \frac{|\{i|C_{i1} = C_{i2}\}|}{N} \tag{2.4}$$

where C_{i1} represents examinee classification based on the first administration and C_{i2} represents classification based on the second. This index is calculated both for all examinees and for examinees who "failed" the first administration, in order to demonstrate the effect of non-representative samples of repeat test takers.

2.2.3.3 Split-Half

This method entails dividing the items of each CAT administration based on administration order into odd and even splits, then re-estimating examinee abilities based on these split-half tests. The percent agreement classification consistency index is then calculated as

$$p_o = \frac{|\{i|C_{ih_1} = C_{ih_2}\}|}{N} \tag{2.5}$$

where C_{ih_1} represents examinee classification based on the first half-test score and C_{ih_2} represents classification based on the second. This raw percent agreement will also be adjusted for the reduced reliability of the shorter half-tests relative to the full-length test.

2.2.3.4 Livingston–Lewis (1995)

The Livingston–Lewis method is implemented by the R package betafunctions using estimated scores and two estimates of test reliability: test–retest and split-half.

In addition to the above classification consistency indices, classification accuracy and expected classification consistency is reported for comparison and calculation of bias. Expected classification consistency is calculated similarly to the Rudner method, but using examinee true θ and the average $SE(\hat{\theta})$ of the two simulated CATs.

2.2.4 Analysis of Classification Consistency Indices

For each of the above indices, an average is computed within each simulation condition. Furthermore, accuracy is quantified by RMSE, and bias ise quantified by the difference between the mean parameter estimate and the true parameter value.

2.3 Results

2.3.1 Individual Factors

The tables below summarize the marginal results (mean classification consistency estimate and RMSE) for each simulation factor: test length (Table 2.2), item parameter estimate precision (Table 2.3), item selection method (Table 2.4), calibration model (Table 2.5), and cut score location (Table 2.6).

Table 2.2 illustrates that both expected classification consistency and estimates from all methods increase when the test is lengthened, consistent with the expectation that a longer test will be more reliable, and a more reliable test will produce more consistent classifications. Additionally, all RMSE values decrease when the test is lengthened, indicating a more reliable test is also associated with more accurate estimates of classification consistency. The TRT-all and Livingston–Lewis methods are somewhat more robust to shorter test length as evidenced by the comparatively low RMSE values in the short test length condition.

Table 2.3 illustrates that random error in item difficulty parameter estimates causes lower expected classification consistency, although interestingly larger parameter estimate imprecision does not lead to lower expected classification consistency. RMSE values, however, monotonically increased for all methods from the no-error condition to the large error condition. This signifies that, even if greater parameter estimate imprecision does not substantially impact the test's actual classification consistency, it can introduce error into *estimates* of classification consistency.

Table 2.2. Mean classification consistency estimate and RMSE for each classification consistency method by test length

Test length	Value	Expected	Rudner	TRT (all)	TRT (fail)	SH (raw)	SH (adj)	LL
20	Mean Value	0.841	0.833	0.823	0.738	0.766	0.798	0.812
	RMSE	—	0.017	0.032	0.147	0.083	0.061	0.036
40	Mean Value	0.874	0.869	0.873	0.811	0.825	0.863	0.864
	RMSE	—	0.014	0.019	0.099	0.056	0.024	0.021

Table 2.3. Mean classification consistency estimate and RMSE for each classification consistency method by item difficulty parameter estimate precision

Parameter estimate error	Value	Expected	Rudner	TRT (all)	TRT (fail)	SH (raw)	SH (adj)	LL
None	Mean Value	0.867	0.865	0.863	0.797	0.815	0.851	0.856
	RMSE	—	0.004	0.013	0.106	0.057	0.027	0.016
Small	Mean Value	0.85	0.845	0.853	0.781	0.790	0.826	0.844
	RMSE	—	0.010	0.015	0.112	0.066	0.038	0.016
Large	Mean Value	0.856	0.842	0.828	0.745	0.780	0.813	0.813
	RMSE	—	0.024	0.041	0.154	0.086	0.066	0.046

Table 2.4 shows that expected classification consistency is slightly (~3%) lower under random item selection. Otherwise, there is not a clear pattern in the results, as some methods demonstrate lower RMSE under random item selection (e.g., TRT-all) while others demonstrate lower RMSE under proportional (SH-adj) or MFI (TRT-fail) item selection criteria.

Table 2.5 shows that the expected classification consistency when using a 2PL model to estimate ability and select items is slightly greater than in the 1PL conditions. Additionally, the split-half methods show higher RMSE values under the 2PL conditions, whereas other methods have more comparable RMSE values under both 1PL and 2PL conditions.

Table 2.6 illustrates that expected classification consistency decreases as the pass/fail cut score gets closer to the test-taker population mean (i.e., $Z = 0$). This result is consistent with expectations that the frequency of classification errors would increase when the cut score is located at a point of higher density in the examinee score distribution. The RMSE values of the methods TRT-all, SH-raw, and SH-adj increase (i.e., accuracy decreases) as the cut score moves closer to the population mean, whereas

Table 2.4. Mean classification consistency estimate and RMSE for each classification consistency method by item selection criterion

Item selection criterion	Value	Expected	Rudner	TRT (all)	TRT (fail)	SH (raw)	SH (adj)	LL
Random	Mean Value	0.836	0.831	0.826	0.744	0.773	0.806	0.814
	RMSE	—	0.007	0.021	0.136	0.069	0.046	0.028
Proportional	Mean Value	0.869	0.863	0.856	0.786	0.809	0.847	0.847
	RMSE	—	0.014	0.030	0.123	0.066	0.034	0.033
Mfi	Mean Value	0.868	0.859	0.863	0.792	0.804	0.838	0.851
	RMSE	—	0.021	0.027	0.117	0.076	0.057	0.028

Table 2.5. Mean classification consistency estimate and RMSE for each classification consistency method by calibration model

Calibration model	Value	Expected	Rudner	TRT (all)	TRT (fail)	SH (raw)	SH (adj)	LL
1PL	Mean Value	0.851	0.848	0.836	0.761	0.792	0.827	0.830
	RMSE	—	0.008	0.029	0.132	0.064	0.036	0.031
2PL	Mean Value	0.864	0.854	0.860	0.787	0.799	0.833	0.845
	RMSE	—	0.020	0.023	0.119	0.076	0.055	0.029

the RMSE of TRT-fail substantially decreases and that of Livingston–Lewis is relatively stable. Thus, when averaging across all other simulation factors, Livingston–Lewis (the only distributional method in the study) seems to be the most robust to cut-score location, whereas TRT-fail is most sensitive. It is also worth noting that by definition, the sample used to estimate classification consistency with the TRT-fail method is different for each cut score, contributing to larger differences in estimate accuracy.

2.3.2 Combined Factors

Figures 2.6 and 2.7 below displays the results of certain specific combinations of simulation conditions. Given the large number of simulation conditions ($2^2 * 3^2 * 4 = 144$), it is not practical to compare all conditions, and so the figures have been chosen to highlight results of potential interest. In all cells of both plots, the horizontal blue line represents the expected classification consistency of that condition. Figure 2.6 visually summarizes the classification consistency estimates of all conditions using the 1PL model for ability estimation and item selection and with no error in the item difficulty parameters. Noteworthy results are that expected classification

Table 2.6. Mean classification consistency estimate and RMSE for each classification consistency method by cut score location

Cut score	Value	Expected	Rudner	TRT (all)	TRT (fail)	SH (raw)	SH (adj)	LL
-1.00	Mean Value	0.906	0.900	0.908	0.700	0.875	0.905	0.885
	RMSE	—	0.010	0.017	0.215	0.035	0.013	0.026
-0.50	Mean Value	0.862	0.852	0.852	0.745	0.799	0.836	0.837
	RMSE	—	0.018	0.025	0.123	0.068	0.038	0.033
0.00	Mean Value	0.833	0.825	0.817	0.810	0.753	0.789	0.813
	RMSE	—	0.018	0.031	0.033	0.086	0.062	0.032
0.25	Mean Value	0.83	0.826	0.815	0.842	0.755	0.791	0.816
	RMSE	—	0.014	0.029	0.022	0.081	0.057	0.028

consistency, and estimates thereof, is/are uniformly higher for both longer tests and for non-random item selection approaches, consistent with the marginal results in the previous section. Additionally, the TRF-fail method (i.e., test–retest using only examinees who scored below the cut score on their first test attempt) greatly underestimates classification consistency at cut scores below the mean, while consistently over-estimating classification consistency at the cut score above the mean under the longer (40 question) test condition.

Figure 2.7 visually summarizes the classification consistency estimates of all conditions in which the test is 40 questions long and the cut score is at the population mean ability. Noteworthy results are that small error in item difficulty parameter estimates ($\rho_{b\hat{b}} \approx 0.95$) has a negligible effect on classification estimate accuracy, whereas large error ($\rho_{b\hat{b}} \approx 0.80$) has a more pronounced effect. There also appears to be an interaction between calibration model and difficulty parameter error, with the ranking of classification consistency methods by accuracy when there is large error depending on whether the 1PL or 2PL model is used. The split-half methods stand out as being particularly impacted by large parameter error in the 2PL conditions. These simulation conditions are also the only ones in which the Rudner method is not the most accurate.

2.3.3 Test Overlap Rate

One factor that could impact the interpretation of the results is the test overlap rate, or the average percentage of items shared between an examinee's first and second test administration. Figure 2.8 plots the overlap rates of all 36 simulated CAT conditions (cut score location is not relevant to test overlap). Test overlap is negligible for all conditions using random assignment and relatively low (<10%) for all conditions using

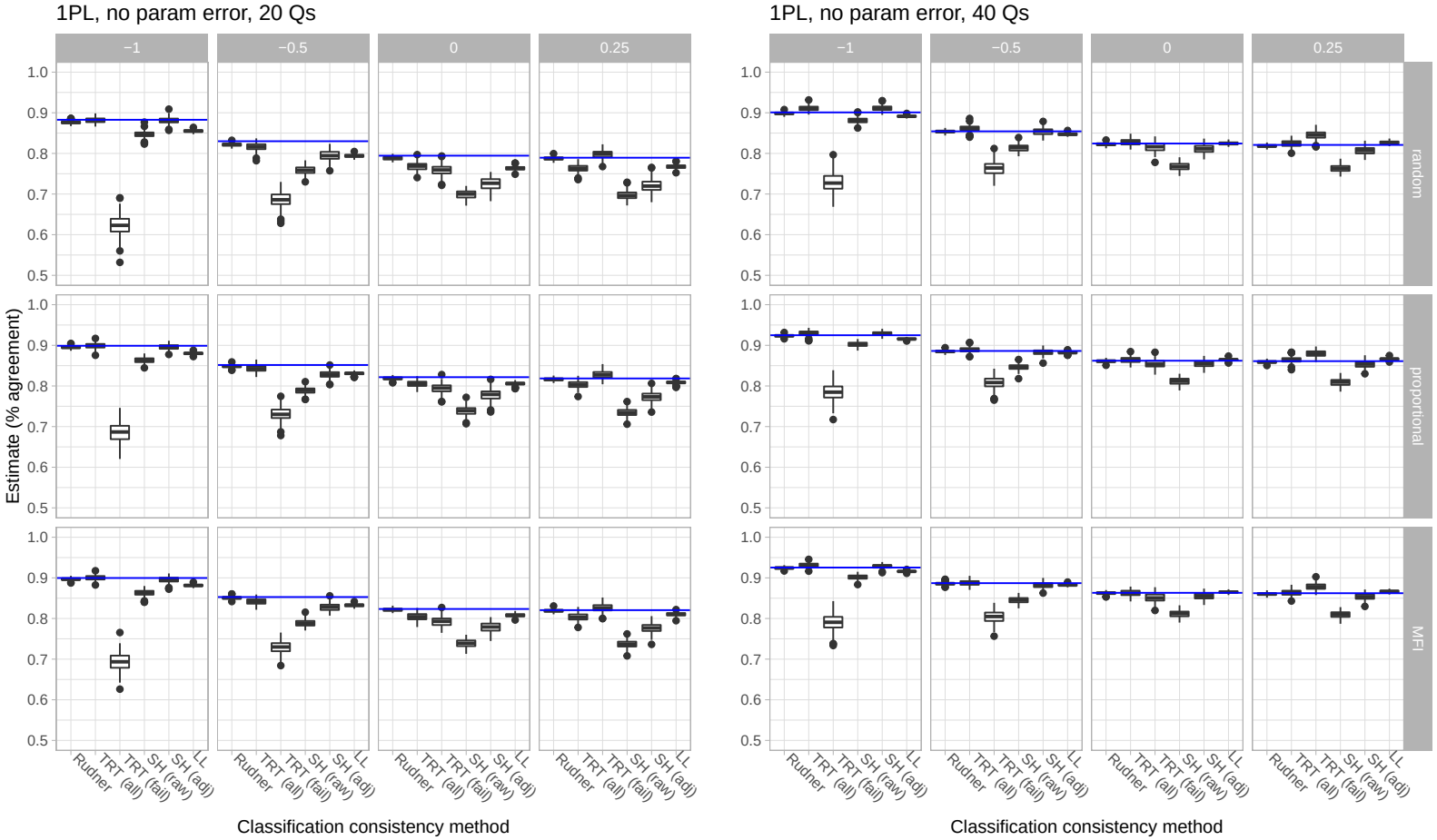

Figure 2.6. Results (% agreement estimates) for all 1PL simulation conditions.

Note: Blue lines represent expected classification consistency.

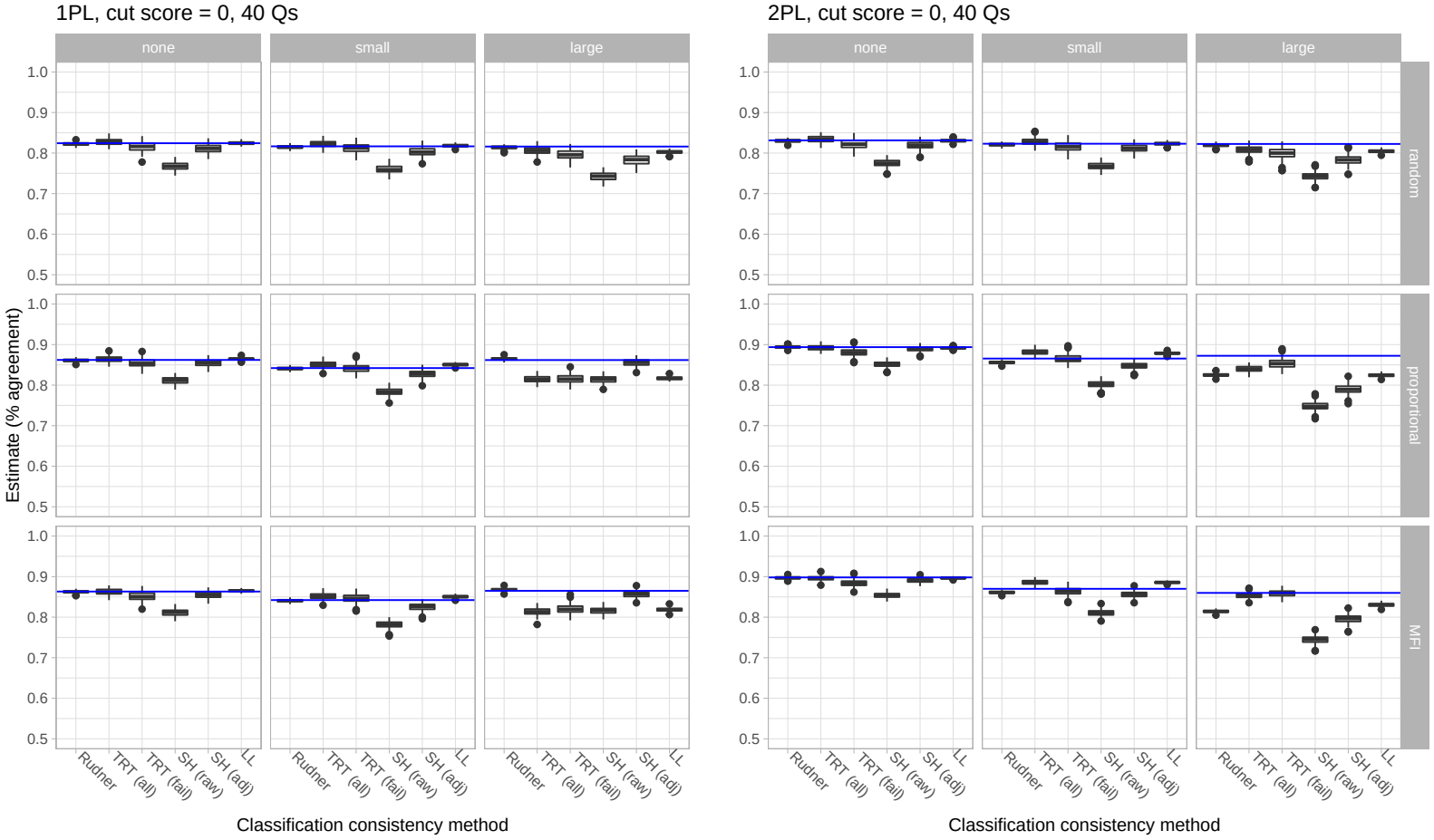

Figure 2.7. Results (% agreement estimates) for all simulation conditions with 40 questions and cut score of $Z = 0$.

Note: Blue lines represent expected classification consistency.

the 1PL. However, in conditions using both the 2PL and MFI item selection, test overlap can exceed 50%. It should be noted that two complete response matrices were independently generated for the simulation study, such that an item occurring on both an examinee's first and second test administration will not necessarily receive the same score each time. In other words, there are no practice or carryover effects, making reoccurring items functionally equivalent to distinct items with identical item parameters. This mitigates the high overlap rates in certain conditions. Nevertheless, repeat test administrations in high-overlap conditions potentially have more similar test information curves due to more similar item parameters, which could impact classification consistency.

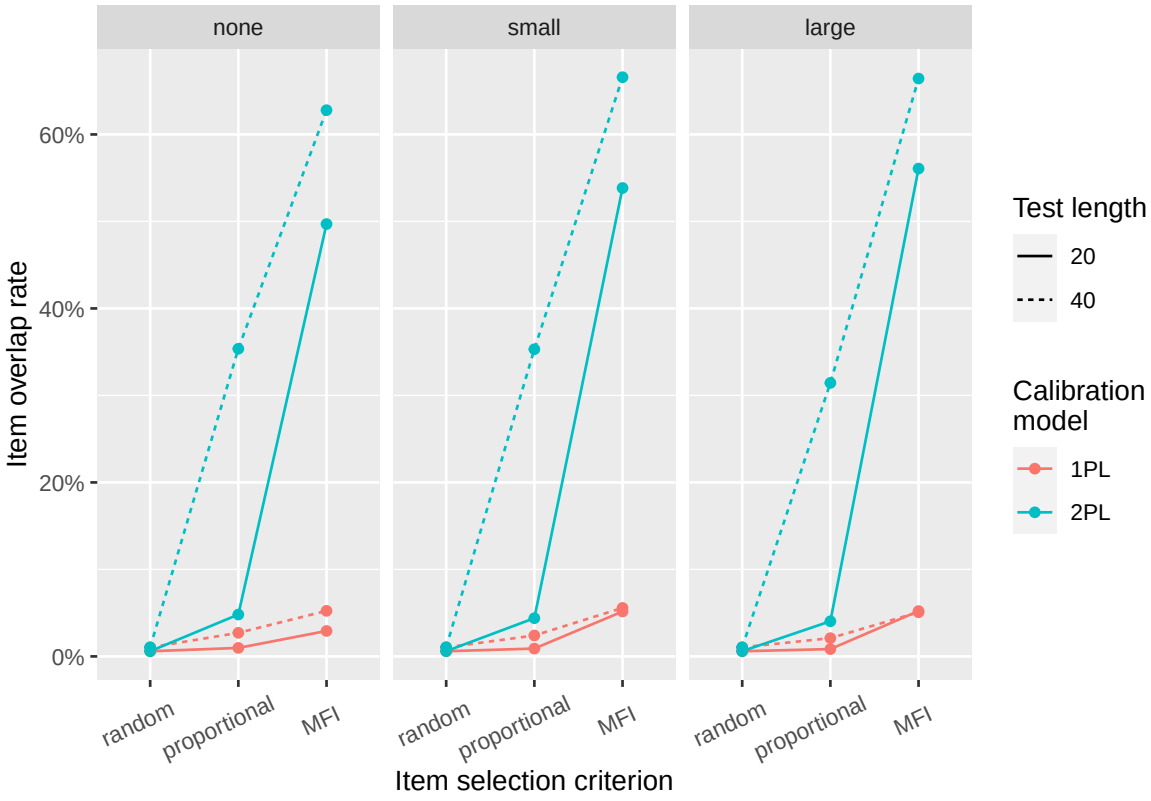

Figure 2.8. Test overlap rates of all 36 simulated CAT conditions

2.4 Discussion

The results show broadly that CTT–based methods for estimating classification are not necessarily inappropriate for a CAT context. The methods considered exhibited sensitivity to different simulation factors. The adjusted split-half approach (adjusting the observed classification consistency based on two half-test scores for reduced

reliability) exhibited a low RMSE (<.05) across most simulation conditions. However, this method was sensitive to item difficulty parameter error when using the 2PL model for calibration. Thus a split-half approach might be particularly appropriate when test developers have high confidence in the accuracy of item parameters or discrimination parameters are not used operationally. Furthermore, split-half methods are only feasible when one has access to split-half scores or the ability to rescore parallel halves of administered tests.

The test–retest approach based on all test takers (TRT-all) also exhibited low bias, even outperforming the gold-standard Rudner method in certain conditions with large parameter error. However, when the test–retest method was applied only to "failing" examinees (TRT-fail), classification consistency was greatly underestimated for cut scores below the population mean, and slightly overestimated for cut scores above the mean. This result underscores the need to carefully and critically consider the samples used in analyses. For many testing programs, it is likely impossible to obtain independent repeat test administrations for the entire examinee population, or even a representative sample thereof. Repeat examinees as a group are therefore likely not representative of the overall examinee population. The test–retest approach to estimating classification consistency is therefore potentially only appropriate when at least some examinees at all ability levels retake the test. Even in such a case, statistical adjustment to account for differences between the general and repeat examinee populations would likely be beneficial.

Other than the TRT-all approach, the Livingston–Lewis method was the only one to demonstrate RMSE<0.05 in every simulation factor when considered independently (i.e., averaging over the conditions of the other simulation factors), suggesting general robustness. However, it is important to note that this method requires an estimate of test reliability as input, and thus could be less accurate if there is appreciable error in the reliability estimate. Thus, while the Livingston–Lewis method is demonstrably appropriate for application to a CAT, it is potentially only appropriate when there is high confidence in the test's reliability estimate, a condition that is potentially difficult to satisfy in cases where IRT-based classification consistency estimates are unfeasible.

Collectively, the results suggest that test developers should be able to use CTT–based methods for estimating the classification consistency of a CAT when IRT–based estimates are unfeasible. Specifically, test–retest using a representative sample, adjusted split-half, and Livingston–Lewis approaches appear to produce reasonably accurate estimates in most simulation conditions and are straightforward to explain to stakeholders. However, each method has its own sensitivities and drawbacks, and thus no single method is universally preferable. CAT developers must consider the circumstances of their testing program, characteristics of the examinee population, and availability of information in deciding on the most appropriate CTT–based

classification consistency method for the given context.

Chapter 3

Paper 2: Estimating Classification Consistency of an Innovative CAT Assessment

3.1 Introduction

When test scores are used primarily to categorize examinees, the consistency of these classifications is arguably more important than the reliability of the continuous scores (Deng & Hambleton, 2013). The Standards for Educational and Psychological Testing (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014) state that "when a test... is used to make classification decisions, estimates should be provided of the percentage of test takers who would be classified in the same way on two replications of the procedure" (p. 46). While the Standards recommend a test–retest approach to estimating classification consistency, it is rare for all examinees (or even a representative sample thereof) to take the same test more than once. Therefore, numerous methods have been proposed for estimating classification consistency from responses to a single test administration. The present study applies several of these methods to the Duolingo English Test, an innovative English proficiency test that presents several challenges to estimating classification consistency.

3.1.1 The Duolingo English Test

The present study is motivated by a gap in the literature concerning the estimation of classification consistency for complex adaptive assessments, of which the Duolingo English Test is used as an example. It is a computer adaptive test (CAT) of general

English proficiency currently accepted by over 3,000 institutions of higher education for admissions purposes (Duolingo English Test, 2020). The test comprises five objectively scored item types and multiple constructed response items in both the speaking and writing modalities (Cardwell, LaFlair, & Settles, 2021). An unreported score is calculated for each examinee on each item type, using an IRT–like procedure for the objectively scored item types and machine learning– and natural language processing–based automated scoring algorithms for the constructed response speaking and writing items; the reported total score is then derived as a linear combination of the separate item type scores (Cardwell, LaFlair, & Settles, 2021). Furthermore, item difficulties of the objectively scored items are determined by CEFR–based machine learning and natural language processing models (Settles, LaFlair, & Hagiwara, 2020). All of these characteristics of the Duolingo English Test complicate calculation of classification consistency estimates. Since it is a CAT and therefore not fixed-form, CTT–based methods are not prima facie appropriate. However, since the test also does not utilize IRT in a traditional sense, extant IRT-based methods are also not necessarily applicable without modification.

Accepting institutions set their own cut scores on the Duolingo English Test's 10–160 reported score scale. Such cut scores define binary classifications based on test scores. The Duolingo English Test also provides score ranges corresponding to the six levels of the Common European Framework of Reference for Languages (CEFR; Duolingo English Test, 2021), constituting five cut scores defining a six-category classification. Furthermore, the Duolingo English Test provides score percentiles for the total score and each of the four reported subscores (Cardwell, LaFlair, & Settles, 2021). Thus, the Duolingo English Test has a mixture of characteristics traditionally associated with both norm- and criterion-referenced tests.

3.1.2 Norm- and Criterion-Referenced Tests

The topic of classification consistency is related to the distinction made in the measurement field between norm-referenced and criterion-referenced tests, in that classification consistency is a common approach to quantifying the dependability of criterion-referenced assessments (Sawaki, 2016). Norm-referenced tests are regarded as those from which scores are used for rank-ordering examinees for purposes such as selection or admission; criterion-referenced tests, on the other hand, are intended to describe an examinee's mastery of a specific domain of knowledge and skills, or to assign an examinee to performance levels, independent of the performance of any other examinees (Sawaki, 2016). It is worth noting that these definitions refer to test score use. While in common parlance, the descriptors "norm-referenced" and "criterion-referenced" are often applied to the tests themselves, these terms, like the concept of validity, are more accurately applied to score uses, of which a test can have several. For example, SAT

and ACT scores are used to classify examinees against college readiness benchmarks, but they are also commonly used to select scholarship and award recipients.

It is also the case that tests can be designed with norm- or criterion-referenced uses in mind. To realize the primary objective of norm-referenced score uses—to compare and select individual examinees—such tests need to ensure sufficient score variability by comprising items of appropriate facility for the intended population (Sawaki, 2016). Criterion-referenced purposes, on the other hand, are best served by items that most accurately discriminate between examinees of different mastery or performance categories (Sawaki, 2016). Thus, norm-referenced-ness and criterion-referenced-ness can in fact be said to constitute an interaction between test properties and the interpretation and uses of its scores.

The preceding discussion of norm- and criterion-referenced score uses implicitly focuses on fixed-form tests, which have been the operational reality for all but the most recent history of standardized testing. It would be challenging for a fixed-form test to simultaneously serve both norm- and criterion-referenced purposes adequately, given their different priorities and resulting differences in test properties and score distributions. This is perhaps a reason that tests themselves have historically been discussed as norm- and criterion-referenced. However, it is more feasible for a test to serve both purposes in a computer adaptive context, which allows for the customization of the test to each examinee's individual ability, potentially enabling both sufficient score variation and adequate classification consistency. As mentioned earlier, the Duolingo English Test is such a computer adaptive test intended to serve multiple purposes, which has implications for the evaluation, communication, and improvement of classification consistency.

3.1.3 Extant Classification Consistency Indices

While there are several dimensions upon which methods for calculating classification consistency indices can be categorized, perhaps the most fundamental is the measurement paradigm on which they are founded. The primary distinction between existing methods is whether they are CTT– or IRT–based. The remainder of this section will summarize the most prominent current methods, with a focus on IRT–based methods, given that the Duolingo English Test uses a latent trait model in its scoring procedures.

3.1.3.1 IRT–based approaches

Item response theory (IRT) is a measurement paradigm that assumes test responses reflect a latent trait of the test taker and derives scores from a model that accounts for characteristics (principally difficulty) of the test items (Embretson & Reise, 2000).

Two foundational aspects of IRT crucial for present purposes are that the error of measurement varies along the latent ability spectrum and that individual ability (i.e., the person parameter) is invariant, or independent of the specific test items used to measure it (De Ayala, 2013).

3.1.3.1.1 Rudner (2000, 2005). While other IRT methods were proposed before and have been developed since, Rudner's method is a popular and conceptually straightforward IRT–based approach to classification consistency. It uses IRT ability estimates and individual or conditional standard errors of measurement (CSEMs) to compute the probability of misclassification for each score point on the test's scale, assuming normally distributed measurement error. For a fixed-form IRT–calibrated test, the standard error is presumably the same for all examinees with a given score estimate, since all examinees respond to the same items and thus have the same test information curve. On a CAT test, however, examinees with the same final score estimates will have responded to different items and may therefore have appreciably different associated standard errors if the CAT algorithm's stopping rule involves anything other than a standard error threshold. In either case, misclassification probabilities can be calculated and plotted to show change in misclassification probability across the score scale or aggregated to compute an overall classification consistency. Lee (2010) suggests the Rudner approach for tests whose reported scores are transformations of θ. The method is implemented in software by the author and in the R package cacIRT (Lathrop, 2015).

3.1.3.1.2 Hambleton & Han (2004). As described in Diao & Sireci (2018), the Hambleton and Han method (in Bourque, Goodman, Hambleton, & Han, 2004) is only applicable in a simulation context as it requires true item parameters in order to simulate two administrations of a test. Thus, it is more a formalization of simulating the two-administration approach than an operationally viable single-administration method. It is nevertheless potentially useful for investigating through simulation studies how manipulating aspects of a measurement procedure could impact classification consistency and accuracy. Furthermore, having this method formalized facilitates the comparison of results across simulation studies.

3.1.3.1.3 Lee (2010). Lee presents a method for estimating the classification consistency of so-called complex assessments, meaning assessments comprising both dichotomous and polytomous items. The author states that the method is applicable to any test calibrated with one IRT model or a combination thereof. The method assumes that test scores are computed as the sum of individual item scores and that item parameter estimates are available. The item parameters are then used to create conditional observed score distributions following either the binomial or compound

multinomial model, depending on whether the test contains polytomous items, and using a recursive algorithm appropriate for the item formats. At this point, the author presents two options for computing marginal classification indices. In what he refers to as the D method, quadrature is used to approximate the integration of the conditional classification function over the population θ. In the P method, classification consistency is computed for each examinee and the average taken. Lee found these two approaches to produce very similar results when applied to real data from two tests.

3.1.3.2 CTT–based approaches

In contrast to IRT–based approaches, approaches to estimating classification consistency based in classical test theory (CTT; Allen & Yen, 1979) do not utilize any item-level psychometric information. Estimates of classification consistency are based either on observed classification consistency, such as between repeat test administrations or split halves of a single administration, or on assumptions about the population true score and error distributions. The CTT methods investigated in this study are described in the Methods section.

3.1.4 Classification Consistency Indices and Complex Assessments

Lee (2010) used the term "complex assessment" to refer to an assessment comprising both dichotomous and polytomous items. Such assessments are now commonplace and easily handled by psychometric software. The current vanguard of complex assessments is arguably approaches such as game-based assessment and machine learning– and natural language processing–based automated scoring. The nature of these assessments, of which the Duolingo English Test is an example, presents challenges to applying existing classification consistency indices. All mainstream classification consistency indices have been developed, either explicitly or implicitly, for fixed-form tests and have not been definitively endorsed for use on CATs. The extent to which they are appropriate for a CAT context is therefore an empirical question that has not been sufficiently addressed. In the case of CTT–based approaches, Duolingo English Test data might violate distributional assumptions, and there is also potential concern about the representativeness of repeat test takers whose data would be used for a test–retest approach. Furthermore, existing IRT–based approaches rely on empirical operational item parameters or model-based standard errors of examinee ability estimates, neither of which is available for the DET due to the innovative nature of its development. And both CTT– and IRT–based approaches assume some degree of parallelism between test forms used in hypothetical replications. Thus an innovative test like the DET will require an innovative approach, or verification of a traditional approach, to accurately

estimate classification consistency.

3.1.5 CAT Item Selection and Classification Consistency

Lee (2010) notes that CTT–based and IRT–based approaches to classification consistency make different assumptions about the error scores, with CTT–based assuming the test forms used in hypothetical measurement replications are randomly parallel, whereas the IRT–based assume test forms are strictly parallel, meaning that each test form has identical item parameters. The author further observes that the approaches with the stricter assumption about form parallelism will necessarily produce higher estimates of classification consistency. It thus seems possible that a testing program that uses IRT models for scoring purposes is not necessarily compatible with an IRT–based approach to classification consistency. In a CAT context, examinees can take multiple paths to arrive at the same final score estimate, meaning the ensemble of item parameters will vary across administrations. This observation naturally leads to the examination of a CAT's item selection algorithm.

The item selection algorithm "is the most critical component" in a CAT administration, and it conventionally encompasses the three components of content balancing, item selection criterion, and item exposure control, as well as the test termination rules (Han, 2018, p. 1). Thus it is the item selection algorithm that determines the psychometric properties of a CAT administration based on how each additional item is selected and at what point the administration ends. Numerous item selection algorithms have been proposed, but those used operationally all make use of either item difficulty or item information (Han, 2018). Additionally, a CAT administration can be specified to terminate after reaching a specific threshold of score estimate precision, after administering a specific number of items, or some combination of these and possibly other criteria. In these ways, the CAT algorithm influences the test information curve of the administration and therefore the precision of the final score estimate, which has clear implications for classification consistency. A testing program can thus influence its classification consistency in predictable ways through the design of its CAT algorithm.

A common suggestion in the measurement field is to maximize test information at the cut score for fixed-length classification tests, although it has been shown that, depending on test length and the location of the population mean score relative to the cut score, maximizing information at the cut score does not necessarily optimize classification consistency (Wyse & Babcock, 2016). Wyse & Babcock (2016) found that the ideal point to maximize test information is generally between the population mean and the cut score. This conclusion is comparable to that of work done in an adaptive classification context, which has shown that the optimal item to administer is located in between the cut score and the examinee's true θ (Nydick, 2014). Rudner (2009) takes

the idea of incorporating classification consistency into the CAT algorithm further in his Measurement Decision Theory (MDT) framework. MDT involves updating estimates of an examinee's probability of belonging to each classification category after each administered item and using these probability estimates to guide item selection and test termination.

The role of the CAT algorithm in determining classification consistency will need to be considered when interpreting the results of the present research and making recommendations for future research and strategies for improving classification consistency as needed.

3.1.6 Research Questions

The present study uses a real-data application to investigate the following research questions concerning the estimation of classification consistency indices for an innovative CAT assessment:

1. How do results of applying CTT–based classification consistency estimates (test–retest, split-half, Livingston–Lewis) to real CAT scores compare to simulation results?
2. To what extent is it possible to apply a classification consistency method of the IRT paradigm to data from the Duolingo English Test? How do the results compare to those of the CTT–based methods?

3.2 Methods

3.2.1 Data Source

Analyses use data from 96,677 certified Duolingo English Test sessions of first-time test takers and 13,938 certified sessions of test takers taking the Duolingo English Test for the second time. All data are from test sessions that took place between March 26, 2021 and September 15, 2021.

3.2.2 Regression Analysis of Score Change Over Time

Before computing classification consistency indices, a multiple regression analysis was conducted on the repeat examinee data to investigate any systematicity in the change in ability estimates between test administrations in order to substantiate the assumption that examinee ability is sufficiently stable within a certain period to warrant using this data for test–retest reliability calculations. The maximal model

regression formula takes the form

$$X_2 - X_1 \sim \boldsymbol{Z} + X_1 * \Delta t * \text{L1} + X_1^2 + X_1^3 + \Delta t^2 + \Delta t^3 + \text{L1}^2 + \text{L1}^3 \tag{3.1}$$

where $\boldsymbol{Z}$ represents the vector of test-taker demographic covariates (age and gender), X_1 and X_2 the reported scale scores on the first and second test administrations respectively, Δt represents the time duration between test administrations, and L1 is a quantitative representation of the similarity between a test taker's native language and English. Second- and third-order terms were included for X_1, Δt, and L1 because their relationship to score improvement is not necessarily linear.

Two data sources were considered for quantifying the linguistic similarity between examinee L1 and English. The United States Department of State's Foreign Service Institute—the organization that trains American foreign service workers—maintains estimates of the average number of weeks of training required for a diplomat to achieve "Professional Working Proficiency" in a language, ranging from 24 weeks for languages such as Dutch and Spanish, to 88 weeks for languages such as Arabic and Japanese (Foreign Service Institute, 2020). Chiswick & Miller (2005) also proposed a quantitative measure of linguistic distance from English based on learning difficulty for L1 English speakers. There is not a one-to-one correspondence between the two measures, with Chiswick and Miller assigning different values to languages to which the FSI assign the same number of weeks and vice versa. The two measures are correlated $r = 0.867$. To capitalize on both measures, the present study uses a novel combination of the two measures produced by the formula

$$LD = \frac{1}{2}(\frac{FSI}{88} + CM) \tag{3.2}$$

where LD represents the new linguistic distance measure, FSI represents the Foreign Service Institute classification in weeks, and CM represents the Chiswick–Miller measure. The list of languages used in the analysis and their corresponding LD values is presented in Table 3.2 in the Appendix. Table 3.3 lists the languages that were excluded from the analysis because they do not have a Chiswick–Miller value.

Stepwise bidirectional model selection using BIC and starting from the maximal model was used to determine the final model. This process was implemented once without removing any outliers, and twice with different outlier detection criteria—once based solely on studentized residuals, and another based on both studentized residuals and Cook's distance. Model fitting and outlier detection was iterated until no data points satisfied the outlier detection criteria.

3.2.3 Classification Consistency Analyses

The following classification consistency indices were computed on the Duolingo English Test data for a binary (i.e., pass/fail) cut score at every point on the reported score

scale.

3.2.3.1 CTT–Based Methods

The following CTT–based methods were used to estimate classification consistency.

3.2.3.1.1 Test–Retest Using reported-scale scores from repeat examinees, the test–retest percent agreement estimate is given by

$$p_o = \frac{|\{i|C_{i1} = C_{i2}\}|}{N} \tag{3.3}$$

where C_{i1} represents examinee classification based on the first administration and C_{i2} represents classification based on the second. This index was calculated both with and without adjustment. The adjusted index was calculated by weighting the classification consistency at each score point on the reported scale by the proportion of examinees in the general population at that score point.

3.2.3.1.2 Split-Half This method entails dividing the items of each CAT administration based on administration order into odd and even splits, then re-estimating examinee abilities based on these split-half tests. These half-test scores are already included in Duolingo English Test examinee response data. The percent agreement classification consistency index is then calculated as

$$p_o = \frac{|\{i|C_{ih_1} = C_{ih_2}\}|}{N} \tag{3.4}$$

where C_{ih_1} represents examinee classification based on the first half-test score and C_{ih_2} represents classification based on the second. This raw percent agreement will also be adjusted for the reduced reliability of the shorter half-tests relative to the full-length test.

3.2.3.1.3 Livingston–Lewis (1995) The Livingston–Lewis method was implemented by the R package betafunctions using reported-scale scores and two estimates of test reliability: test–retest and split-half. The most recently reported test–retest reliability estimate of the DET total score is 0.90 (Cardwell, LaFlair, & Settles, 2021). Split-half reliability was estimated directly from the data used in the present study.

3.2.3.2 Bootstrapping

As mentioned above, the approach of many IRT–based methods for estimating classification consistency is to approximate the error distribution around a score estimate

and calculate the proportion of the error distribution on either side of the cut score. This aim can also be accomplished through bootstrapping.

An examinee's DET total score is computed as a weighted sum of seven component scores:

$$X_i = w_1c_{1i} + w_2c_{2i} + w_3c_{2i} + w_4c_{4i} + w_5c_{5i} + w_6c_{6i} + w_7c_{7i} = \mathbf{w} \cdot \mathbf{c}_i \tag{3.5}$$

For each subset of the analytic data set made up of examinees with the same total score $X = T$, the seven component scores were re-sampled with replacement to produce a vector of alternative total scores $\mathbf{X}^*_T = \mathbf{w} \cdot \mathbf{c}^*_T$. This distribution is taken to be the error distribution around the total score T, and the proportion of the distribution on either side of the cut score is used to estimate classification consistency at that cut score conditional on total score.

This procedure was replicated $B = 250$ times to produce 250 bootstrap estimates of conditional classification consistency for each observed total score T at each possible cut score J. The final conditional classification consistency estimate is then the average over the bootstrap sample estimates, $CC_{tj\cdot} = \overline{CC}_{tjb}$. The overall classification consistency at a particular cut score j can then be estimated as a weighted average of conditional classification consistency estimates $CC_j = N^{-1} \sum_{t=1}^{T} n_t CC_{tj\cdot}$.

3.3 Results

3.3.1 Score Change Over Time

Figure 3.1 depicts the relationship in the raw data between time elapsed between the first and second test attempts of repeat test takers and the change in overall score. The data are from 13,938 examinees who received certified results on their first test attempt (i.e., the test results were not invalidated due to a technical error, rule violation, etc.) and later received certified results on a second test attempt. There is an evident increase in score change between 0 and 25 days, followed by a stagnation until approximately 100 days. After 100 days between test attempts, there is another slight increase in score change, although the data sparsity in this range makes the trend uncertain.

Three regression models fit to the score change data are summarized in Table 3.1. Model 1 represents the stepwise selection procedure described in the Methods section applied to the full repeat test-taker dataset without removing any potential outliers. Models 2 and 3 additionally incorporated the removal of potential outliers in the model fitting process. Model 2 was fit by removing data points based on studentized residuals, resulting in 256 observations being removed. Model 3 used both studentized

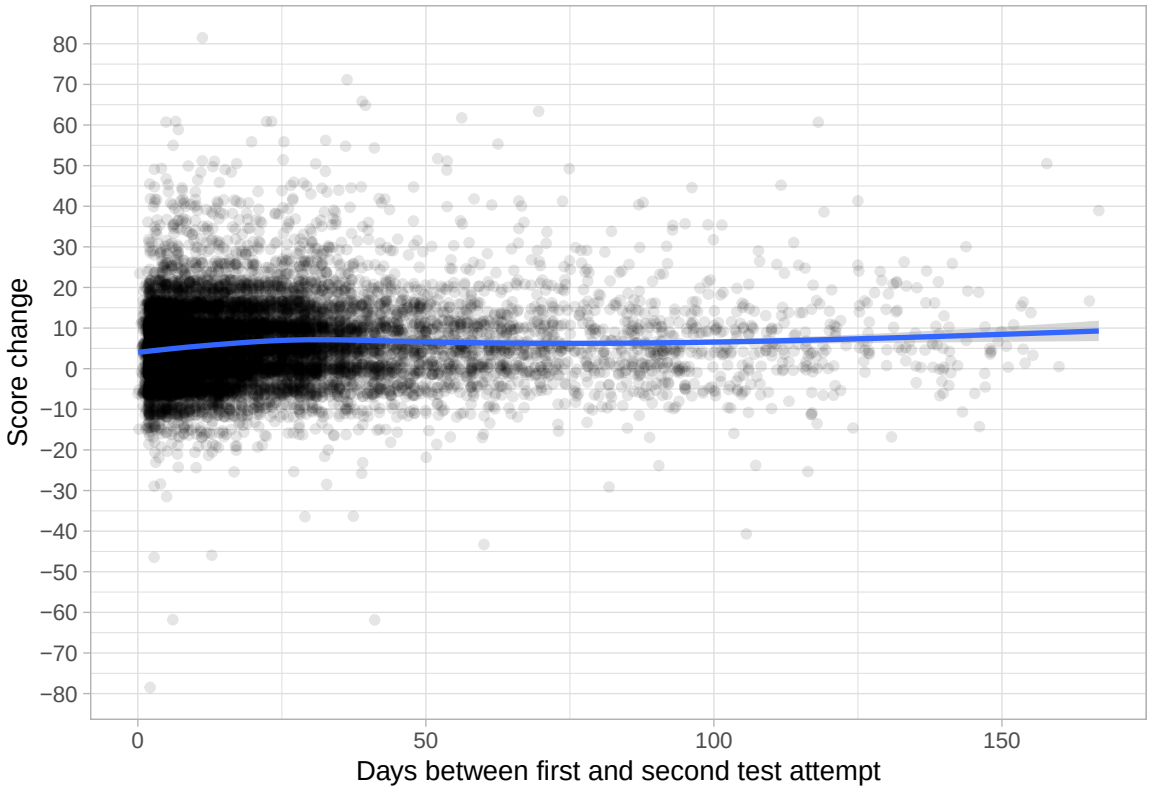

Figure 3.1. Score change over time between first and second test attempts

residuals and Cook's distance as criteria for removing potential outliers, resulting in 930 observations being removed. However, despite removing different numbers of potential outliers, the resulting models are quite similar in the predictors retained and the sign/magnitude of their coefficients. In all models, examinee demographics (age and gender) are dropped, as are most of the interaction terms. However, higher-order terms are retained for both time duration and linguistic distance, suggesting a non-linear relationship with score improvement. However, it should be noted that for all models, the adjusted-R^2 was less than 0.07, indicating that the predictors collectively account for very little variance in the score change of repeat examinees. Additionally, the complexity of the models make interpretation difficult, and so some example predictions would be illustrative.

Figure 3.2 below shows predictions of score change, on the 10–160 Duolingo English Test scale, from all three models for 1 to 50 days between the first and second test attempts. All predictions fix $X_1 = 100$, which is the median and approximate mean score of repeat examinees on the first test attempt. The three colored lines in each panel of the plot correspond to three different levels of LD: 0.303 (e.g., Romanian), 0.583 (e.g., Vietnamese), and 1.0 (e.g., Korean). For all values of LD under all models, predicted score gains start small (<3) and increase noticeably until about 10 days between test attempts, and then effectively level off after two weeks. Under all three models, predicted score gains for $LD \in 0.303, 0.583$ are similar and consistently higher than for $LD = 1$. However, the shaded regions represent the 95% confidence intervals of the predicted values, and reflect the high proportion of variance in score change that remains unaccounted for by the models.

The results of the score change regression analysis indicate that there is little difference in predicted score change for inter-test durations greater than ~10 days. Score gains within 10 days are unlikely to reflect true change in English language proficiency, and are more plausibly attributable to factors such as test familiarity, motivation, and measurement error. The low adjusted-R^2 and the fact that 37% of repeat examinees in the dataset did not achieve a higher score on the second test attempt are consistent with measurement error being a primary factor in score change. There is thus no evidence to suggest widespread, systematic change in true ability on the test construct between test attempts, making the cutoff for inclusion in the repeat examinee dataset effectively arbitrary. Subsequent analyses will use data from repeat examinees with inter-test durations up to 50 days. Observations with score changes more than 2.5 SDs from the mean score change (i.e., outside of the range [-22.3, 33.9]) are also excluded as outliers, leaving 89% of the original repeat examinee dataset.

Table 3.1. Table comparing regression models fit to the score change data

	delta_score	step_fit	
	(1)	(2)	(3)
$log(\Delta t)$	3.697^{***}	2.820^{***}	2.976^{***}
	(0.530)	(0.456)	(0.459)
X_1	-0.208^{***}	-0.038^{***}	-0.027^{**}
	(0.043)	(0.011)	(0.011)
LD	-53.406^{***}	-49.495^{***}	-71.732^{***}
	(13.879)	(11.866)	(11.285)
X_1^2	0.001^{***}		
	(0.0002)		
LD^2	102.044^{***}	89.073^{***}	127.533^{***}
	(22.073)	(18.856)	(18.074)
$log(\Delta t)^3$	-0.027^{***}	-0.027^{***}	-0.038^{***}
	(0.009)	(0.007)	(0.008)
LD^3	-58.693^{***}	-49.106^{***}	-70.028^{***}
	(11.225)	(9.581)	(9.272)
$log(\Delta t)^2 : X_1$	-0.025^{***}	-0.017^{***}	-0.016^{***}
	(0.005)	(0.004)	(0.004)
Intercept	23.919^{***}	14.907^{***}	17.359^{***}
	(3.509)	(2.569)	(2.460)
Observations	11,003	10,747	10,073
Adjusted R^2	0.069	0.045	0.045
Note:			*p<0.1; **p<0.05; ***p<0.01

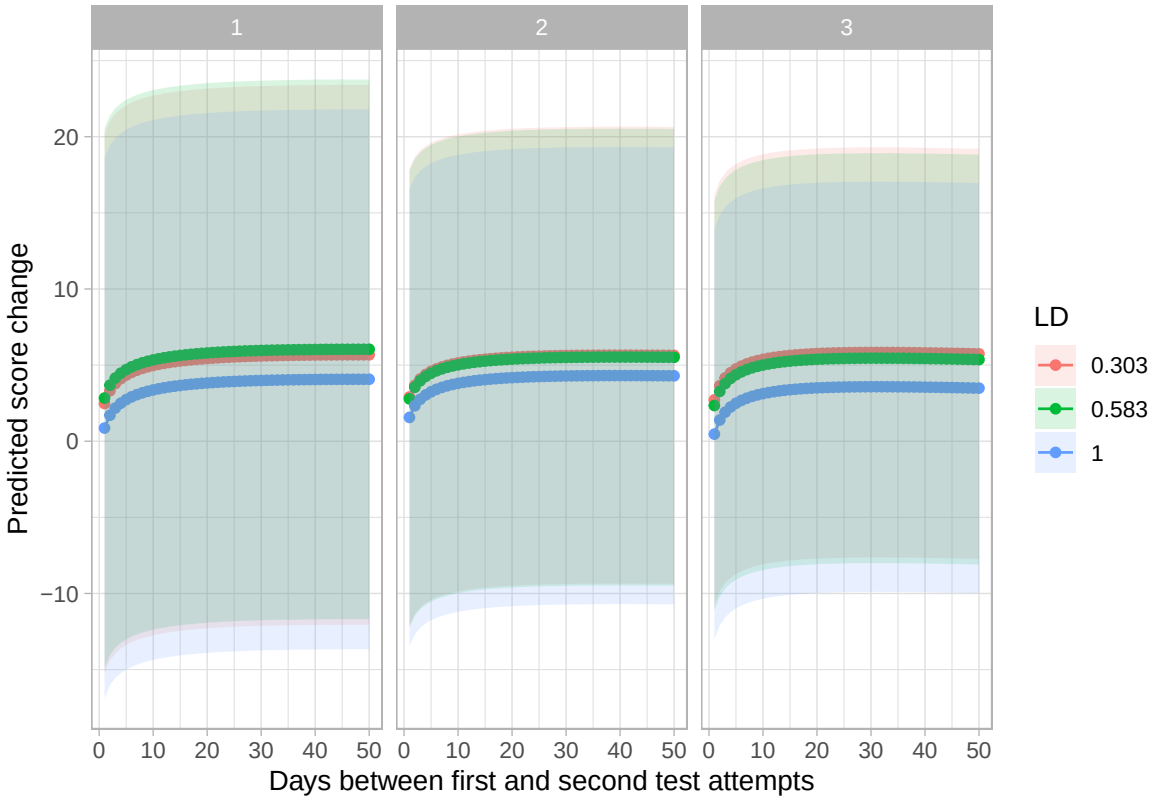

Figure 3.2. Model predictions of score change by duration between test attempts

3.3.2 Classification consistency

Figure 3.3 below depicts the densities of the Duolingo English Test total score distributions for first-time test takers, repeat test takers (first time), and repeat test takers (second time). It is clear that repeat test takers score substantially lower than the general test-taker population on both their first and second test attempts. However, there are also repeat test takers with first-attempt scores at almost all scale points, demonstrating that repeat test takers do not strictly self-select in the sense that no test taker above a certain threshold repeats the test.

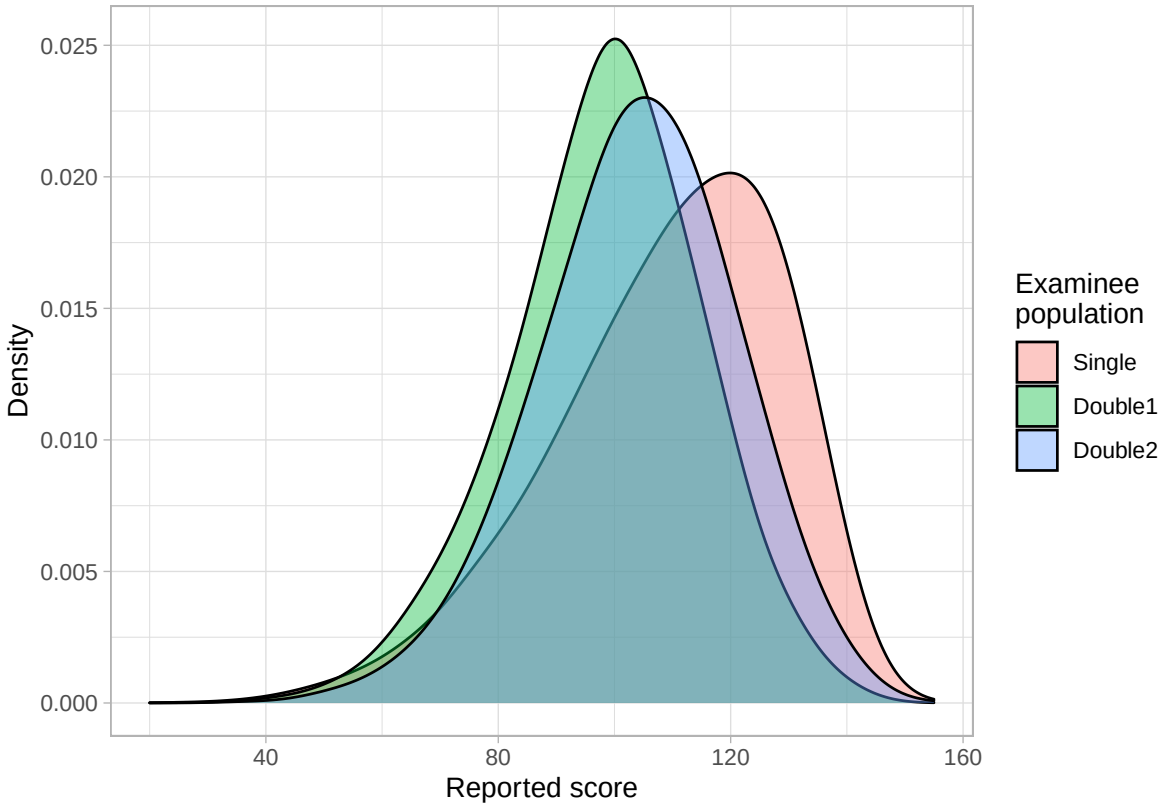

Figure 3.3. Total score distributions for first-time test takers (Single), repeat test takers' first time (Double1), and repeat test takers' second time (Double 2)

Figure 3.4 depicts the percent agreement classification consistency estimate using each method at each point on the Duolingo English Test 10–160 score scale for which there is sufficient data. At the extremes of the scale, all classification consistency estimates approach 100%, primarily due to the low density of scores in these regions, leading to high probability of consistent classification because the vast majority of test takers are far from the cut score. In the 90–120 score range, which encompasses a slight majority of actual test results and a large majority of institutional admissions cut scores, the

results broadly mirror those of the simulation study reported in the previous chapter. The raw test–retest estimate is generally the lowest, but the adjusted test–retest estimate is comparable to the raw split-half estimate. The adjusted split-half estimate was consistently higher than most others. The Livingston–Lewis estimate depended greatly on the estimate of test reliability used in its calculation. At all score points between 90 and 120, all classification consistency estimates (with the exception of TRT-raw) differ by less than five percentage points.

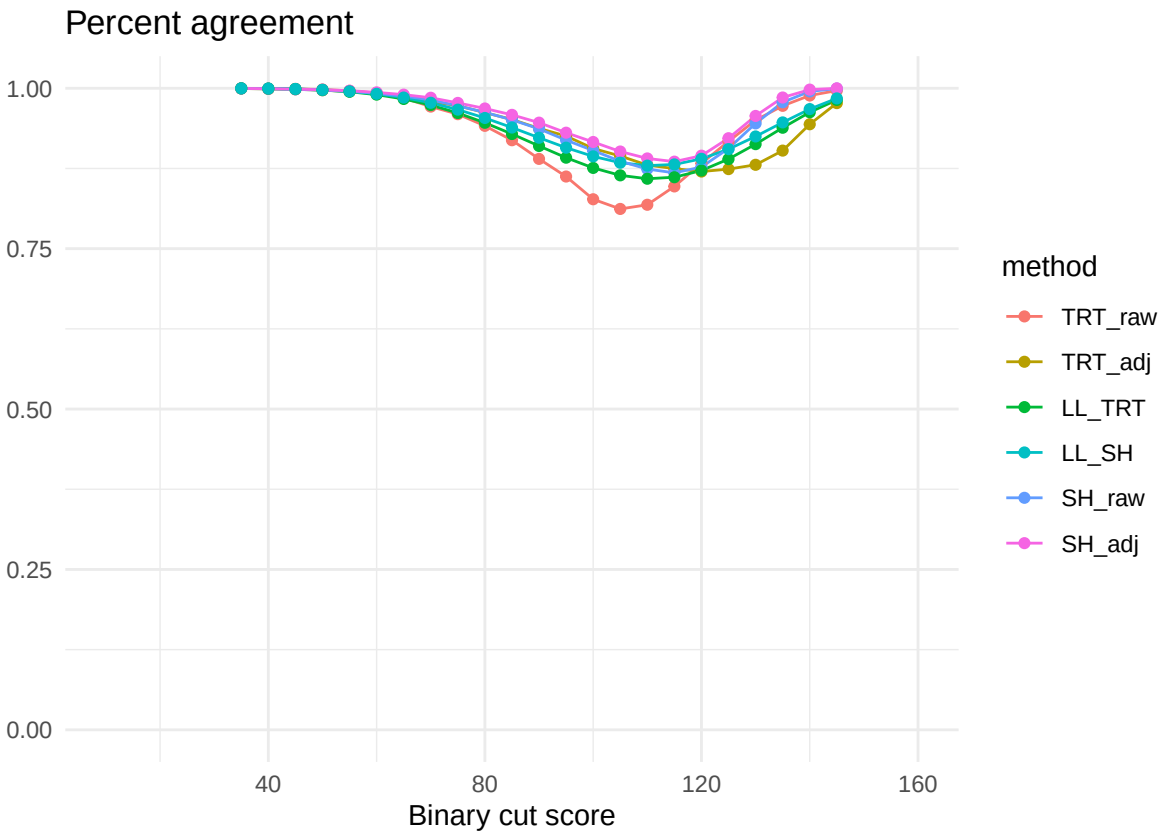

Figure 3.4. Percent agreement estimate by method at each score point on the Duolingo English Test scale

Figure 3.5 displays the same results as Figure 3.4, but as Cohen's kappa to adjust for the probability of chance classification consistency. The relative ranking of methods in the 90–120 score range is largely unchanged. The decreasing kappa values outside of this range illustrate that the high observed percent agreement in these regions is more attributable to chance as the distance increases between the cut score and the mean of the score distribution. With the exception of the raw (unadjusted) test–retest method, all classification consistency methods produce Cohen's kappa estimates above .60 for score points in the 90–120 range, indicating acceptable classification consistency.

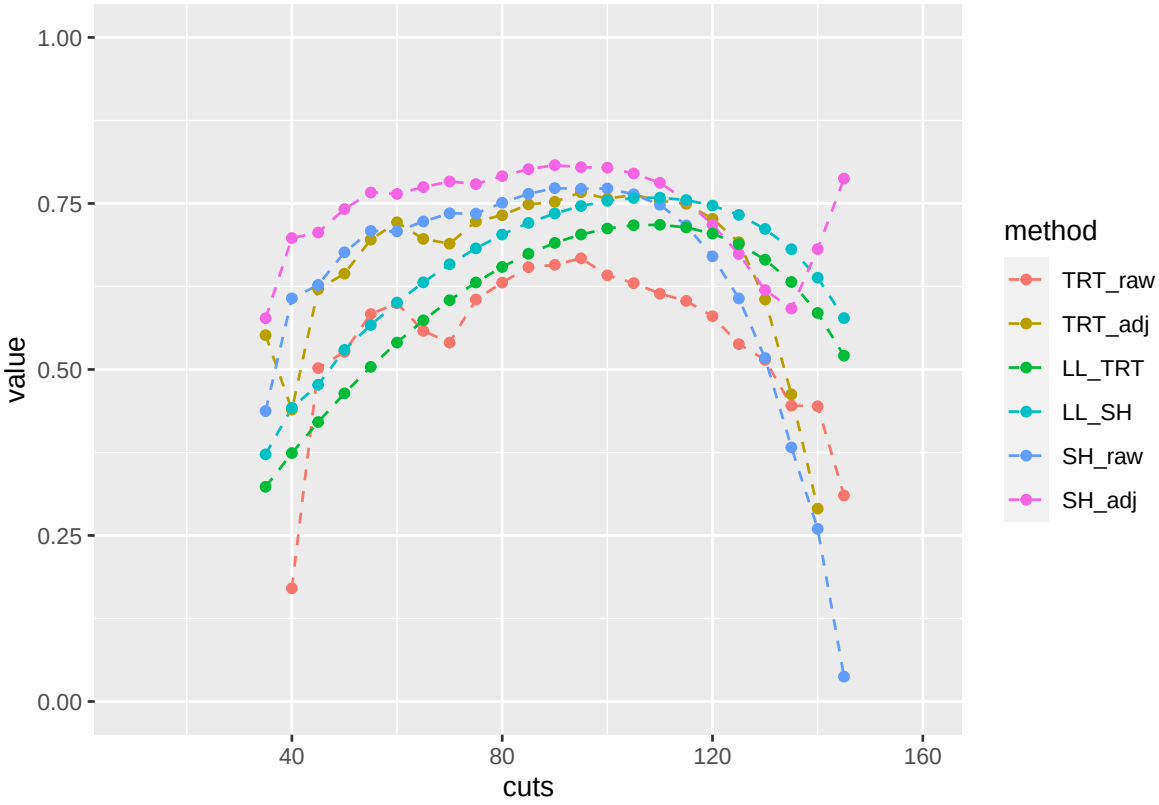

Figure 3.5. Cohen's kappa estimate by method at each score point on the Duolingo English Test scale

3.3.3 Bootstrapping

Figure 3.6 depicts the results of the bootstrapping approach overlaid on the percent agreement plot of the other classification consistency methods. The solid black line represents the results of bootstrapping performed as described in the methods section. It is clear that this method produces much higher estimates of classification consistency than other methods in a large range of the score scale. This indicates that the bootstrapping approach could be overestimating total score precision, potentially due to the fact that the component scores used in the bootstrapping are themselves estimates. To rectify this bias, the bootstrapping approach was repeated with sampling weights applied to each component score, with the weight of the ith observation given by $m_i = |z_i| + 1$ where z_i represents the z-score of the observation. This weighting increases sampling from the tails of the distribution. The results of this weighted bootstrapping approach are represented by the dotted line in the figure. It can be seen that these results are more in line with those from the other methods. However, more research is needed to provide a theoretically sound justification for the choice of

weights.

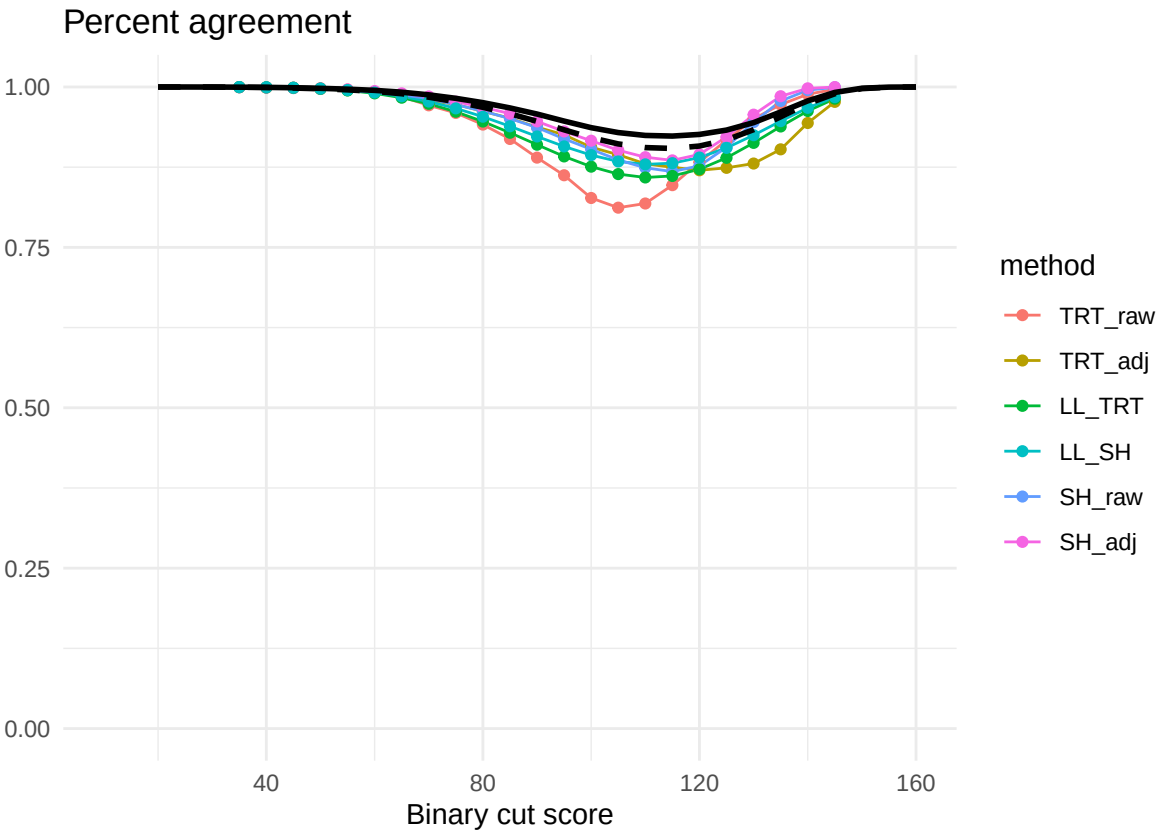

Figure 3.6. Results of bootstrapping approach compared to other classification consistency estimates

3.4 Discussion

The language distance (LD) score proved to be a significant predictor in regression models of score change between administrations, presenting a promising area for future research. As a novel measure, the combination of FSI and Chiswick–Miller (2005) values requires further verification. However, despite its significance as a predictor variable, a test taker's LD is highly confounded with other cultural and socioeconomic factors, making it difficult to conclude that linguistic (dis)similarity alone is responsible for the significance of the LD predictor. Additionally, the low adjusted-R^2 values and non-linear nature of model-predicted score gains indicate that there is no evidence of meaningful, systematic change in true ability between test administrations within the time spans represented in the data (maximum 167 days). This result supports the

use of repeat test-taker data in estimating test properties such as score reliability and classification consistency.

The various CTT-based classification consistency methods applied to the DET data (test–retest, split-half, and Livingston–Lewis) produce results that are broadly consistent with findings from the simulation study from the previous chapter. Specifically, the results are consistent with the simulation study finding that the test–retest approach using only low-scoring examinees can greatly underestimate expected classification in certain conditions. The real data application also produced substantially lower classification consistency estimates at cut scores within about half a standard deviation of the repeat test-taker mean score. Otherwise, the various methods produced comparable classification consistency estimates, indicating that each method is potentially appropriate. But it is impossible to know which method most accurately estimates classification consistency.

3.4.1 Bootstrapping approach

The bootstrapping approach produced results with a similar pattern to those of the CTT–based methods, but seemed to substantially overestimate classification consistency (assuming that true classification consistency is bounded by the other methods). There are multiple possible explanations for this observation. The procedure applied in the present study begins with stratifying the data by total score on the reported score scale, assuming that (given a sufficient sample size) the observed component scores of test takers within the stratum accurately represent the distribution of all possibly observable component scores for a test taker at that ability level. Then, the component scores are effectively resampled independently and combined to estimate alternative total scores. The distribution of alternative scores are taken as the error distribution around a given true score. Thus, the overestimation of classification consistency implies the underestimation of the variance of the error distribution. One or more of the following steps of the bootstrapping approach could be causing the underestimation of the conditional error variances: estimation of component scores from test responses, stratification of the sample by total score, and estimation of alternative total scores as weighted combinations of resampled component scores.

If there is any error in the estimation of the component scores, this issue would also be present in the other classification consistency methods, and thus it is not clear how component score estimation error could be the primary cause of inconsistent results from the bootstrapping approach. The stratification of the sample on total score (as a proxy for true score) could, however, be artificially reducing the variance of the alternative score distributions by imposing a constraint on the linear combination of component scores in each stratum. An adjustment could be necessary to produce a more accurate estimate of the conditional error distribution, conceptually similar to

methods for correcting bias in bootstrap confidence intervals that can change both the width and midpoint of the confidence interval (see e.g., Efron, 1987). Lastly, the way of combining component scores could be a contributing factor. When combining multiple estimates, weighting each estimate inversely to its corresponding variance is known to produce a consistent estimator with minimum variance (Graybill & Deal, 1959). However, the component score weights used to estimate total scores do not incorporate component score estimate precision, which could have unanticipated effects on the mean and variance of the alternative score distributions. More research is needed to identify the cause(s) of the bootstrapping results inconsistency and justify an appropriate correction.

3.5 Conclusion

Each of the classification consistency methods considered in this study has pros and cons, and is potentially appropriate for a particular test. A test–retest approach is only feasible if there are a sufficiently large number of repeat test takers, and the repeat test takers represent most of the ability levels found in the overall test-taker population. A split-half approach requires that test-taker responses can be separated and re-scored. The Livingston–Lewis approach relies heavily on an estimate of score reliability, and any inaccuracy thereof would substantially affect the resulting classification consistency estimates. And the bootstrapping approach is only applicable to tests for which the total score is derived as a weighted sum of component scores. Additionally, it is possible that the bootstrapping approach overestimates true classification consistency, and so further research is needed to investigate the properties of the bootstrap estimate and potentially develop corrective adjustments such as sampling weights. Notwithstanding these caveats and limitations, the simulation study of the previous chapter and real-data application of the present study provide evidence that CTT-based classification consistency methods are not fundamentally incompatible with non-fixed-form tests, and that bootstrapping presents a potential alternative in contexts where there is insufficient information to apply traditional IRT–based approaches.

3.6 Appendix

Table 3.2. Linguistic distance (LD) measures for languages used in score change regression analysis

Language	Language Family	Count	FSI Weeks	Chiswick–Miller	LD
Norwegian	Indo-European	13	24	0.333	0.303
Romanian	Indo-European	228	24	0.333	0.303
Swedish	Indo-European	13	24	0.333	0.303
Dutch	Indo-European	15	24	0.364	0.318
Italian	Indo-European	125	24	0.400	0.336
Portuguese	Indo-European	492	24	0.400	0.336
Danish	Indo-European	13	24	0.444	0.359
Spanish	Indo-European	1490	24	0.444	0.359
French	Indo-European	481	30	0.400	0.370
Malay	Austronesian	7	36	0.364	0.386
Swahili	Niger–Congo	9	36	0.364	0.386
German	Indo-European	74	36	0.444	0.427
Indonesian	Austronesian	118	36	0.500	0.455
Russian	Indo-European	230	44	0.444	0.472
Amharic	Afro-Asiatic	50	44	0.500	0.500
Bulgarian	Indo-European	49	44	0.500	0.500
Czech	Indo-European	8	44	0.500	0.500
Persian	Indo-European	526	44	0.500	0.500
Finnish	Uralic	2	44	0.500	0.500
Hebrew	Afro-Asiatic	78	44	0.500	0.500
Croatian	Indo-European	2	44	0.500	0.500
Hungarian	Uralic	21	44	0.500	0.500
Khmer	Austroasiatic	6	44	0.500	0.500
Mongolian	Mongolic	13	44	0.500	0.500
Polish	Indo-European	34	44	0.500	0.500
Serbian	Indo-European	10	44	0.500	0.500
Tagalog	Austronesian	57	44	0.500	0.500
Thai	Tai–Kadai	180	44	0.500	0.500
Turkish	Turkic	292	44	0.500	0.500
Bengali	Indo-European	217	44	0.571	0.536
Greek	Indo-European	29	44	0.571	0.536
Hindi	Indo-European	336	44	0.571	0.536
Nepali	Indo-European	54	44	0.571	0.536
Sinhala	Indo-European	81	44	0.571	0.536
Burmese	Sino-Tibetan	38	44	0.575	0.537
Lao	Tai–Kadai	4	44	0.667	0.583
Vietnamese	Austroasiatic	137	44	0.667	0.583
Arabic	Afro-Asiatic	819	88	0.667	0.833
Mandarin	Sino-Tibetan	3831	88	0.667	0.833
Cantonese	Sino-Tibetan	213	88	0.800	0.900
Japanese	Japonic	275	88	1.000	1.000
Korean	Koreanic	333	88	1.000	1.000

Table 3.3. Languages excluded from regression analysis due to lack of Chiswick–Miller (2005) value

Language	Language Family	Count	FSI Weeks
Haitian Creole	Creole	NA	36
Azerbaijani	Turkic	65	44
Tibetan	Sino-Tibetan	4	44
Estonian	Uralic	3	44
Armenian	Indo-European	11	44
Icelandic	Indo-European	7	44
Georgian	Kartvelian	6	44
Kazakh	Turkic	40	44
Kyrgyz	Turkic	NA	44
Kurdish	Indo-European	17	44
Latvian	Indo-European	4	44
Lithuanian	Indo-European	17	44
Macedonian	Indo-European	2	44
Slovak	Indo-European	6	44
Slovenian	Indo-European	1	44
Somali	Afro-Asiatic	13	44
Albanian	Indo-European	25	44
Tamil	Dravidian	114	44
Telugu	Dravidian	537	44
Tajik	Indo-European	7	44
Turkmen	Turkic	4	44
Ukrainian	Indo-European	54	44
Urdu	Indo-European	310	44
Uzbek	Turkic	23	44

Chapter 4

Paper 3: Results Reporting of an Innovative High-Stakes CAT

4.1 Introduction

As the primary point of contact between test developers and results[1] users, results reports bridge test development and results use. Results users must be able to understand information in results reports in order to make appropriate use of results while avoiding misinterpretation and misuse. But despite the central role of results reports in interpreting and using results for important decision-making purposes, "score report design and development... receives scant attention in the psychometric literature" (Zapata-Rivera & Katz, 2014, p. 443).

Results reporting best practices are receiving more attention in recent years (Zenisky & Hambleton, 2016). This trend reflects an increasing awareness of results reports as an integral part of a testing program that should be addressed during test development to ensure alignment between test characteristics and results reporting practices (Zapata-Rivera & Katz, 2014). Results reports provide the basis for results user decisions, and thus are a crucial link in a test's validity argument, as well as a tool for communication with stakeholders (Clauser & Rick, 2016). A nuanced and theoretically grounded perspective on results reports can guide research and design for report development

[1] In this paper I primarily use the terms "results" and "results report," as suggested by O'Donnell & Zenisky (2020), to encompass all information about an examinee's performance that is communicated to stakeholders. This includes scores, but also more broadly verbal descriptions of examinee performance and ability, as well as samples of examinee performance. Any use of the term "score report" should be understood to potentially include results beyond numerical scores. I will use the term "score" to refer specifically to numerical scores, although quotes from other sources may use this term in a more general sense.

(Clauser & Rick, 2016), as well as the integration of results report development into the validation process (Zapata-Rivera & Katz, 2014).

4.1.1 Results Reporting and Validity

In Kane's (2006, 2013) argument-based validity framework, score interpretation and use are conceptualized as a series of inferences and related arguments starting with an examinee's test responses and building to the intended interpretation/use. An implicit assumption is that the inference-maker has access to, and comprehends, the evidence on which the inferences are based. Indeed, in discussing Toulmin's (1958) model of inference, a framework for evaluating arguments that is foundational to Kane's argument-based approach to validity, Kane (2006) asserts that Toulmin "treated his model as a dialogue between an advocate and a challenger" (p. 28). This adversarial approach is evident in Kane's framework and establishes the process of validation as an attempt to convince someone who is knowledgeable and motivated enough to critically analyze a test's development. Such an approach, of trying to convince a critical peer, is understandable from the perspective of ensuring that test developers uphold high standards. But it nonetheless neglects the issue of potential misinterpretation of test results by test users who are not knowledgeable or motivated enough to critically evaluate the use of a test for a given purpose.

Kane (2006) only mentions score reports to clarify that the warrants (i.e., premises) and backing (i.e., evidence) for the inferences of an interpretive argument are typically not included in score reports, but rather in a technical report or similar supporting documentation. But as the medium to transmit information to results users to make judgements and decisions about examinees, "score reports are where the 'rubber hits the road' in the validity argument for a test" (Zapata-Rivera & Katz, 2014, p. 442). Under the conceptualization of validity as the collected evidence justifying particular score interpretations and uses (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014), a significant risk to valid score interpretation occurs when the score user does not receive and comprehend the appropriate information, resulting in potentially improper uses. In this sense, score reports are a component of any validity argument, and are thus subject to the need for validation. The score reporting literature thus represents the measurement community's endeavor to actualize tests' validity arguments by bridging the gap between test use validation and the needs and abilities of results report audiences.

While Kane's (2006, 2013) work represents mainstream thinking on validity in the US measurement community, other validity frameworks are influential in different geographical and/or professional contexts. Bachman & Palmer (2010) is one such framework popular in the language testing field. The authors operationalize test use

validation through the assessment use argument (AUA). "AUA is anchored in consequences and communication plays a critical role in turning technical documentation into information/actions that relevant stakeholders can utilise" (Chalhoub-Deville & O'Sullivan, 2021, p. 65). So the role of communicating information to stakeholders is certainly present in some validity frameworks, although that is not to say that the issue is sufficiently addressed in real-world validation. In their discussion of the history and present state of validity, Chalhoub-Deville & O'Sullivan (2021) acknowledge that test documentation and publications are typically written for an audience of "professional peers" (p. 154), even though the ultimate goal of testing programs is "to render technical information meaningful" and actionable to stakeholders. The authors question why stakeholder communication is not yet a distinct area of scholarship within the measurement field, and their proposal of an integrated argument-based approach to validation prominently features the communication engagement argument as one of the four key arguments in the model. Thus, greater and more explicit attention to stakeholder communication, of which results reports are a part, is already part of current forward-thinking validity theorization, and likely to factor more prominently in future applied validation.

Concerning the use of ELP tests to make postsecondary admissions decisions, the ultimate inference admissions officers make on the basis of test results is whether an applicant's proficiency in the language of instruction will pose a significant barrier to degree completion. In a doctoral dissertation of expansive scope, Banerjee (2003) surveyed the literature on the predictive validity of ELP tests for postgraduate academic success, investigated the decision making processes of two admissions officers, and interviewed a sample of international students over the course of their degree to ascertain in detail their academic success. Among her many insights and conclusions, Banerjee noted that the admissions officers had a good understanding of the language abilities students needed for particular academic programs, but they were unable to articulate the relationship between those abilities and ELP test results (i.e., what individuals with a particular score should be able to do). She asserted that it is the responsibility of the test developer to provide such information (e.g., score band descriptors) to test users.

4.1.2 Results Reports as Interventions

Another perspective on results reports that complements their role in validity is to view them as an extremely brief and small-scale intervention. This perspective draws on the concepts and approaches of program evaluation, including that of a theory of action (NCME, 2018), to emphasize the fact that results reports constitute an action by test developers on test users, which is intended to effect change in the report reader's knowledge state and lead to desired behavioral outcomes (O'Donnell

& Zenisky, 2020). Following the analogy of a theory of action, it is thus necessary to evaluate the components of a results report, explicitly state the intended short- and long-term effects on report audiences, and posit a causal mechanism connecting the two.

A more evaluation-oriented approach to results report development involves identifying stakeholder groups, assessing their needs, and determining whether the results reporting program is meeting those needs. Indeed, the existing models for results report development draw inspiration from the field of user experience design, which focuses on meeting the informational needs of users (i.e., stakeholders). The framework presented by Zapata-Rivera & VanWinkle (2010) and expanded upon by Zapata-Rivera & Katz (2014) explicitly incorporates audience analysis – an approach to formalizing the needs and characteristics of score report audiences.

Given that score reports constitute a link in the chain of validity inferences and a tool for communicating with exam stakeholders, the methodological, evidence-based design of score reports must address issues of both validation and evaluation. Therefore, a comprehensive score report evaluation that seeks to address both validity concerns and user experience will blend elements of validation-research and program-evaluation approaches.

4.1.3 Best Practices in Results Reporting

In order to evaluate and improve the results reporting practices of a particular testing program, it is necessary to understand the best practices and current thinking in the field. This section will summarize relevant requirements from the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014) and advice from the academic literature on results reporting.

4.1.3.1 The *Standards for Educational and Psychological Testing* (2014)

The *Standards* (2014) are a natural starting place for information on testing-related best practices. Concerning score reporting advice from the *Standards*, Zenisky & Hambleton (2016) assert that guidance on score reporting in the *Standards* has become "more concentrated and direct" (p. 585). In their application of the Hambleton & Zenisky (2013) model, Clauser & Rick (2016) summarized the guidance of the Standards as recommending "that reports provide clear explanations of the meaning and limitations of all reported scores (Standards 5.1, 5.4), address likely misinterpretations (Standard 5.3), and provide interpretations that are appropriate to the intended audience (Standard 6.10)" (p. 4). But the advice from the Standards concerns general principles and does not dictate a single best way to design a score report (Zenisky

& Hambleton, 2016). A well-designed report will provide target audiences with the needed information in a comprehensible manner, which can be accomplished through numerous score report designs.

4.1.3.2 Research on Results Reporting

There are no precise, universally-applicable requirements for a good results report, since different methods of communicating test results will be appropriate and effective depending on the test purpose and report audience (O'Donnell & Zenisky, 2020). The research literature can therefore provide, at best, general principles and guidelines on reporting results in a particular context, supported with evidence of audience preference for, and communicative effectiveness of, different report features. For example, in the context of medical certification, Lipner & Brossman (2017) advise that score reports should display the most important information most prominently, use clear and simple language with glossaries, and provide information that could be used for improvement of future performance. Clauser & Rick (2016) found through focus groups and surveys a preference for a concise front page of the score report that clearly communicates the score and pass/fail result, as well as a desire for more normative score information. The authors note, however, that audience preferences are not always consistent with current best practices (Clauser & Rick, 2016). Many of the authors' results report designs therefore opt for qualitative descriptors in their visual displays of examinee performance. This approach illustrates the need to balance audience desires with the reasonable prevention of misinterpretation and misuse.

Zenisky & Hambleton (2016) note that the increasingly common digital delivery of results reports allows for interactivity, such as by including hyperlinks to additional information that can facilitate score interpretation. Additionally, they note that reporting should align with test blueprints and reflect reliable and valid results interpretation. A test score's measurement error should be clearly communicated and explained, using methods such as confidence bands, even though external stakeholder focus groups often suggest removing the confidence bands because they can be confusing (Zenisky & Hambleton, 2016).

4.1.4 Results Reporting of the Duolingo English Test

The Duolingo English Test is an internet-delivered computer adaptive test that measures general English language proficiency and the subskills thereof (Cardwell, LaFlair, & Settles, 2021). The test comprises seven distinct item types, of which five are objectively scored (i.e., there is a definitively correct target response) and the other two are extended speaking and writing items, the responses to which are scored automatically using linguistic properties of the examinee's written or spoken response

(Cardwell, LaFlair, & Settles, 2021). Scores on these seven item types are combined to compute a reported overall score, as well as four reported subscores: Literacy, Conversation, Production, and Comprehension. In addition to these numerical scores, the Duolingo English Test also publishes total score ranges corresponding to each of the six levels of the Common European Framework of Reference for Languages (CEFR; Council of Europe, 2018) and associated level descriptors in the form of can-do statements (Duolingo English Test, 2021). Furthermore, test results users (i.e., college and university admissions officers) are also provided with examinees' responses to unscored speaking and writing items.

4.1.4.1 Ungraded speaking and writing samples

One of the many unique features of the Duolingo English Test is the inclusion of one ungraded speaking item and one ungraded writing item, called the Speaking Sample and Writing Sample, respectively (Duolingo English Test, 2021). The Speaking Sample and Writing Sample are administered at the end of the test, after all items that contribute to the numerical scores. For both items, the examinee is given a choice of two prompts. For the Speaking Sample, the examinee must speak for a minimum of one minute with a maximum of three minutes. For the Writing Sample, the examinee has between three and five minutes to compose their response. The video recording of the examinee's spoken response and text of the written response accompany the numerical scores shared with results users. The Speaking Sample and Writing Sample thus arguably constitute test results in their own right—information about an examinee captured through a standardized assessment procedure and presented to a third party for interpretation. However, in this case the results are raw performance samples, and it is unclear to what extent and in what way they are interpreted and used by admissions officers.

4.1.4.2 Anticipated audiences of Duolingo English Test results

In O'Donnell and Zenisky's (2020) concentric circles of interest in test results, the examinee is usually at the center of the circle, representing the audience with the strongest interest in the test results and the desire for the most detailed and granular feedback. This is arguably the case for the Duolingo English Test as well. Those taking the test to apply to higher education institutions need to understand their scores in order to decide which schools to apply to based on their performance relative to institutional cut scores, whether it is worth retaking the test to improve their results, and how to prepare to retake the test. Thus the needs and abilities of examinees are of high importance when designing results reports.

Circles farther from the center in O'Donnell and Zenisky's (2020) model of interest in test results represent results report audiences increasingly distant from the test

taker and increasingly interested in coarser grained or aggregated results. One of these audiences is educational institutions, which use test results to make decisions about the test taker. In the case of the Duolingo English Test, this refers to higher education institutions, which use test results to determine if applicants meet the criteria for admission to a degree program. Admissions offices thus constitute an important results report audience, as their use of test results carry significant consequences for the test taker. The present research will focus primarily on the educational admissions audience of Duolingo English Test results.

4.1.4.3 Primary results report

The prototypical results report of the Duolingo English Test is the results certificate (Figure 4.1). The certificate begins with the picture and name of the examinee, as well as the date of the test administration for which results are reported. There is also a logo of the Duolingo English Test at the top of the page to indicate the provenance of the certificate. Moving down the page, the certificate next prominently displays the examinee's overall score in a comparatively large font size. The overall score is accompanied by a short explanation that this score represents "the test taker's ability to use English in a variety of modes and contexts." There is also a number line to position the overall score on the score scale and a set of can-do statements providing a qualitative interpretation of the examinee's English proficiency. Below this section on the overall score, there are four smaller sections with the examinee's four subscores: Literacy, Conversation, Comprehension, and Production. Each subscore is accompanied by a short explanation of what the subscore represents (e.g., "The test taker's ability to read and write") and a number line of the same design and proportions as that for the overall score. The certificate concludes with a link to the Duolingo English Test webpage about test score interpretation.

While the results certificate depicted in Figure 4.1 constitutes the quintessential results report most commonly discussed in the literature, it is not the only way that test results are communicated. All test takers with certified results (meaning they did not violate test rules or experience a technical issue that prevented proper test administration) will receive a results certificate, and the certificate is presumably the sole way in which test takers will learn the results of their test. However, in the case of results users such as admissions officers, there are multiple ways for results to be received, which can vary between and even within institutions. While some results users will see the results certificate, many receive scores through dashboards and Constituent Relationship Management (CRM) systems.

Figure 4.1. Example of the Duolingo English Test results certificate

4.1.4.4 Admissions officer view of scores

Arguably the greatest impact of high-stakes standardized test results arises from their use by individuals and organizations other than the examinee to make decisions about the examinee. In the case of the Duolingo English Test, the primary use of test results is currently for admissions to English-medium higher education programs, and thus the primary results user population is college and university admissions offices. Each institution that accepts Duolingo English Test results for admissions purposes has access to a dashboard that aggregates all test results shared with them by examinees. When admissions officials access the dashboard, they will see a list with one row for each test session (see Figure 4.2). Clicking on the down arrow (a) will open an expanded view of an examinee's results (Figure 4.3), clicking on "View" (b) will open the examinee's ungraded speaking and writing samples (Figure 4.4), and clicking on "View" (c) will open the examinee's results certificate (Figure 4.1). Dashboard users can also download all visible test session results in a single CSV file by clicking "Export" (d).

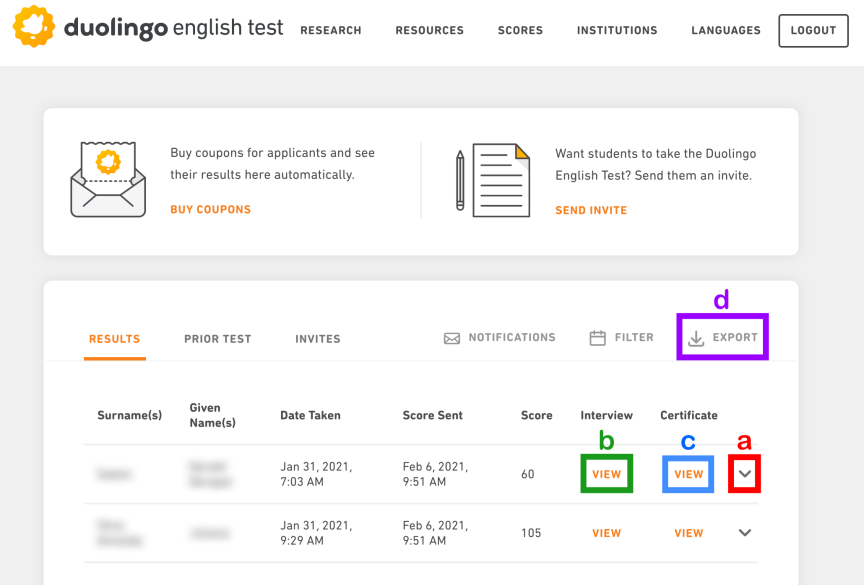

Figure 4.2. Example of institutional dashboard for accessing Duolingo English Test results

Although all accepting institutions have access to a Duolingo English Test dashboard, this is not to say that the dashboard is the primary way that admissions officers view test results when evaluating applicants. A 2014 survey responded to by 603 American higher education institutions found that 64% of responding institutions were at that time using a customer relationship management (CRM) system (American

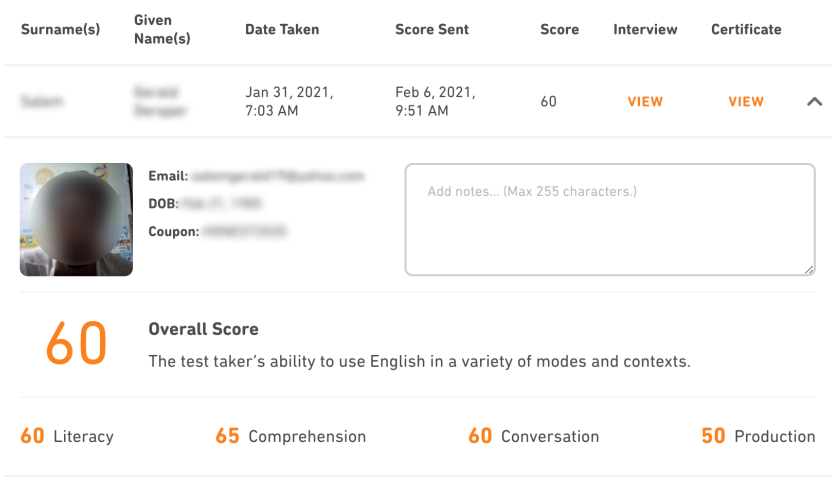

Figure 4.3. Expanded summary view of an examinee's Duolingo English Test results in an institutional dashboard

Association of Collegiate Registrars and Admissions Officers, 2014), a type of software used by an organization for keeping track of interactions with clients (Menard, 2020). The AACRAO survey found that recruiting and admissions were the most commonly reported uses of a CRM system by respondents. CRM systems allow educational institutions to organize all documents and information related to an applicant, assign applications to reviewers, and record all interactions with an applicant (de Juan-Jordan, Guijarro-Garcia, & Gadea, 2018). The CRM market in the United States is currently highly fragmented, with the largest CRM vendor holding 16% market share and many other vendors having less than 10% market share (Menard, 2020). The Duolingo English Test currently supports integration with a single CRM system, Slate, by the company Technolutions (Technolutions, 2021). Slate currently holds about 14% market share among CRMs used for admissions purposes (Menard, 2020).

For admissions officers at schools that use the Slate CRM, their view of Duolingo English Test results might look like that in Figure 4.5, with the results appearing under the tab "Duolingo" within an applicant's application, visible at the bottom left of the figure. This layout of the test results is essentially the same as that available in the Duolingo English Test dashboard (Figure 4.4). The only notable difference is that results users do not have direct access to the results certificate within Slate, and therefore will not see the visualizations of the overall score and subscores on a number line, the verbal descriptors accompanying the examinee's overall score, nor the link to additional information on score interpretation. Additionally, it is possible that institutions could customize the display of test results in Slate, meaning not all institutions necessarily have the same view as shown in Figure 4.5.

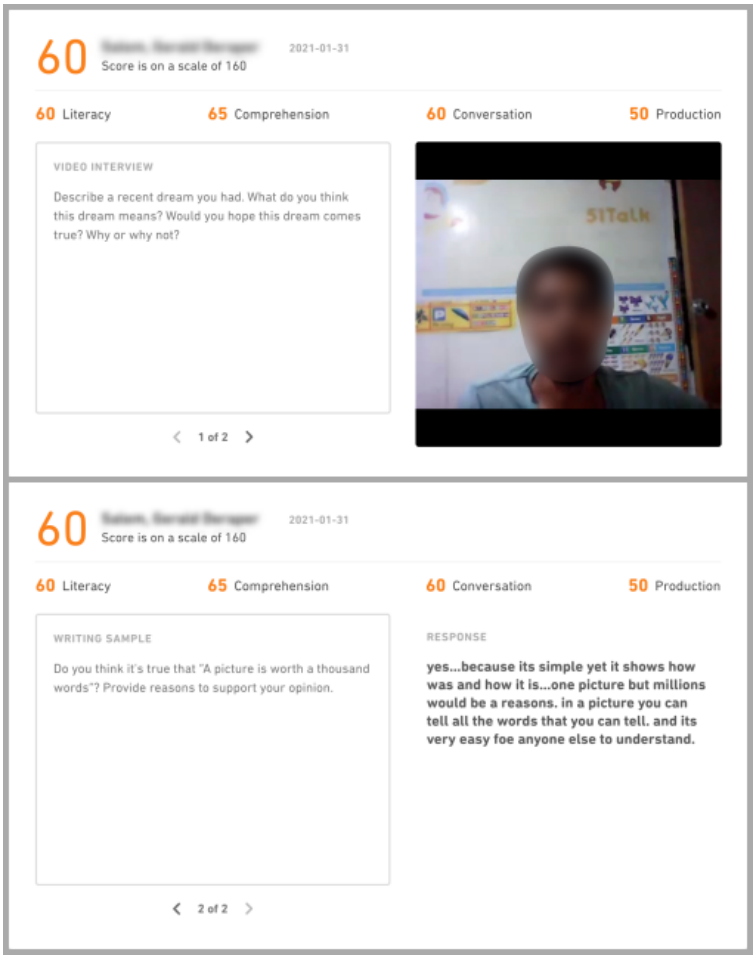

Figure 4.4. View of an examinee's ungraded speaking (top) and writing (bottom) samples

At schools that do not use Slate, there are theoretically several ways that admissions officers might see Duolingo English Test results. As mentioned above, each school has a dashboard that admissions officers could use to view an applicant's results. It is also possible to export results for multiple examinees in a CSV file, which can then be imported into another CRM or application system. And finally it is conceivable, and there is anecdotal evidence suggesting, that some schools might save each applicant's results certificate as a PDF and attach it to their application file. However, it is not known how prevalent each of these options are among schools that accept the Duolingo English Test, and thus the method by which results users see scores is a topic for investigation in the present research. It is also worth noting that if results users are obtaining results from the certificates, then these individuals are viewing the certificates of numerous examinees. However, results reporting research and

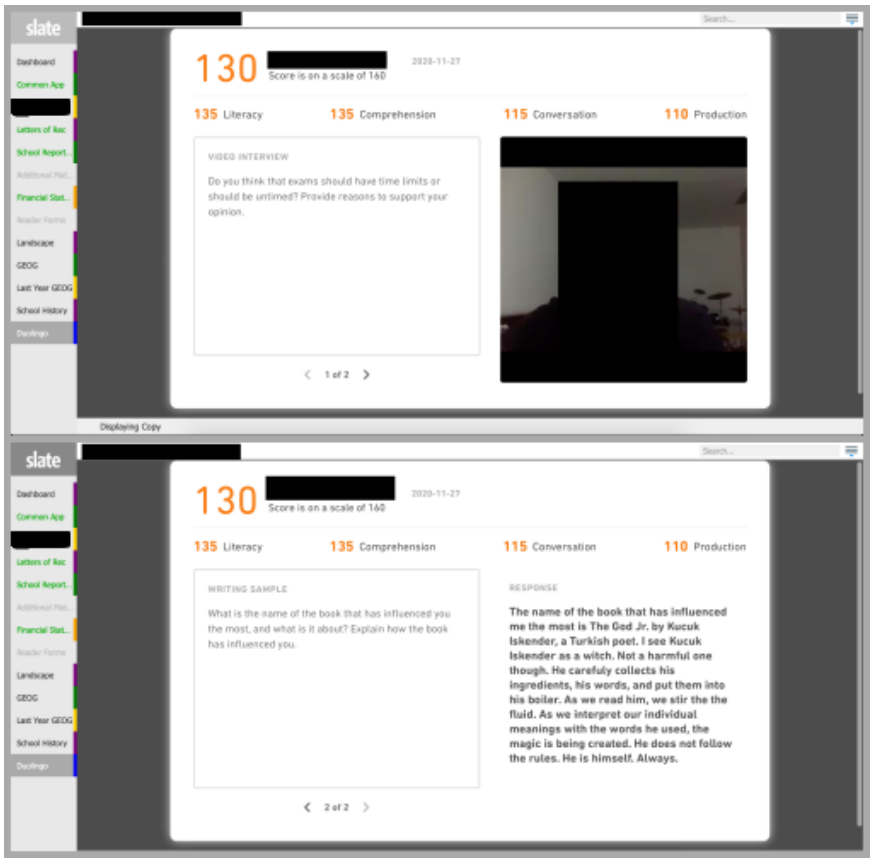

Figure 4.5. View of Duolingo English Test results in Slate

recommendations often discuss results reports in the context of being read carefully by a motivated reader for the first time. It is thus potentially worth investigating whether the results certificate format is effective for, and has the intended effect on, a reader who regularly interacts with the results report format.

4.1.4.5 Supplemental information to aid in results interpretation

All audiences of Duolingo English Test results have access to supplemental information to assist them in interpreting and using the results. The document *Duolingo English Test: Official guide for test takers* (Duolingo English Test, 2020) is a primarily test taker–oriented resource available for free online that contains information on score interpretation, sharing scores with institutions, and how institutions use results. Resources that are more appropriate for results users include the *Duolingo English Test: Technical Manual* (Cardwell, LaFlair, & Settles, 2021), which provides explanations of how the test is administered and scored, and also gives the percentiles of the total score and each subscore. There is also a research report (LaFlair, 2020) explaining

the rationale and procedure for developing the test's subscores. These resources could help results report audiences interpret and use test results, although it is unknown to what degree audiences access these resources and how they use them. The present research will seek to address these questions.

4.1.5 Models of score report evaluation and development

While general advice on the design of results reports is helpful, a fully developed model can provide a comprehensive framework for the principled development, evaluation, and revision of a testing program's results reports. The models proposed by Zenisky & Hambleton (2016) and Zapata-Rivera & VanWinkle (2010) are two such models of score report development. These models go beyond guidelines and general best practices of score report design by providing a structured process for developing and testing score report designs for a particular context based on audience needs. Both models prioritize the articulation of relevant audiences and their needs and preferences.

The model described by Zenisky & Hambleton (2016) comprises four phases: (1) "laying the groundwork for all report development efforts to follow" (p. 591) by reviewing considerations of results reporting inherent in the test's development, identifying results report audiences, assessing the needs of audiences, and conducting a review of relevant literature and documents; (2) developing one or more results report templates based on the outcomes of Phase 1; (3) field testing results report templates with audiences and iteratively revising them as necessary based on feedback; and (4) ongoing evaluation maintenance of the results report template in operational use. The present research focuses primarily on the first phase, particularly assessing the needs of the main audience of results reports—higher education admissions officers.

4.1.6 Research Questions

The present research seeks to answer the following three main research questions and corresponding sub-questions:

1. What are the known or supposed informational needs of North American admissions officers as a results report audience of high-stakes English language proficiency (ELP) tests used for admissions purposes?
    - What are the intended and unintended uses of ELP test results by this audience?
    - What other sources of relevant information (e.g., results from other tests) are available to each audience to inform their interpretation and use of ELP test results?
2. To what extent are current ELP tests' results reporting practices meeting their informational needs?

- Are different ELP tests' results seen as equally valid for supporting inferences about applicants' language abilities?
- Do DET results reports present the desired information in a way that is accessible (i.e., noticeable and comprehensible)?
- To what extent do results users understand and care about score reliability?
- Are results users utilizing the ungraded speaking and writing samples, and how are they interpreting these results?
- Do results users desire any additional information about individual examinees or subgroups of the examinee/applicant population (i.e., aggregate data)?
- Do results users desire a particular format for results presentation (e.g., static vs. interactive, succinct vs. rich detail)?
- How do results users feel DET results reporting compares to that of alternative tests?
- How can the unique characteristics of a digital-first testing program (e.g., web traffic/clicks) improve understanding of results report use to inform results report development?

3. What steps can be taken to improve DET results reporting?
    - Can any actionable conclusions be drawn from the answers to the previous two questions?
    - What additional research is needed to inform DET results reporting?

4.2 Methods

As suggested by the Zenisky & Hambleton (2016) model, understanding audience needs and preferences ultimately requires engaging with these audiences. Therefore, the research questions were addressed through a focus group conducted with a sample of 8 admissions officers from the United States and Canada.

4.2.1 Participants

Individuals currently working as admissions officers at postsecondary institutions in the United States and Canada were invited via email to take part in the focus group. Potential participants were identified through personal connections. Eight individuals took part in the focus group, all of whom were working at institutions that accept the Duolingo English Test for admissions purposes. At the beginning of the focus group, participants were asked to fill out a short survey about their professional background in higher education admissions. The survey contained the following questions:

1. How many years of professional experience do you have in higher ed admissions?
2. At how many higher ed institutions have you held an admissions-related role?

3. Which of the following descriptions apply to any of the institutions in which you have worked in an admissions-related role? -Type of institution: Doctoral university, Master's college/university, Baccalaureate college, Associate's college, Research-intensive, Liberal arts -Size/location of institution: Very small (<1,000 students), Small (1,000 ~ 2,999 students), Medium (3,000 ~ 9,999 students), Large (>10,000 students), Urban campus, Suburban campus, Rural campus, Canadian/UK university

4.2.2 Focus group procedure and questions

The focus group was conducted through Zoom and lasted approximately one hour. Participants discussed the following six questions related to the interpretation and use of ELP test results, both in general and specifically for the Duolingo English Test, spanning the process from adoption of a new ELP test to making decisions about applicants using test results.

1. In your experience, who are the stakeholders involved in deciding to accept an English proficiency test for admissions purposes, and on the basis of what information?
2. How do institutions determine the admissions cut score?
3. At what point in the process of reviewing an international application do you consider English proficiency test results? How much time might be spent on considering these results?
4. I'm going to show you example results reports of the Duolingo English Test. Which parts of the report do you think are most important for admissions officers to make a decision about an applicant? Is there any part of the report you think is not useful?
5. What conclusions about an applicant's abilities do admissions officers draw based on Duolingo English Test results?
6. What Duolingo English Test supplemental resources (e.g., DET technical manual, documents, webpages, video recordings, webinars etc.) do admissions officers utilize to inform their interpretation and use of Duolingo English Test results?

The questions were discussed sequentially and with minimal interruption or participation by the researcher, other than to introduce the subsequent question. Participants were invited to respond to each question and/or the responses of others, either orally or by writing in the Zoom chat as they preferred. An automated transcript of the meeting was produced by Zoom using built-in automatic speech recognition (ASR), and the transcript was later edited by the researcher to correct ASR errors (e.g., "duo lean go" –> "Duolingo") and misattribution of utterances, and to remove false starts and hesitation markers (e.g., "It's, you know, umm, it's just that…" –> "It's just that…").

4.3 Results

4.3.1 Participants

Participants' length of professional experience ranged from 8 to 26 years, with a mean and median of 16 years (see Figure 4.6).

```
0 | 8
1 | 11
1 | 667
2 | 1
2 | 6
```

Figure 4.6. Stem-and-leaf plot of participants' years of professional experience in postsecondary admissions

Three participants (38%) have only worked in admissions at one postsecondary institution, another three (38%) have worked at two institutions, and two participants (25%) have worked at three institutions. No two participants were working at the same institution at the time of the focus group. The focus group participants thus collectively represent 126 years of professional admissions experience at 15 postsecondary institutions in the United States (n=6) and Canada (n=2).

Figure 4.7 summarizes participants' professional experience by the characteristics of institutions they have worked at with respect to highest degree offered, size (i.e., enrollment), and setting (urban, suburban, or rural). Percentages of categories within a variable can add to more than 100% because participants were allowed to endorse multiple response options to reflect work experience at multiple institutions. Participants have worked primarily at doctoral and masters-level institutions, with an equal number reporting experience at research-intensive (50%) and liberal arts (50%) institutions. Participants have worked primarily at large institutions in urban and suburban settings. No participants had worked at medium (3,000~9,999 students) or very small (<1,000 students) institutions, nor had any worked at an institution in a rural setting. Thus, the results of the focus group are potentially most transferable to large, graduate degree–granting, DET–accepting postsecondary institutions in (sub)urban American and Canadian settings.

4.3.2 Focus group questions

In this subsection, the themes of participant responses to each focus group question are discussed. For each question, a table is presented connecting each extracted theme

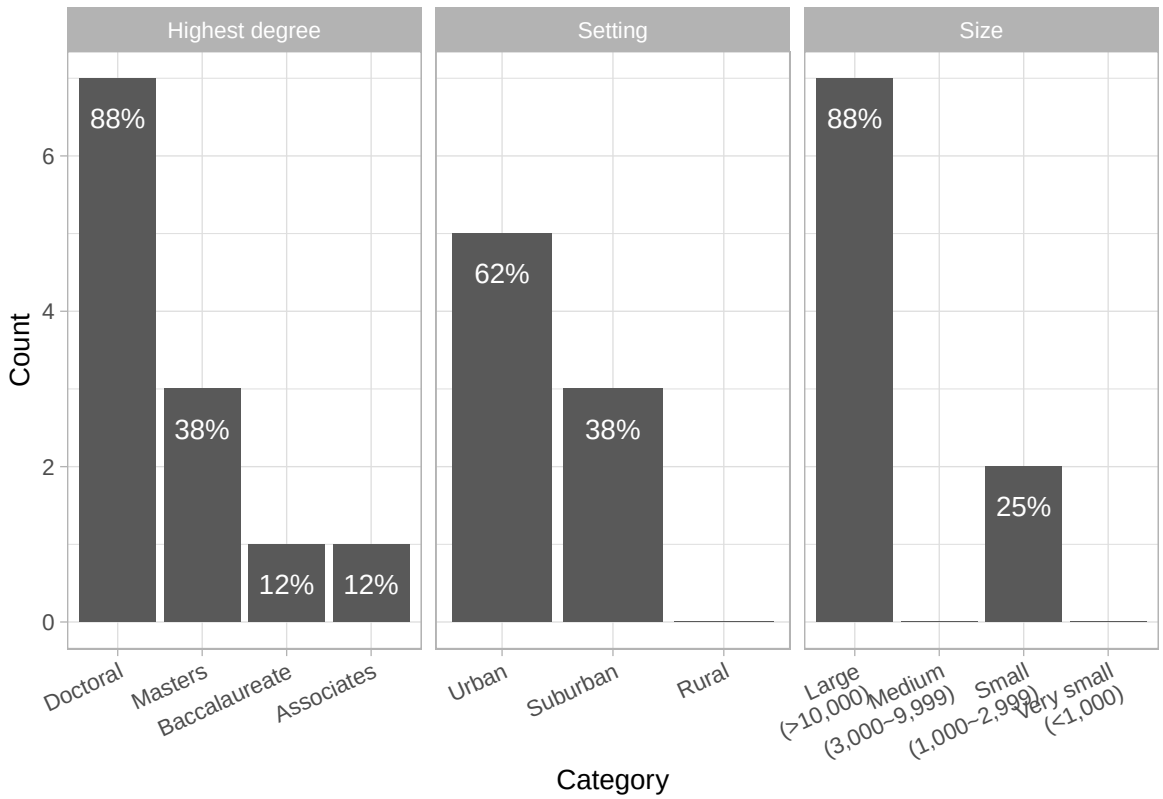

Figure 4.7. Highest degree and orientation of institutions at which participants have worked in admissions

with representative participant quotes. All extracted responses to each focus group question are presented in chronological order in the Appendix.

4.3.2.1 Question 1

> *In your experience, who are the stakeholders involved in deciding to accept an English proficiency test for admissions purposes, and on the basis of what information?*

The most salient theme in participant responses to the first question (see Table 4.1) was the high degree of variability across institutions in the process of deciding to accept a new ELP test. Some participants reported complete autonomy of the admissions office to decide to accept a particular test, while others discussed needing approval from the faculty senate or senior university leadership. But one participant asserted that in most (if not all) cases, the initiative to accept a new test starts in admissions. Multiple participants attributed inter-institutional differences in the approval process to public vs. private status, with private institutions' admissions offices having more autonomy to accept a new test and not needing to involve other stakeholders. Public institutions, meanwhile, were reported to require approval from stakeholders from outside the admissions office. The evidence considered in making the decision to accept a new test also seemed to depend at least in part on the stakeholders involved; one participant reported needing to present research and "quite a bit of data" to the faculty committee.

Table 4.1. Themes in participant responses to focus group question 1

Response Theme	**Representative Quotes**
Variability in degree of autonomy when deciding to accept a new test	• "US universities, [it's] typically ultimately a faculty governance decision." • "The autonomy with which my office operates is more or less absolute, like we just decide what we want to do." • "And then my previous institution, which was [a] large public research, [we] had to go all the way up to Provost staff for approval. But that authority is delegated at my current institution" • "University leadership and how they function and how much they're willing to delegate definitely impacts how the decision making happens."
Differences between public and private institutions	• "I would also say that's more for public institutions, because at a private, we wouldn't need to involve the faculty." • "I think regardless of whether or not you need faculty approval, it almost always starts with the admissions office or someone [in] enrollment management."
Evidence considered depends in part on stakeholders involved	• "I don't think we consulted with anyone when we decided to begin taking the DET, [we] were just like, Yeah, sounds like a good idea." • "that faculty committee, also then took in sort of our program staff's recommendations and research, and there had to be quite a bit of data presented as well, so looking at different kinds of validation success measures."
Distinguishing admissions standard and acceptable evidence	• "There's the difference between meeting an admission standard, and how do you evidence that standard. It might be DET, might be TOEFL, might be IELTS." • "...as the senior admissions officer, I can make that decision [to accept a new test]. I can't change the standard."
Importance of concordance between tests	• "I can't change the standard. So it had to have valid concordances between other accepted measures."

In addition to details of the approval process, another interesting theme in responses to Question 1 was the distinction between an admission standard (i.e., the level of language proficiency needed for admission) and standardized test results as evidence of meeting that standard. This perspective frames the decision to accept a new test as answering the question, "Can this test provide sufficient evidence that an applicant has met our language proficiency standard?" This perspective also emphasizes the view of ELP tests used for admissions purposes as interchangeable sources of evidence. Thus, as one participant reported, a new ELP test must "have valid concordances between other accepted measures."

4.3.2.2 Question 2

How do institutions determine the admissions cut score?

Participants mentioned at least six factors that an institution might consider when determining their admissions cut score(s) on a newly accepted ELP test (see Table 4.2). The first and arguably most straightforward factor is the practices of "competitor" institutions. Other institution-external factors include external data/studies, presumably those conducted by other institutions or by the test developer, as well as the reported concordance with other tests, a topic raised in response to the previous question. Multiple participants discussed the importance of concordances between accepted tests and how the concordance constrains or even dictates the choice of cut score. Participants felt that it is not possible to tell applicants and other stakeholders that multiple tests are interchangeable while using cut scores for the tests that contradict the publicly available concordances. Therefore, according to one participant, it is "pretty standard at many institutions" to base the cut score(s) for a newly-adopted test entirely on the concordance with already-accepted tests.

Table 4.2. Themes in participant responses to focus group question 2

Response Theme	**Representative Quotes**
What other schools do	• "By looking at what everybody else does." • "I think it's also looking at our direct competitors..."
Concordance with other tests	• "...based off of looking at concordance..." • "when we look to adopt any tests we, you know, look for the concordance and whatever that concords to, that's what we go with. It's certainly far from perfect, ...that's pretty standard at many institutions, sort of have to do that because we're providing general information to a whole group of people and if we accept multiple tests, then how could we justify having anything other than concorded scores?" • "...to me nothing else makes sense, so it's unfortunately still sort of driven by these other tests..." • "...we definitely wanted to try and keep them all as consistent as possible so it didn't seem like one test was more advantageous or easier than another..."
Leadership/institutional priorities regarding incoming class profile	• "...with a high bar we were cutting students who are performing very strongly in other areas, in testing, GPA, etc. and potentially kind of not matching up to what they wanted to see in terms of the incoming class profile, so at that point we kind of didn't necessarily lower the bar but we were just definitely a little bit more flexible in how we would decide who made the cut and who didn't. So I think that it's definitely kind of what leadership is saying, in terms of overall institutional priorities, will influence where the line is drawn on some of these things."
Faculty feedback	• ...quite influenced by faculty feedback, talking about the English proficiency in the classroom." • "...also considering our own programs and what the faculty is looking for."

Collecting sufficient data for internal analyses	• "Then we collected, I think, two years, two years of data, looked at student outcomes with that performance, primarily looking at GPA and retention for persistence, and we saw good data there." • "...others sort of said well if you don't kind of widen the scope enough, you're not going to get enough variance to really then study what's predictive and what's not, ...So we ended up opting for the second in this case, ...let's widen the bridge just a little bit so that we can get a little more variance and then that might really help our validity studies down the line to then hone in on do we need to adjust where the score is in the future years."
External data/studies	• "Our volume is such that we still have to rely on external data and reports in order to ensure that. We just didn't have the volume to ensure anonymity of current students when doing our own assessments."

In addition to such institution-external factors, participants also reported internal factors that could influence a test's cut score. Multiple participants reported taking faculty feedback into consideration, such as observations about the English proficiency of the students in their classes. Interestingly, one participant commented on "institutional priorities" concerning the profile of the incoming class, and while they explicitly said they did not "lower the bar" but were just "a little bit more flexible," it seems that the desire to achieve a certain class profile can impact how ELP test results are interpreted and used in the admissions process. Multiple participants also mentioned the need to accumulate data from matriculating students in order to evaluate a new test in terms of retention and academic achievement. One of these participants explicitly discussed the range restriction issue, indicating that their institution made the conscious decision to set a more lenient cut score at first (and thereby potentially admitting under-qualified applicants) in order to ensure a sufficient range of language proficiency for predictive validity studies that can inform an appropriate cut score in the long term.

While most participants indicated that their institution has a cut score/required minimum for the DET, one participant reported that their institution does not, saying the following:

> "we don't have a cut score either or minimum score ... we've started talking about it in terms of middle 50% similar to the SAT/ACT, so all

> of our English language testing we now report out the 25th and 75th percentile. And just to kind of keep them aligned with the way that we talked about an SAT or ACT."

This quote illustrates that not all institutions fit within the paradigm of explicit cut scores. And while such institutions may be in the minority, it should be noted that they exist, and that the way admissions officers at such institutions interpret and use ELP test results may differ categorically from those at institutions with cut scores.

4.3.2.3 Question 3

> *At what point in the process of reviewing an international application do you consider English proficiency test results? How much time might be spent on considering these results?*

According to participants, when and for how long English test results are considered by an admissions officer can depend on the application volume/bandwidth of the admissions office, the overall strength of the application, and the stringency of the institution's language proficiency requirement (e.g., whether a conditional offer of admission or exception to the policy are possible). For institutions that strictly enforce a required minimum score, ELP test results will be reviewed early in the process, and if an applicant has not met the requirement, the admissions officer will not spend much time reviewing the file (including ELP test results). Conversely, at institutions where conditional acceptances are extended, ELP test results might be the last application element to be reviewed. One participant explained that more time might be spent reviewing ELP test results when an applicant has a mix of positive and negative application elements. Collectively, these responses suggest that comparatively little time is spent reviewing results when the application will clearly be rejected or accepted, and relatively more time is spent considering results for "borderline" cases.

Concerning the review of Duolingo English Test results specifically, more time might be spent reviewing DET results, compared to those of other tests, because of the speaking and writing samples. Participants discussed reviewing applicants' speaking and/or writing samples to decide on borderline cases or to verify other evidence of English proficiency. Reviewing such samples inevitably takes more time than examining only numerical test scores.

Table 4.3. Themes in participant responses to focus group question 3

Response Theme	**Representative Quotes**
Factors affecting timing and duration of ELP test results review	• "We do offer conditional letters of offer for students who have not yet proven they've met our proficiency score, as long as their academics are such that we would admit them. So in that respect, we potentially could be later in the process." • "I think it's right in the beginning, if a student doesn't meet your minimum, you know, that's a clear indicator that this application is pretty much dead in the water, unfortunately, at that point. But once you're past that, you know, it's like they meet your minimums [...] I think how much time you spend on it, it's going to depend on all the other factors as well because it might be a no-brainer everything lines up, you know they're great so you spent two seconds, okay, but for our student who you have factors that are good and make you feel confident I think there's some others that might be making a little worried. That's where you might spend a little bit more time looking at and taking a deeper dive on it." • "[It depends on] whether a file emerges as competitive and compelling, you know, because if it doesn't, we're not getting a whole lot of time on it. Period." • "It may also be evaluated differently for different access points in the process. So for us, ...If you don't meet the score you don't meet the score and will be an automatic denial. We'll reach out and ask, or as [other participant] said, we'll look for other opportunities for the student to have demonstrated proficiency but if they do, they might meet the threshold for general or university admission, but they may not meet the language proficiency requirement for one of our more competitive programs or a conservatory program. So, there are differences there as well and I think that's very similar, that's, that's common in graduate admissions as well."

Interpreting ELP test results in context	• "But once you're past [ensuring they] meet your minimums and you're looking at things for them to see what's different in this course they've taken, do they have ACT scores, trying to get the whole picture of where you're trying to place their English proficiency level" • "I would say in addition to the score itself we also kind of look at the date on which it was taken, so if it was a much older score we might, you know, take it a little bit more with a grain of salt, that scores can go up over time, right, and then also looking at, you know, what's the languages spoken at home, what's the language medium of the school that they're attending, the language of the country they're coming from, you know, I think those things can play a role in providing context to the actual test results that we find." • "We've moved in the direction of comparing very carefully—you know, obviously it's not an edited careful writing sample—but we compare that and essays, and we've tried increasingly to copy to students files every email that they send us, because often their English, you know, may be nowhere near as good in these emails that they're sending to us. And it's really revealing, and so like triangulating those things. Again, this is not scientific, but that's a way in which we're able to use the DET that we are not necessarily able to use other English proficiency tests because we have, like the raw material, we have the primary source of, here's the writing, not just a score."
DET writing and speaking samples	• "the DET actually has been a little different than some of the other tests given the access to the interview [...] the DET and the ability, during a holistic admissions process we have made accommodations for students who we've seen, we've watched the video, we've reviewed the writing. And we've looked at that a little bit more intently than we had with other test scores where it's just purely a number or numbers, but it is pretty small. So, we're able to do that." • "...with the DET, because we have, you know, minimal though it may be, a writing sample..."

The DET speaking and writing samples are used in different ways. One participant suggested they use them to inform decisions of leniency or conditional admissions. Another described how the writing sample is used, in conjunction with other writing from the applicant such as the application essay and emails, to "triangulate" an applicant's writing ability. Multiple participants indicated that English test results are interpreted in the context of other information about an applicant—such as SAT/ACT scores, courses taken, native language, language of instruction, application essay(s), and even emails received from the applicant—in order to draw conclusions about the applicant's English proficiency. Collectively, these responses indicate that admissions officers are, at least some of the time, interpreting ELP test results in the context of the entire application package and demographic information about the applicant, and not relying solely on the test scores to draw conclusions about an applicant's language proficiency.

4.3.2.4 Question 4

> *I'm going to show you example results reports of the Duolingo English Test. Which parts of the report do you think are most important for admissions officers to make a decision about an applicant? Is there any part of the report you think is not useful?*

For this focus group question, participants were also shown images of the DET institutional dashboard (Figures 4.3 and 4.4), a CRM view (Figure 4.5), and the DET results certificate (Figure 4.1). Several participants reported rarely using the institutional dashboard. They might use it to access interview/writing samples or search for results for a time-sensitive application (e.g., an athlete). Therefore they find the sorting and searching functionalities useful. Most discussion was around the visibility of coupon codes used by test takers, with concern that this could cause subconscious bias. Knowing that a test taker used a coupon code was not seen as necessary or useful for application review.

It seems that all participants' institutions use CRMs for application review (multiple participants explicitly mentioned using a CRM, while no participants reported primarily using the DET institutional dashboard or results certificates to access applicants' DET results). According to participant responses, institutions can customize the appearance of test results, such that some do not have access to the DET speaking and writing samples within the CRM and must use the DET dashboard to access them. Some institutions just populate the numerical scores into a page that provides an overview of an applicant. It is therefore not possible to assume that all admissions officers across different institutions are seeing ELP test results displayed in the same way, and in fact the responses suggest this is decidedly not the case.

Table 4.4. Themes in participant responses to focus group question 4

Response Theme	**Representative Quotes**
Institutional dashboard not frequently used	• "Never used the dashboard, only occasionally for like an athlete, who's app I know is coming through, but we don't yet have. And I'm so happy to hear about this new sorting and searching that we can do on the dashboard, because it's been really hard in the past to try to find what I'm trying to find" • "So I don't use the dashboard, but we pull everything in using the API and our system so that [it] appears on our reader screen or in a student's file"
Institutional dashboard coupon code	• "So, the only piece that sticks out to me is the coupon section here. ...there's reasons you might want to know that, because maybe you want to know if the student, you know, did use it, but also, you know, are you biasing your— Now you know this student receives, you know, a waiver of their test or anything like that." • "But that was one thing that we were very adamant about removing from view for any of our readers, we didn't want anyone to know whether or not they had applied for a fee waiver, you can't see if they've applied for financial aid, you can't see things like this that if there was a coupon or whatnot, so we were adamant that we remove that."
CRM view speaking/writing samples	• "we'll see it [=speaking sample video] when we click on the link that we get, which obviously takes you to this page to watch the videos. I guess it's helpful in the sense that it's good to have all those scores there when you watch the video. But you could get rid of everything else and I would still find it useful because there's the video." • "We'll go to the dashboard if we wanted to pull up an individual student's video to watch." • "We do listen to every record and so we have it [=speaking sample video] there, as a part of the application, but whether or not we look at it— again because for us it's not a required piece of information, or required part of the application—it depends on the strength of the file."

CRM view speaking/writing samples (cont'd)	• "Does it slow down Slate? Is that why you guys don't load it into each file? What I hear you saying, you use it infrequently, but it was so easy for us to do, and I feel like, there it is at our fingertips." • "with the volume of applications we get, it's unlikely that we would go and look at the video. It's really only if we've got particular questions and...when you do have an unusual applicant you want to kind of dig into them, but for the most part, then it's like a quick check into dashboard, 'is a student okay? yes/no' and move forward. We don't go digging any deeper than that most of the time, and you're not toggling to the DET dashboard to check"
CRM view customizability	• "We do use Slate and we pull it in a different way so that just populates a table so we just see the score and the subscores. But because we tend to watch the videos on a very small number of students, so we don't actually use this particular view of it [with interview/ writing samples]" • "And I don't know if it is always like in the order that's on there, on the left-hand side [of the Slate view], for us it's the last thing too" • "everything is on the [CRM] dashboard and we kind of go look at it there."
CRM view date	• "The only thing that stands out to me looking at this particular view is the date. What is that date? Is the date that it was taken? Is it the date that it was shared?"
Results certificate	• "We have used [results certificates] a few times this year because it's a relatively new test, so our office is not super familiar with the test, [...] it's a nice visual reminder that these are point estimates, they're not exact scores, right? And so there can be that little error band that I think helps remind people of that, and then as well as the descriptors I think also helps people familiarize themselves with what [...] a score mean[s]. It was really a small handful of students for whom we're really on that border or there is a lot more of conversation, people want to dig in deeper where there was sort of contentious argument about a particular student."

Additionally, one participant noted that in the CRM view shown, which is essentially identical to the institutional dashboard view, there is a date in grey text above the speaking/writing samples, and it is not entirely clear what this date refers to. This is an example of stakeholder feedback on results report interpretability that can potentially lead to design modifications.

Among focus group participants, DET results certificates seem to be rarely (if ever) used when reviewing applications, since most participants reported using CRMs at their institutions. The certificates can be useful for training/familiarization purposes, particularly seeing the ability descriptors associated with different scores. The error bar on the number line representation was also discussed as a useful reminder that scores are estimates.

4.3.2.5 Question 5

> *What conclusions about an applicant's abilities do admissions officers draw based on Duolingo English Test results?*

In response to question 5, multiple participants reported thinking of the language proficiency requirement in an essentially binary way—is the applicant linguistically prepared to succeed at the institution or not? While one of these participants also brought up the possibility of providing "additional support" to applicants who are not yet "where they need to be" linguistically, indicating that their conceptualization of ELP for academic purposes is not strictly binary, they concluded their thought by saying that "it is kind of [a straightforward cut off]" (See Table 4.5). This perspective is consistent with research findings (e.g., Carlsen, 2018) and theoretical arguments (e.g., Chalhoub-Deville, 2003) that there is not a linear relationship between proficiency in the medium of instruction and academic success, but rather that students need only achieve a certain proficiency threshold to "unlock" access to academic content. Participants also described feeling an ethical responsibility to act in the interests of applicants by not admitting applicants who are unlikely to succeed due to a lack of sufficient language proficiency.

Table 4.5. Themes in participant responses to focus group question 5

Response Theme	**Representative Quotes**
Language proficiency binary inference	• "So [language proficiency is] a very different component of an application in that it's, 'can they do this or not?', in a much more black and white [way]." • "And in some cases it is a case of, Can we offer them additional support to help them get to where they need to be? it's not always just to kind of a straightforward cut off. But I think it is kind of."
Ethical responsibility	• "I think we have a kind of a responsibility to students, to think about ...whether they're going to get through four years at our institution and do well. [...] We just have that responsibility to make sure that they're going to land on their feet when they come to us." • "For me it feels like an ethical responsibility to some extent." • "Our faculty members are also relying on us [admissions officers] to ensure the students are able to keep up in the classes. To [other participant]'s points, we have an ethical responsibility to the students and process."
Different tests support same inferences	• "One of the things that we think about in the admission process is, is a student going to be successful at our institution and they're going to be able to, you know, to walk into a classroom and be comfortable and do well and interact with their classmates and their professors. [...] If you've got direct entry into a four year undergraduate program in the US and it's going to move fast, you're gonna have a lot of reading, you're going to have classroom interactions, you're going to have to be presenting, doing all kinds of things from the outset" • "...not to speak for this group, I think we all get the DET and then any English test in that same lens that we're describing here. It's not like, Oh, it was TOEFL, I would draw these conclusions, but Duolingo tasks I draw these conclusions. I get I don't speak for everybody. That's, you know, a lot of head nodding, we all could draw the same conclusions based on an English language proficiency test"

Different tests support same inferences (cont'd)	• "I agree with what you're saying, but also there is a distinction, and I suspect you agree with this as well, between the components of certain exams, and you'll see that in what their scores, what their sections of the exam actually are. And a distinct component [of] the DET is access to the interview, and the writing and so forth."
Unique inferences from speaking and writing samples	• "the students that get the attention in the language piece of an application are usually the students on the border. We're having to look at them for something else. And at that point it's when the DET results, particularly the interview, and as I think [other participant] had said, triangulating the pieces that we're seeing maybe in the subscores with other components of the application become helpful." • "...what we are getting from the DET, again, is this raw material that we don't get from other tests. Now, again, I'm admitting something about our process that is not particularly scientific, but, you know, here's me reading a spontaneous writing sample where I'm like, 'how sophisticated is the language that they're able to use off the top of their heads?' And also, I don't know, there's something to the way that they actually carry themselves, in this spoken part. And none of this is information that I can quantify. And all of this is information that I cannot get from other tests. So it is different to me in that way." • "...we've had a lot of push back at a senior level, about the potential bias for one way to interpret interviews and writing samples and the simple fact that we are not trained EL/TESOL professionals. We look at it with this kind of amateur eye and say, 'well this is fine to me'. ...It's why we definitely have a hesitation in our office about using that raw information. I think there's value to it at times, but as a kind of a standard metric for admission. Then you know we do it in, definitely more heavily on the actual scores and subscores and total scores."

Unique inferences from speaking and writing samples (cont'd)	• "a few years ago, would watch all of the interviews that we got from Duolingo, from Initial View, from Vericant. We would have our student tour guides watch all of them, and they would give a rating that essentially said, the student's going to be strong in the classroom, you know, whatever, on average, not going to be able to perform [...] But since then, what we've done— We very much felt uncomfortable with that because we were asking, you know, 18-, 19-, 20-year-old students to gauge English language proficiency, when they're studying engineering, or business, not language. And so we have since decided to do away with that. We really don't watch any of them. We'll watch them in select cases, and we mostly accept them to let students send them to us because we were worried [...] that if we suddenly stopped accepting them, would there be like kind of a customer service issue with kids trying to send them to us and being upset that they can't. So, but essentially we don't really watch them anymore. And we spend more time looking at the concordance between the subsections and our performance on campus, so pulling in GPA, pulling in... the grade for our first year English language courses, things like that." • "But when you're talking about such a narrow band of competitive applicants, like, I'm like, 'okay well everybody has a score in the same range and so what are the differences?' The scores don't help to illustrate those, and I don't know that they should right? And so we return to this very human, very imperfect sort of element. That's where we found ourselves." • "but [other participant] what I will say is not all institutions are looking at such a narrow band of students." • "Institutions are admitting them at a much higher acceptance rate than many of you. And so, when we're going to those components of the DET,... we would all choose to look or not look at those with, again, a very different lens. We're going to those components hoping that we're going to find 'Yes', if that's the case. But it is imperfect; it is absolutely imperfect and completely subject to human flaw."

Concerning Duolingo English Test results specifically, participants believed DET results can support the same general inferences about test-taker English proficiency as results from other ELP tests used for admissions purposes. However, one participant pointed out that "there is a distinction [...] between the components of certain exams, [...and] what their sections of the exam actually are." While the remark is not entirely unambiguous, it is potentially referring to the subscores of the DET—Literacy, Conversation, Comprehension, and Production (LaFlair, 2020)—which differ in construct from the subscores of other ELP tests commonly accepted for admissions purposes; tests such as TOEFL and IELTS still report subscores structured around the traditional language subskills of speaking, listening, writing, and reading. So at the finer-grained subscore level, it is arguable that DET subscores cannot be treated as interchangeable with subscores of other tests, and are necessarily used to support different inferences about linguistic subskills. Additionally, the unscored speaking and writing samples were mentioned as a unique aspect of the DET, and this topic generated substantial discussion among the focus group participants.

The DET's unscored speaking/writing samples were reported by participants to be seen as constituting additional information about an applicant that would not be available from alternative ELP tests, and that this information is potentially useful for unique inferences. As also mentioned in response to Question 3, some participants talked about using DET unscored samples to "triangulate" an applicant's language proficiency—comparing the samples with the numerical scores, grades in relevant previous courses, other test results, and even emails the applicant might have sent to the school. In this case, an admissions officer is not so much using the samples for a categorically different inference about the test taker, but to substantiate and contextualize the inference from the numerical scores. However, participants also mentioned considering the perceived sophistication of spontaneous, unaided written production, as well as how the applicant can "carry themselves" in the speaking sample. In these latter cases, the target inferences are potentially deviating from those supported by the numerical test results and supported by validity evidence. Other participants were quick to point out the potential for bias and unfairness.

Multiple participants expressed concern about potential bias from having raw performance samples interpreted by individuals with no background in language assessment. One participant reported that due to this concern, their institution puts much greater weight on the numerical scores. Another participant reported that in the past, their institution had undergraduate student employees watch and rate applicant videos, but that this approach was eventually stopped due to concerns of bias. Now, their institution "essentially [doesn't] watch them anymore." However, other participants maintained that the unscored samples can sometimes provide a necessary way to differentiate applicants with otherwise comparable applications, prompting another

participant to remark that "not all institutions are looking at such a narrow band of students." This latter comment underscores the fact that postsecondary institutions differ in attributes such as selectivity and volume of international applications, which plausibly impact how the institutions will interpret and use raw language performance samples.

4.3.2.6 Question 6

> *What Duolingo English Test supplemental resources (e.g., DET technical manual, documents, webpages, video recordings, webinars etc.) do admissions officers utilize to inform their interpretation and use of Duolingo English Test results?*

Concerning the use of supplemental resources about the DET such as white papers, webpages, and webinar recordings, a prominent theme in participant responses was that they largely do not use such resources during full-scale application review. Rather, supplemental resources were reported as useful primarily for two purposes. One is for issues associated with initial acceptance of the test such as deciding to accept it and setting a cut score (see also Questions 1 and 2). The other is for training admissions officers to interpret results of the test when evaluating applications. One participant explained that this "pre-review preparation training" takes place "maybe over the summer" (Table 4.6). Another participant further explained someone in the admissions office such as "the international admissions lead or training lead" will typically "distill a lot of this stuff into, probably, a paragraph or a few sentences of what an admissions officer would need to know." These responses suggest that, at least at some institutions, most admissions officers never personally look at the majority of supplemental resources for an ELP test, instead relying on the distilled information produced by a colleague.

During discussion of this question, a participant mentioned again the importance of concordances between accepted ELP tests, which was a theme in several preceding questions. This further underscores the perceived importance of such concordances given their role in multiple aspects of the admissions process. The participant asserted that concordances are one of "the most frequently used" supplemental resources. It is conceivable that for the majority of admissions officers, who do not personally study the supplemental resources of an ELP test, the concordances constitute a heuristic for understanding the meaning of scores from a new ELP test by associating them with score points on a scale with which they are likely more familiar.

Table 4.6. Themes in participant responses to focus group question 6

Response Theme	**Representative Quotes**
Supplemental materials not accessed during application review	• "I will say from my perspective that not really anything. ...we've already set our benchmark of what we're looking for. So I'm generally not going back to any of those additional resources. I feel like they're useful but not necessarily as I'm using the actual DET [...] So there's very little that I would necessarily be using from the resources, as I go through the individual evaluation process for an applicant." • "but as soon as it reaches the evaluation and selection stage that's not something that, I mean, I'd never watch a video recording or webinar about it at that point." • "But otherwise, you know, those are probably going to ... link to a website if they have any other questions, or they'll come to whoever the sort of main lead is if they needed to dive deeper into other technical resources. But that typically doesn't happen during the review phase; it's usually, you know, maybe over the summer, sort of pre-review preparation training, that kind of thing."
Supplemental materials used for standard setting and training	• "I would say that all of that information is really useful when you're trying to set a score minimum" • "The only other time that we use it is as training purposes for new stuff." • "I would say it's extremely useful for maybe whoever the international admissions lead or training lead is, and often what we do is we will distill a lot of this stuff into, probably, a paragraph or a few sentences of what an admissions officer would need to know when reviewing the file because they've got, you know, 50 other things that they're also paying attention to. But it is helpful for us to establish those initial summaries."

Supplemental materials used for standard setting and training (cont'd)	• "Keeping in mind accreditation audits and all the administrative changes that we've mentioned and how subject all of this is to that, it's kinda, even though we're not using it every day, and it was very helpful initially, you have to have it there. And you have to stay on top of it, it can't be stale data. ...if you think of the test as the movie that the kids want to see, well, this is the theater that needs to hold them. From a training perspective as well, it's giving offices the opportunity not to create/recreate, and to have to keep up with, our own training manuals. It is available from the source's mouth itself. So that's always helpful and, you know, it's a lot of turnover from recruitment staff to admission staff, and just having the information and knowing the direct source to get it from, from a training perspective, saves a lot of time from our end."
The importance of concordances	• "...I would say probably the most frequently used, of course, the concordances and subscores or explanations. I think those are probably the two things we use the most."
Appreciation for clear supplemental materials	• "I want to use this opportunity to give you guys all a compliment in saying that everything is presented in such a clear way that I feel like [...] we don't have to do a whole lot of training because it's so easy to understand what you're putting in front of us. So thank you for that." • "And from a recruitment perspective, for me to be able to dive in a little bit deeper and have some more background information rather than just saying to a student like here's the minimum admission requirements that we're looking for for your English proficiency based on the tests that you're presenting. I'm now able to go and do my own research and my own understanding without relying on my admission staff to give that information to me so it allows me to be more autonomous and just more proactive with my learning, and this is something that I encourage my staff to do because we're not directly involved with all of the hard paper and scores that we get. [...] clear, concise, easy to digest information, and just the webinars, has been phenomenal for learning."

4.4 Discussion

Each of the original research questions are discussed separately below in light of focus group participant responses. While the results are based on a small convenience sample, they potentially transfer to experienced admissions professionals at similar North American institutions.

4.4.1 Research Question 1

> *What are the known or supposed informational needs of North American admissions officers as audiences of score reports of high-stakes English language proficiency (ELP) tests used for admissions purposes?*

In response to multiple focus group questions, participants referenced ethical concerns and obligations, giving the impression that they are highly motivated by a sense of ethical and professional duty to make decisions that are in the best interests of both the applicants and their institutions. Additionally, participants indicated that they think about the ELP construct in an essentially binary way—either an applicant has the linguistic ability to undertake studies in English or they do not. Given these two points, a fundamental informational need of North American admissions officers is understanding the likelihood that an applicant possesses the required minimum English proficiency to undertake studies at the institution and in the program in question. This informational need requires both demonstrating that the test is generally valid for this intended inference and presenting individual test taker's results in a way that facilitates accurate interpretation and use.

In addition to a binary, procedural inference about an applicant's ability to handle English-medium postsecondary studies, several focus group participants also discussed the use of DET unscored speaking and writing samples to differentiate applicants with otherwise similar profiles. This indicates that for some results users, there is a need for additional information about applicants that is not necessarily language-related. Participants also discussed using the speaking and writing samples to judge how an applicant "carries themselves" and to "triangulate" an applicant's language proficiency by comparing to other elements of the application or even to written communication with the applicant. In these cases, the results users are potentially using language performance samples to draw inferences other than those for which the numerical test scores were intended and validated.

Related to the use of unscored language samples, participants also discussed the use of other application components—such as SAT/ACT scores, secondary school grades, and demographic information about the applicant and their place of origin—to draw conclusions about an applicant's language proficiency. This suggests that

some results users are not relying on test scores alone to evaluate an applicant's proficiency, but are looking for more context to add nuance. For example, studies have shown significant positive correlations between tests of theoretically unrelated constructs such as writing and mathematics (e.g., Stricker, 2004), so perhaps results users are considering SAT/ACT scores to disaggregate language proficiency and general academic or test-taking abilities. Alternatively, results users may be using applicant demographic information to draw conclusions (even subconsciously) about an applicant's opportunity to learn English (e.g., due to characteristics of their home country's education system or their socioeconomic status) and thereby interpreting ELP test results from the perspective of language-learning aptitude or noncognitive attributes such as motivation. The only conclusion that can be drawn is that more research is needed to elucidate what additional and unique information about language proficiency score users feel they are getting from sources such as general academic test scores and applicant demographics, and how they use such information to inform their interpretations.

While this section has thus far discussed information that admissions officers need as results users, participants also discussed information that they do not need. At multiple points, participants raised concerns about the potential misuse of information. First, participants drew attention to the coupon codes visible in the DET institutional dashboard. The overwhelming consensus among participants seemed to be that this information should not be visible, or at least institutions should have the option to hide it by default, as knowing that a test taker used a coupon code (a method of providing DET test takers with a fee waiver) could potentially bias an application reviewer. Participants also discussed concerns about the use of the DET unscored speaking and writing samples, noting that admissions officers are not trained in language assessment but are left to draw their own conclusions from raw language performance samples (or even engage students to rate the samples), thus opening the door to possible inconsistency and bias. These responses indicate that participants were aware of their own limitations in terms of training and implicit biases, and do not endorse the idea that "more is better" with respect to information about test takers. This is a positive finding in that it can possibly empower test developers to selectively withhold from results users information with great potential to be misused.

4.4.2 Research Question 2

> *To what extent are current ELP tests' results reporting practices meeting their informational needs?*

To the extent that participants were aware and willing to report in a focus group setting, the ELP tests their institutions accept (including the DET, IELTS, and TOEFL) seemed to be meeting the basic informational need of evaluating whether an

applicant has most likely attained the minimum required level of general language proficiency. Concerning this basic inference, participants did not report making a distinction between tests with respect to their validity in supporting the inference or the constructs represented by the numerical test scores. However, as noted in the Methods section, participants were all employed at DET-accepting institutions, and thus are likely to hold a favorable view of the DET. It is entirely plausible that admissions officers at institutions that do not accept the DET hold different views of the interchangeability of the DET and other ELP tests. Nevertheless, it seems plausible that admissions officers at all institutions generally view and treat accepted ELP tests as interchangeable, since to do otherwise could introduce bias based on the ELP test an applicant has taken.

Participants did not mention any difficulties in comprehending or interpreting DET numerical scores as they are presented in DET results reports. However, participants did report using DET unscored speaking and writing samples to support inferences about applicants' language ability that they cannot make based on other ELP tests. They also raised concerns about their ability to interpret such raw language samples without guidance or training, suggesting a potentially unmet informational need.

At multiple points during the focus group, at least some participants demonstrated considerable understanding of the statistical/research methods component of language assessment literacy (LAL; Kremmel & Harding, 2020) by discussing concepts such as language proficiency as a latent variable, the impact of reduction of range on score variance, and the fact that scores are inherently estimates subject to measurement error. However, most participants did not express a deep understanding, nor a concern, for topics such as score reliability or classification consistency (potentially due to lack of opportunity to express such thoughts). It is nevertheless worth noting that participants described other factors besides their own language assessment literacy that could lead to less than theoretically ideal interpretations and uses of ELP test results. One participant mentioned institutional priorities with regards to an incoming class profile, which could influence an admissions office to be more lenient in applying an ELP requirement. But more than any other factor, participants emphasized the importance of concordances between accepted ELP tests, and how they feel constrained to set admissions cut scores that do not contradict these concordances (e.g., the required cut score for TOEFL should concord with the required cut score for the DET). From this perspective, a robust, defensible, and up-to-date concordance between tests is a primary informational need for all institutions that accept multiple ELP tests for admissions purposes.

Collectively, these findings underscore the complexity of identifying and meetings stakeholders' informational needs. The message that test developers want to convey, and the critical information that test users need for decision making, should ideally be

linked explicitly to choices about data manipulation and graphic elements to include on reports. The design of results reports and supplemental materials should be draw on the rich literature of graphic design and data visualization, such as the well-known work of Tufte (Tufte, Goeler, & Benson, 1990; Tufte & Robins, 1997), and also reflect the realities of the intended audience, such as the widespread use of CRMs in admissions offices. Then usability studies of various approaches should be employed to determine how stakeholders are (mis)using the provided information, which then leads to iterative revision of results reports and other materials as prescribed by Zenisky & Hambleton (2016). Clearly, such an undertaking requires more than just psychometric or linguistic expertise, but rather necessitates the collaboration of professionals from various fields (Chalhoub-Deville & O'Sullivan, 2021) such as graphic design and user experience research.

4.4.3 Research Question 3

What steps can be taken to improve DET results reporting?

There are some immediately actionable findings from the focus group. In particular, focus group participants called for coupon codes to be hidden or hideable in the institutional dashboard to mitigate implicit bias, and for the meaning of the date written above test results to be clarified. These findings demonstrate the utility of investigating results interpretation and use among intended audiences, as doing so can reveal unanticipated issues and inform substantive changes to address them.

Other findings do not clearly indicate the need for a specific change in DET results reporting practices, but rather reveal topics requiring additional investigation, including the display of test results within CRMs, the use of supplementary materials, and the use of unscored speaking and writing samples. Potentially many admissions officers do not look at DET results certificates (Figure 4.1) on a regular basis, and so efforts to promote appropriate results use through the design of this results report could be ineffective. Furthermore, many admissions officers review ELP test results within a CRM, which allow for some (unknown) degree of customization of how the test results are displayed. Additional research is necessary to understand the variations in how test results are displayed within CRMs, and how differences in this display could impact results interpretation and use. Similarly, the focus group revealed that, at least in some institutions, supplementary materials (e.g., white papers, webinar recordings) about an ELP test are only reviewed by one or a few members of the admissions office, who then relay a summary of the information to their colleagues. The information that most admissions officers receive about a given ELP test is therefore subject to the interpretation of someone who is most likely not an expert in either language or assessment. It could be illuminating to research how the information in supplemental materials is summarized and relayed to admissions officers. The findings of such

research could potentially inform the creation of more targeted materials for training admissions officers, to ensure they are receiving accurate and theoretically sound information about a test.

A topic that generated substantial discussion among participants was the use of the DET unscored speaking and writing samples. Participants saw these samples as a unique and valuable element of the DET that allow for greater insight into an applicant. But they also raised concerns about the potential misuse of these samples, given that admissions officers are generally not trained extensively in language assessment. These findings suggest that there is considerable room for improvement in the results user experience of using the unscored samples. Additional research is needed to determine how misuse of the samples can be minimized while maintaining their added value for results users.

4.4.4 Limitations

As discussed previously, the present study was based on a relatively small convenience sample of admissions officers from schools in the United States and Canada. Additionally, the participants had a substantial amount of professional experience in postsecondary admissions (min = 8 years, mean = 16 years), have worked primarily at large urban and suburban institutions, and all work at institutions that already accept the Duolingo English Test. Therefore it cannot be assumed that the responses of focus group participants are representative of all admissions officers in the United States and Canada, much less other regions of the world. Future research could employ a larger and more representative sample of North American or global admissions officers. Additionally, the present study did not endeavor to assess the accuracy of admissions officers' interpretations of ELP test results. Future research should also account for, and investigate, the possibility that results users are interpreting the results incorrectly. Finally, the breadth and depth of topics that could be discussed were limited due to the time constraint of one hour. Future research should allow for more time and/or address a narrower scope of topics.

4.5 Conclusion

Understanding the needs and realities of results report users—in this case, admissions officers—is crucial in effectively promoting valid test results use. Focus group participants demonstrated LAL by discussing ELP as a latent ability and acknowledging measurement error in test scores. However, use of test results involves a constellation of factors, including LAL and external forces such as time and pressure from other stakeholders. Additionally, prototypical results reports and supplemental materials might be rarely accessed. If found to be widespread, these phenomena must be

accounted for in efforts to promote valid ELP test use in postsecondary admissions. In particular, test developers should respond to the widespread use of CRMs by developing CRM-compatible displays of test results that leverage the extensive literature on graphic design and data visualization. Additionally, this study demonstrates how investigating the informational needs, priorities, and abilities of results report audiences can lead to concrete changes to results reports, as well as reveal new topics in need of research in order to continuously monitor and improve the efficacy of results reports as suggested by Zenisky & Hambleton (2016).

4.6 Appendix: Focus group question responses in chronological order

4.6.1 Question 1

- "US universities, [it's] typically ultimately a faculty governance decision."
- "I would also say that's more for public institutions, because at a private, we wouldn't need to involve the faculty."
- "The autonomy with which my office operates is more or less absolute, like we just decide what we want to do. I don't think we consulted with anyone when we decided to begin taking the DET, [we] were just like, Yeah, sounds like a good idea. And I think one important distinction there, though, is that we don't require any English proficiency test."
- "We require it though. And, like I made the decision, myself and my team, we made the final decision on accepting DET, but we require something."
- "There's the difference between meeting an admission standard, and how do you evidence that standard. It might be DET, might be TOEFL, might be IELTS."
- "I think regardless of whether or not you need faculty approval, it almost always starts with the admissions office or someone [in] enrollment management. The faculty, they aren't going to come to admissions generally."
- "...that faculty committee, also then took in sort of our program staff's recommendations and research, and there had to be quite a bit of data presented as well, so looking at different kinds of validation success measures. It was probably on the order of maybe one or sometimes even two year process to change what we accept."
- "And then my previous institution, which was [a] large public research, [we] had to go all the way up to Provost staff for approval. But that authority is delegated at my current institution so that as the senior admissions officer, I can make that decision.... I can't change the standard. So it had to have valid concordances between other accepted measures."
- "a nuance here is, is it being accepted permanently, or is it just to be accepted for a temporary basis. I think many universities skip the going to the highest levels that [said] look, for 2021, we need the Duolingo English Test or whatever tests, we're going to take [it] to meet the needs that we can fill our class. ... Whether it sticks from there, when that emergency ends maybe it goes away. I think that's something that we're seeing happen right now."
- "University of leadership and how they function and how much they're willing to delegate definitely impacts how the decision making happens."

4.6.2 Question 2

- "By looking at what everybody else does."
- "We determined that we were going to do a 120 overall, and no subsection below 110 for Duolingo. And that was based off of looking at concordance. . . . Duolingo ...were able to provide subsection scores for tests that didn't actually have subsection scores, so that we were able to... backtrack to see what we had been accepting to get a sense for what we would be accepting as our actual policy."
- "I think a lot of us in here, we're using the test initially as an ancillary admission measure, it wouldn't be necessarily the only thing that we'd be basing proficiency on. Then we collected, I think, two years, two years of data, looked at student outcomes with that performance, primarily looking at GPA and retention for persistence, and we saw good data there. I have no idea when [we] started using a TOEFL score of 79, but that's been our standard, and because of that when we look to adopt any tests we, you know, look for the concordance and whatever that concords to, that's what we go with. It's certainly far from perfect, ...that's pretty standard at many institutions, sort of have to do that because we're providing general information to a whole group of people and if we accept multiple tests, then how could we justify having anything other than concorded scores?"
- "The minimum that we require, to me nothing else makes sense, so it's unfortunately still sort of driven by these other tests, or at least that's the perception. Um, but yeah we handle things pretty much the same way that [others] just described."
- "there was a little bit of squishiness to it, right, because not all the concordances are exact, and so if there's a range, we did have to have those conversations, if it was well, this concords to a 92 to 96, what do we say is the score? ... we definitely wanted to try and keep them all as consistent as possible so it didn't seem like one test was more advantageous or easier than another, but also tried to keep it easy to remember too, ... so we might do a little bit of rounding, and then we had to make those decisions, do you pick 90 or 95? . . . [there] really was no other way to do that without any previously existing scores to do any research on."
- "we don't have a cut score either or minimum score ... we've started talking about it in terms of middle 50
- "I wonder if I'm going to answer every question the same way, but leadership is something that kind of that influences how we come up with that and an institutional priority ...was quite influenced by faculty feedback, talking about the English proficiency in the classroom. And as a result, then we kind of raise the bar a little bit and put some extra steps in place to evaluate students English

proficiency. And then the second one was probably a little bit more concerned with the fact that with a high bar we were cutting students who are performing very strongly in other areas, in testing, GPA, etc. and potentially kind of not matching up to what they wanted to see in terms of the incoming class profile, so at that point we kind of didn't necessarily lower the bar but we were just definitely a little bit more flexible in how we would decide who made the cut and who didn't. So I think that it's definitely kind of what leadership is saying, in terms of overall institutional priorities, will influence where the line is drawn on some of these things."

- "some people were much more risk averse, and so they said if you're going to take a new test, we want to be very strict with where we draw the line, and others sort of said well if you don't kind of widen the scope enough, you're not going to get enough variance to really then study what's predictive and what's not, ...So we ended up opting for the second in this case, ...let's widen the bridge just a little bit so that we can get a little more variance and then that might really help our validity studies down the line to then hone in on do we need to adjust where the score is in the future years."
- "our volume is such that we still have to rely on external data and reports in order to ensure that. We just didn't have the volume to ensure anonymity of current students when doing our own assessments."
- "I think it's also looking at our direct competitors, and ...different types of programs being offered at our university, there's such a wide range of programs. So just kind of figuring out what's going to meet for every program and working with the faculty as well. It's sometimes grades only determines that, and then there's sometimes, you know, the grades plus component. Our university just went through an entire English proficiency policy change. So that has been very different for us. So, it's looking at competitors but also considering our own programs and what the faculty is looking for."

4.6.3 Question 3

- "the DET actually has been a little different than some of the other tests given the access to the interview. We do offer conditional letters of offer for students who have not yet proven they've met our proficiency score, as long as their academics are such that we would admit them. So in that respect, we potentially could be later in the process. And then as far as the time spent, again the DET and the ability, during a holistic admissions process we have made accommodations for students who we've seen, we've watched the video, we've reviewed the writing. And we've looked at that a little bit more intently than we had with other test scores where it's just purely a number or numbers, but it is pretty small. So, we're able to do that."

- "I hate to say this but I would have to go with a good old fashioned, it depends. I think it's right in the beginning, if a student doesn't meet your minimum, you know, that's a clear indicator that this application is pretty much dead in the water, unfortunately, at that point. But once you're past that, you know, it's like they meet your minimums and you are looking at things for them to see what's different in this course they've taken, do they have ACT scores, trying to get the whole picture of where you're trying to place their English proficiency level. You know, that's where it comes into play. So, I think how much time you spend on it, it's going to depend on all the other factors as well because it might be a no brainer everything lines up, you know, you know they're great so you spent two seconds, okay, but for our student who you have factors that are good and make you feel confident I think there's some others that might be making a little worried. That's where you might spend a little bit more time looking at and taking a deeper dive on it."
- "I would say in addition to the score itself we also kind of look at the date on which it was taken, so if it was a much older score we might, you know, take it a little bit more with a grain of salt, that scores can go up over time, right, and then also looking at, you know, what's the languages spoken at home, what's the language medium of the school that they're attending, the language of the country they're coming from, you know, I think those things can play a role in providing context to the actual test results that we find."
- "[It depends on] whether a file emerges as competitive and compelling, you know, because if it doesn't, we're not getting a whole lot of time on it. Period. But for those that do with the DET, because we have, you know, minimal though it may be, a writing sample. We've moved in the direction of comparing very carefully—you know, obviously it's not an edited careful writing sample—but we compare that and essays, and we've tried increasingly to copy to students files every email that they send us, because often their English, you know, may be nowhere near as good in these emails that they're sending to us. And it's really revealing, and so like triangulating those things. Again, this is not scientific, but that's a way in which we're able to use the DET that we are not necessarily able to use other English proficiency tests because we have, like the raw material, we have the primary source of, here's the writing, not just a score."
- "It may also be evaluated differently for different access points in the process. So for us, ...If you don't meet the score you don't meet the score and will be an automatic denial. We'll reach out and ask, or as [other participant] said, we'll look for other opportunities for the student to have demonstrated proficiency but if they do, they might meet the threshold for general or university admission, but they may not meet the language proficiency requirement for one of our more competitive programs or a conservatory program. So, there are differences

there as well and I think that's very similar, that's, that's common in graduate admissions as well."

4.6.4 Question 4

4.6.4.1 Institutional dashboard

- "Never used the dashboard, only occasionally for like an athlete, who's app I know is coming through, but we don't yet have. And I'm so happy to hear about this new sorting and searching that we can do on the dashboard, because it's been really hard in the past to try to find what I'm trying to find ..., because we don't decision out of the CRM that we use. So, I also think admissions officers, however scores are reported, no matter the test, you become very familiar with it."
- "So, the only piece that sticks out to me is the coupon section here. ...there's reasons you might want to know that, because maybe you want to know if the student, you know, did use it, but also, you know, are you biasing your— Now you know this student receives, you know, a waiver of their test or anything like that. So that's the only kind of a red flag or yellow flag that kind of makes me go, should that be on here? should it not be on here?"
- "So I don't use the dashboard, but we pull everything in using the API and our system so that [it] appears on our reader screen or in a student's file. But that was one thing that we were very adamant about removing from view for any of our readers, we didn't want anyone to know whether or not they had applied for a fee waiver, you can't see if they've applied for financial aid, you can't see things like this that if there was a coupon or whatnot, so we were adamant that we remove that. That's probably a much bigger deal for need-sensitive need-aware institutions, than those of us [that are need-] blind. Because at the point that a student would be invited to apply for any merit based scholarship with us, none of that would be pertinent information... But, again, it depends. That's all, how the admissions operation is structured"
- "the only reason that it [=coupon code] might be pertinent is if the student were requesting an application fee waiver. But we grant those on an honor system... Our take is, we're not going to burden them anymore, we just grant it."
- *Message in chat*: "Overall score is what our admissions officers look at for our undergraduate programs."

4.6.4.2 Consitutient Relations Management (CRM) software

- "we'll see it [=video interview] when we click on the link that we get, which obviously takes you to this page to watch the videos. I guess it's helpful in the

sense that it's good to have all those scores there when you watch the video. But you could get rid of everything else and I would still find it useful because there's the video."

- "We do use Slate and we pull it in a different way so that just populates a table so we just see the score and the subscores. But because we tend to watch the videos on a very small number of students, so we don't actually use this particular view of it [with the interview and writing sample visible in Slate]. We'll go to the dashboard if we wanted to pull up an individual student's video to watch. So typically our Slate, part of it is just this, the overall score and subscores, and then we'll use the dashboard for for digging into a video or writing sample"
- "We do listen to every record and so we have it [interview video] there, as a part of the application, but whether or not we look at it— again because for us it's not a required piece of information, or required part of the application—it depends on the strength of the file. And I don't know if it is always like in the order that's on there, on the left-hand side [of the Slate view], for us it's the last thing too. I don't know if it's always that way if that's set. But it, but it sort of makes sense, from our approach."
- "Yeah, we're much the same ..., everything is on the [CRM] dashboard and we kind of go look at it there. The only thing that stands out to me looking at this particular view is the date. What is that date? Is the date that it was taken? Is it the date that it was shared? Is the date that— I'm looking at it, so I just, for me now, I'm like looking at that now, I'm really intent to be looking at it. I'm like what is that thing for? But otherwise, I don't tend to look at it in this way"
- "Does it slow down Slate? Is that why you guys don't load it into each file? What I hear you saying, you use it infrequently, but was so easy for us to do, and I feel like there it is at our fingertips."
- "So it's unusual for us, we just, with the volume of applications we get, it's unlikely that we would go and look at the video. It's really only if we've got kind of particular questions and...when you do have an unusual applicant you want to kind of dig into them, but for the most part, then it's like a quick kind of check into dashboard, 'is a student okay? yes/no' and move forward. We don't go digging any deeper than that most of the time, and you're not toggling to the DET dashboard to check"

4.6.4.3 Results certificate

- "We have used that [=results certificate] a few times this year because it's a relatively new test, so our office is not super familiar with the test, and so I think... it's a nice visual reminder that, you know, these are point estimates, they're not exact scores, right? And so there can be that little error band that I

think helps remind people of that, and then as well as the descriptors I think also helps people familiarize themselves with what is, you know, a score of x y z really mean. So we did, but... we didn't do it for every student. It was really a small handful of students for whom we're really on that border or there is a lot more of conversation, people want to dig in deeper where there was sort of contentious argument about a particular student."

4.6.5 Question 5

- "I always describe or explain to students that with language proficiency... it's less to do about assessing, you know, readiness for a program or [anything] other than 'how was their language skill?'. So it's a very different component of an application in that it's, 'can they do this or not?', in a much more black and white [way]. I think hopefully you've heard from us that the students that get the attention in the language piece of an application are usually the students on the border. We're having to look at them for something else. And at that point it's when the DET results, particularly the interview, and as I think [other participant] had said, triangulating the pieces that we're seeing maybe in the subscores with other components of the application become helpful."
- "one of the things that we think about in the admission process is, is a student going to be successful at our institution and they're going to be able to, you know, to walk into a classroom and be comfortable and do well and interact with their classmates and their professors and, you know, I think we have a kind of a responsibility to students, to think about ...whether they're going to get through four years at our institution and do well. And so I think for non-native speakers. That is, you know, this is a part of it, they're going to be able to... because no one's going to slow down for them, they need to be able to, unless we're kind of talking about a pathway program or something. If you've got direct entry into a four year undergraduate program in the US and it's going to move fast, you're gonna have a lot of reading, you're going to have classroom interactions, you're going to have to be presenting, doing all kinds of things from the outset, and so, you know, We just have that responsibility to make sure that they're going to land on their feet when they come to us. And in some cases it is a case of, Can we offer them additional support to help them get to where they need to be? it's not always just to kind of a straightforward cut off. But I think it is kind of."
- "for me it feels like an ethical responsibility to some extent. Is the question here, just to kind of dive a little deeper specifically about the conclusions we draw based just on the DET? Because I think, you know, not to speak for this group, I think we all get the DET and then any English test in that same lens that we're describing here. It's not like, Oh, it was TOEFL, I would draw these conclusions, but Duolingo tasks I draw these conclusions. I get I don't speak

for everybody. That's, you know, a lot of head nodding, we all could draw the same conclusions based on an English language proficiency test, not what type of test necessarily is and what the date is."

- "I agree and I disagree at the same time because what we are getting from the DET, again, is this raw material that we don't get from other tests. Now, again, I'm admitting something about our process that is not particularly scientific, but, you know, here's me reading a spontaneous writing sample where I'm like, 'how sophisticated is the language that they're able to use off the top of their heads?' And also, I don't know, there's something to the way that they actually carry themselves, in this spoken part. And none of this is information that I can quantify. And all of this is information that I cannot get from other tests. So it is different to me in that way."
- "I looked as though it was agreeing with me and then adding additional information about why the DET gives us even more. I didn't see any disagreement there, all agreement and then some."
- "That's a good point, [other participant]. I agree with what you're saying, but also there is a distinction, and I suspect you agree with this as well, between the components of certain exams, and you'll see that in what their scores, what their sections of the exam actually are. And a distinct component [of] the DET is access to the interview, and the writing and so forth."
- "I would add to that, I think it's a valid point, but we've had a lot of push back at a senior level, about the potential bias for one way to interpret interviews and writing samples and the simple fact that we are not trained EL/TESOL professionals. We look at it with this kind of amateur eye and say, 'well this is fine to me'. ...It's why we definitely have a hesitation in our office about using that raw information. I think there's value to it at times, but as a kind of a standard metric for admission. Then you know we do it in, definitely more heavily on the actual scores and subscores and total scores."
- "Our faculty members are also relying on us (Admissions Officers) to ensure the students are able to keep up in the classes. To [other participant]'s points, we have an ethical responsibility to the students and process."
- "similar to that we, a few years ago, would watch all of the interviews that we got from Duolingo, from Initial View, from Vericant. We would have our student tour guides watch all of them, and they would give a rating that essentially said, you know, the student's going to be, like, strong in the classroom, you know, whatever, on average, not going to be able to perform, and that was when I first got to my institution. But since then, what we've done— We very much felt uncomfortable with that because we were asking, you know, 18-, 19-, 20-year-old students to gauge English language proficiency, when they're studying engineering, or business, you know, not language. And so we have since

decided to do away with that. We really don't watch any of them. We'll watch them in select cases, and we mostly accept them to let students send them to us because we were worried, so many kids would be, had sent them to us, that if we suddenly stopped accepting them, would there be like kind of a customer service issue with kids trying to send them to us and being upset that they can't. So, but essentially we don't really watch them anymore. And we spend more time looking at the concordance between the sub sections and our performance on campus, so pulling in GPA, pulling in... the grade for our first year English language courses, things like that."

- "But when you're talking about such a narrow band of competitive applicants, like, I'm like, 'okay well everybody has a score in the same range and so what are the differences?' The scores don't help to illustrate those, and I don't know that they should right? And so we return to this very human, very imperfect sort of element. That's where we found ourselves."
- "but [other participant] what I will say is not all institutions are looking at such a narrow band of students."
- "Institutions are admitting them at a much higher acceptance rate than many of you. And so, when we're going to those components of the DET,... we would all choose to look or not look at those with, again, a very different lens. We're going to those components hoping that we're going to find 'Yes', if that's the case. But it is imperfect; it is absolutely imperfect and completely subject to human flaw."

4.6.6 Question 6

- "I will say from my perspective that not really anything. ...we've already set our benchmark of what we're looking for. So I'm generally not going back to any of those additional resources. I feel like they're useful but not necessarily as I'm using the actual DET ...the webinar that I went to this week was useful, but in a kind of a general information sense, not as I'm actually looking at an applicant review in an application. So there's very little that I would necessarily be using from the resources, as I go through the individual evaluation process for an applicant."
- "I would agree. I would say that all of that information is really useful when you're trying to set a score minimum.... but as soon as it reaches the evaluation and selection stage that's not something that, I mean, I'd never watch a video recording or webinar about it at that point."
- "The only other time that we use it is as training purposes for new stuff."
- "I would say it's extremely useful for maybe whoever the international admissions lead or training lead is, and often what we do is we will distill a lot of this stuff into, probably, a paragraph or a few sentences of what an admissions officer

would need to know when reviewing the file because they've got, you know, 50 other things that they're also paying attention to. But it is helpful for us to establish those initial summaries, so I would say probably the most frequently used, of course, the concordances and subscores or explanations. I think those are probably the two things we use the most. But otherwise, you know, those are probably going to ... link to a website if they have any other questions, or they'll come to whoever the sort of main lead is if they needed to dive deeper into other technical resources. But that typically doesn't happen during the review phase; it's usually, you know, maybe over the summer, sort of pre-review preparation training, that kind of thing."

- "Keeping in mind accreditation audits and all the administrative changes that we've mentioned and how subject all of this is to that, it's kinda, even though we're not using it every day, and it was very helpful initially, you have to have it there. And you have to stay on top of it, it can't be stale data. ...if you think of the test as the movie that the kids want to see, well, this is the theater that needs to hold them. From a training perspective as well, it's giving offices the opportunity not to create/recreate, and to have to keep up with, our own training manuals. It is available from the source's mouth itself. So that's always helpful and, you know, it's a lot of turnover from recruitment staff to admission staff, and just having the information and knowing the direct source to get it from, from a training perspective, saves a lot of time from our end."
- "I want to use this opportunity to give you guys all a compliment in saying that everything is presented in such a clear way that I feel like there's relatively little necessity, we don't have to do a whole lot of training because it's so easy to understand what you're putting in front of us. So thank you for that." Two participants expressed enthusiastic agreement in the chat
- "a lot of the work is being done from our admissions team, so a lot of that decision making happens with them in our senior and then it goes up to Senate. And from a recruitment perspective, for me to be able to dive in a little bit deeper and have some more background information rather than just saying to a student like here's the minimum admission requirements that we're looking for for your English proficiency based on the tests that you're presenting. I'm now able to go and do my own research and my own understanding without relying on my admission staff to give that information to me so it allows me to be more autonomous and just more proactive with my learning, and this is something that I encourage my staff to do because we're not directly involved with, you know, all of the hard paper and scores that we get. And of course our institutions are so, so different, but to [other participant]'s point, clear, concise, easy to digest information, and just the webinars, has been phenomenal for learning."

Chapter 5

Final Discussion

5.1 Overview of Purpose and Main Findings

The goals of this dissertation were to evaluate the performance of non-IRT–based methods of estimating classification consistency when applied to a computer-adaptive test (CAT), and to investigate the needs and behaviors of admissions officers as an audience of English language proficiency (ELP) test results. The first paper used a simulation study to address the former goal. Test-taker responses to a CAT were simulated under different conditions of test length, parameter estimate error, item selection criterion, and scoring model. Classification consistency was then estimated using three classical test theory (CTT)–based methods (test–retest, split-half, and Livingston–Lewis), two of which were applied both with and without adjustment. The results suggest that CTT–based methods are potentially appropriate in a CAT context, with each method showing sensitivity to different simulation factors. The adjusted split-half and Livingston–Lewis methods performed particularly well in terms of RMSE. However, given the different requirements of each method (e.g., the Livingston–Lewis method requires a test reliability estimate as input), the optimal method will depend on the characteristics and available information of a specific context. Nevertheless, the results suggest that CTT–based methods should not be dismissed out of hand when estimating classification consistency of a CAT, and such methods may prove more feasible in complex adaptive assessment contexts such as game-based assessment or assessments that incorporate automated scoring of speaking and/or writing responses.

The second paper was a direct extension of the first, applying the methods investigated in the simulation study to an operational data set from the Duolingo English Test. Additionally, a bootstrapping-inspired approach was explored based on resampling the component scores that are used to compute the overall test score. The results

were broadly consistent with the simulation study in that the unadjusted test–retest approach produced markedly low estimates at cut scores near and below the population mean, but higher estimates than other methods at score points further above the population mean. Also, the Livingston-Lewis and adjusted split-half methods produced roughly comparable estimates at cut points near the population mean. However, the Livingston–Lewis method exhibited sensitivity to the reliability estimate used as input. The bootstrapping-inspired approach followed the general pattern of the other methods but produced markedly higher estimates over much of the score scale. Additional investigation, such as a simulation study, are needed to refine this approach to estimating classification consistency. Nevertheless, the results of this paper provide further support to the use of non-IRT–based methods for estimating classification consistency of a CAT, although each method poses unique challenges (e.g., the need for split-half scores).

The third paper investigated the informational needs and priorities of postsecondary admissions officers as an audience of results reports from ELP tests. Although not explicitly connected to the first two papers, this study speaks to the need to understand how test users interpret and use psychometric information about tests, such as classification consistency indices discussed in the first two papers. The study employed a focus group approach with eight participants who were all admissions officers working at North American universities. Participants discussed six questions relating to the adoption, interpretation, and use of ELP tests for admissions purposes. Responses revealed several aspects of participants' priorities and behaviors as consumers of results reports. Participants emphasized a sense of ethical and professional responsibility, manifesting in attention to issues of fairness and a desire for students to be successful. Participants demonstrated some aspects of language assessment literacy (LAL), such as conceptualizing ELP as a latent construct and acknowledging measurement error. However, it was also suggested that participants' use of ELP test results is determined by factors other than LAL, such as institutional priorities (e.g., incoming class profile) and the perceived need to present ELP test options as interchangeable. Related to the latter point, the importance of concordance between ELP tests was raised repeatedly. Concerning the use of test results reports and supplemental information about a test, participants reported that applicants' results are typically imported directly into software for reviewing applications, and thus the results reports received by test takers are rarely seen by admissions officers; supplemental test information (e.g., test technical manual, research reports) are accessed infrequently. The study results suggest that it would behoove test developers and language assessment researchers to pay attention to the various factors influencing the use of test results for admissions purposes in order to better design results reports and supplemental materials that promote appropriate test use.

5.2 Limitations

As with all research, the studies presented in this dissertation were limited in some ways, both foreseen and unforeseen. The primary limitations of the first study were the capabilities of the simulation package and computational speed. The package used for the simulation, catR (Magis & Barrada, 2017), was limited in the item selection criteria it could successfully implement[1]. The investigation of how item selection criteria influence classification consistency was thus restricted to random selection and two criteria based on Fisher information. Furthermore, the number of test takers, items, and manipulated factors included in the simulation were limited by available computation power. The scope of the simulation was reduced to acheive a more manageable simulation run time, but even as presented herein and utilizing parallel processing, the simulation took approximately 10 days to run.

The primary limitation of the second study was the non-experimental nature of the data. As the data were not collected in the context of a controlled experiment and test takers self-selected to repeat the test, it is effectively impossible to ensure the representativeness of the repeat test-taking population. Additionally, it is also not possible to control certain aspects of the test-retaking conditions, such as time between test attempts or consistency of testing conditions (e.g., device used, time of day) across attempts. For these reasons, it is impossible to determine a "true" classification consistency against which to compare estimates from various methods.

The primary limitation of the third study was access to the target population, North American admissions officers. It is challenging to even identify, much less recruit, members of this population, making it difficult to obtain a sample that is representative of the population. Furthermore, the logistics of scheduling and limits to the time that admissions officers can be reasonably asked to devote to an uncompensated study limited the possible duration of the focus group. As such, there was not sufficient time to fully explore every participant's thoughts on each discussion topic, much less unplanned interesting topics that came up during discussion.

5.3 Methodological Research Directions

Building on the previous section, there are a few ways that future research could expand upon the studies presented here. One such category is simulation approaches. The first study could potentially be replicated with a different simulation software and/or using more computing power, thereby allowing for a more complex simulation (e.g., by simulating polytomous/continuous item scores or component scores that are

[1]While the package documentation lists 13 available item selection criteria, many of these options caused simulation runs to freeze.

combined to derive the total score) or, at the very least, consideration of other item selection criteria. A simulation could also be used to further explore and validate the simulation-inspired approached applied in the second study.

Another category of research directions concerns sampling and data collection. Classification consistency methods could be better explored if data are controlled in the context of a controlled experiment, such that a true classification consistency index can be calculated. Research on the use of test results by admissions officers would also benefit from a larger, more representative sample. More and/or longer focus group sessions would also allow for deeper exploration of topics and corroboration of themes across focus groups. Alternative data collection methodologies (e.g., survey, interview) could also facilitate the collection of data on topics of interests.

5.4 Substantive Research Directions

In addition to methodological considerations, there are also substantive avenues for expanding upon the studies. Concerning classification consistency, there is ample work to be done to investigate the estimation of classification consistency in the context of non–fixed-form assessments, including computer-adaptive assessments (CATs) and game-based assessments (GBAs). While the first two studies of this dissertation explored the application of a number of CTT–based methods in a CAT context, varying several aspects of the CAT, both simulation and real-data studies could be done to investigate the relationship of a wider variety of CAT properties to classification consistency. Moreover, there is research to be done to investigate the application of classification consistency methods to other non-traditional assessment contexts such as GABs, which could present unique issues and challenges.

With regard to test results use by admissions officers, the third study can be considered as a foundation that revealed general aspects of the use of ELP tests in the postsecondary admissions process, as well as the potential priorities and behaviors of admissions officers in interpreting ELP test results. Future research should build upon this base to more specifically investigate how admissions officers interpret psychometric information such as classification consistency, and how/whether such information factors into decisions to accept an ELP test and decisions about individual applicants. Such "decision processes" of test results users appear to be sorely under-researched in the educational measurement/psychometric literature. Research on this topic could serve to develop more effective supplemental materials to facilitate valid use of test results by stakeholders such as admissions officers.

References

Alger, S. (2016). Is this reliable enough? Examining classification consistency and accuracy in a criterion-referenced test. *International Journal of Assessment Tools in Education*, *3*(2), 137–150.

Allen, M. J., & Yen, W. M. (1979). *Introduction to measurement theory*. Prospect Heights, IL: Waveland Press, Inc.

American Association of Collegiate Registrars and Admissions Officers. (2014). *2014-2015 state of CRM use in higher education report*.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (Eds.). (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Bachman, L., & Palmer, A. (2010). *Language assessment in practice*. Oxford University Press.

Banerjee, J. V. (2003). *Interpreting and using proficiency test scores* (PhD thesis). University of Lancaster.

Barrada, J. R., Olea, J., Ponsoda, V., & Abad, F. J. (2008). Incorporating randomness in the fisher information for improving item-exposure control in CATs. *British Journal of Mathematical and Statistical Psychology*, *61*(2), 493–513.

Bourque, M. L., Goodman, D., Hambleton, R. K., & Han, N. (2004). *Reliability estimates for the ABTE tests in elementary education, professional teaching knowledge, secondary mathematics and english/language arts (final report)*. Leesburg, VA: Mid-Atlantic Psychometric Services.

Brennan, R. L. (2004). *Manual for BB-CLASS: A computer program that uses the beta-binomial model for classification consistency and accuracy. Version 1*. CASMA Research Report.

Breyer, F. J., & Lewis, C. (1994). Pass-fail reliability for tests with cut scores: A simplified method. *ETS Research Report Series*, *1994*(2), i–30.

Brown, J. D., & Hudson, T. (2002). *Criterion-referenced language testing*. Cambridge University Press.

Cardwell, R. L., LaFlair, G. T., & Settles, B. (2021). *Duolingo English Test: Technical manual*. Duolingo.

Carlsen, C. H. (2018). The adequacy of the B2 level as university entrance requirement. *Language Assessment Quarterly*, *15*(1), 75–89.

Chalhoub-Deville, M. (2003). Fundamentals of ESL admissions tests: MELAB, IELTS, and TOEFL. In D. Douglas (Ed.), *English language testing in US colleges and universities* (pp. 11–35). NAFSA Washington, DC.

Chalhoub-Deville, M., & O'Sullivan, B. (2021). *Validity: Theoretical developments and integrated arguments*. Equinox.

Chiswick, B. R., & Miller, P. W. (2005). Linguistic distance: A quantitative measure of the distance between English and other languages. *Journal of Multilingual and Multicultural Development*, *26*(1), 1–11.

Clauser, A., & Rick, F. (2016). *Designing and evaluating score reports for a medical licensing examination*. Washington, D.C.: Paper presented at the annual meeting of the National Council on Measurement in Education.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*(1), 37–46.

Council of Europe. (2018). *Common European Framework of Reference for Languages: Learning, teaching, assessment: Companion volume with new descriptors*.

De Ayala, R. J. (2013). *The theory and practice of item response theory*. Guilford Publications.

de Juan-Jordan, H., Guijarro-Garcia, M., & Gadea, J. H. (2018). Feature analysis of the "customer relationship management" systems for higher education institutions. *Multidisciplinary Journal for Education, Social and Technological Sciences*, *5*(1), 30–43.

Deng, N., & Hambleton, R. K. (2013). Evaluating CTT- and IRT-based single-administration estimates of classification consistency and accuracy. In R. E. Millsap, L. A. van der Ark, D. M. Bolt, & C. M. Woods (Eds.), *New Developments in Quantitative Psychology* (Vol. 66, pp. 235–250). Retrieved from http://link.springer.com/10.1007/978-1-4614-9348-8_15

Deygers, B., Zeidler, B., Vilcu, D., & Carlsen, C. H. (2018). One framework to unite them all? Use of the CEFR in European university entrance policies. *Language*

Assessment Quarterly, *15*(1), 3–15.

Diao, H., & Sireci, S. G. (2018). Item response theory–based methods for estimating classification accuracy and consistency. *Item Response Theory*, *19*, 6.

Duolingo English Test. (2020). *Excited to announce that the Duolingo English Test is now accepted by over 3,000 programs around the world! Thank you to the students and to the higher education community for all your support this year. We look forward to.* Facebook status update.

Efron, B. (1987). Better bootstrap confidence intervals. *Journal of the American Statistical Association*, *82*(397), 171–185.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists.* Mahwah.

Foreign Service Institute. (2020). *Foreign language training – United States Department of State.* Retrieved from https://www.state.gov/foreign-language-training/

Graybill, F. A., & Deal, R. (1959). Combining unbiased estimators. *Biometrics*, *15*(4), 543–550.

Haakstad, H. (2020). *betafunctions: Functions for working with two- and four-parameter beta probability distributions.* Retrieved from https://CRAN.R-project.org/package=betafunctions

Haertel, E. H. (2006). Reliability. In R. L. Brennan (Ed.), *Educational Measurement* (Fourth, pp. 65–110). American Council on Education.

Hambleton, R. K., & Zenisky, A. L. (2013). Reporting test scores in more meaningful ways: A research-based approach to score report design. In *APA handbook of testing and assessment in psychology, vol. 3: Testing and assessment in school psychology and education* (pp. 479–494). Washington: American Psychological Association.

Han, K. C. T. (2018). Components of the item selection algorithm in computerized adaptive testing. *Journal of Educational Evaluation for Health Professions*, *15*.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational Measurement* (Fourth, pp. 1–73). American Council on Education.

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

Kremmel, B., & Harding, L. (2020). Towards a comprehensive, empirical model of language assessment literacy across stakeholder groups: Developing the language assessment literacy survey. *Language Assessment Quarterly*, *17*(1), 100–120.

LaFlair, G. T. (2020). *Duolingo English Test: subscores*. Duolingo.

Lathrop, Q. N. (2015). *cacIRT: Classification accuracy and consistency under item response theory*. Retrieved from https://CRAN.R-project.org/package=cacIRT

Lay, A., Patton, E., & Chalhoub-Deville, M. (2017). A case for the use of the ability-in language user-in context orientation in game-based assessment. *Language Testing in Asia*, *7*(1), 1–17.

Lee, W.-C. (2010). Classification consistency and accuracy for complex assessments using item response theory. *Journal of Educational Measurement*, *47*(1), 1–17.

Lipner, R. S., & Brossman, B. G. (2017). *Changing the design of examinee score reports*. Scottsdale, AZ: Paper presented at the ATP Innovations in Testing conference.

Livingston, S. A. (1972). Criterion-referenced applications of classical test theory 1, 2. *Journal of Educational Measurement*, *9*(1), 13–26.

Livingston, S. A., & Lewis, C. (1995). *Estimating the consistency and accuracy of classifications based on test scores*. 20.

Luecht, R. M. (2015). Applications of item response theory: Item and test information functions for designing and building mastery tests. In S. Lane, M. R. Raymond, & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 485–506). Routledge.

Magis, D., & Barrada, J. R. (2017). Computerized adaptive testing with R: Recent updates of the package catR. *Journal of Statistical Software, Code Snippets*, *76*(1), 1–19.

Marshall, J. L., & Haertel, E. H. (1975). *A single-administration reliability index for criterion-referenced tests: The mean split-half coefficient of agreement*. Washington, D.C.: Paper presented at the Annual meeting of the American Educational Research Association.

Menard, J. (2020). CRM market share (HigherEd). Retrieved from https://www.listedtech.com/blog/crm-market-share

Nydick, S. W. (2014). The sequential probability ratio test and binary item response models. *Journal of Educational and Behavioral Statistics*, *39*(3), 203–230.

O'Donnell, F., & Zenisky, A. L. (2020). Results reporting for large-scale assessments [digital ITEMS module 21]. *Educational Measurement: Issues and Practice*, *39*.

R Core Team. (2020). *R: A language and environment for statistical computing*. Retrieved from https://www.R-project.org/

Rudner, L. M. (2000). Computing the expected proportions of misclassified examinees. *Practical Assessment, Research, and Evaluation*, *7*(1), 14.

Rudner, L. M. (2005). Expected classification accuracy. *Practical Assessment, Research, and Evaluation*, *10*(1), 13.

Rudner, L. M. (2009). Scoring and classifying examinees using measurement decision theory. *Practical Assessment, Research, and Evaluation*, *14*(1), 8.

Sawaki, Y. (2016). Norm-referenced vs. Criterion-referenced approach to assessment. In D. Tsagari & J. Banerjee (Eds.), *Handbook of Second Language Assessment*. Berlin, Boston: De Gruyter.

Settles, B., LaFlair, G. T., & Hagiwara, M. (2020). Machine learning–driven language assessment. *Transactions of the Association for Computational Linguistics*, *8*, 247–263.

Stricker, L. J. (2004). The performance of native speakers of English and ESL speakers on the computer-based TOEFL and GRE general test. *Language Testing*, *21*(2), 146–173.

Subkoviak, M. J. (1976). Estimating reliability from a single administration of a criterion-referenced test. *Journal of Educational Measurement*, 265–276.

Subkoviak, M. J. (1988). A practitioner's guide to computation and interpretation of reliability indices for mastery tests. *Journal of Educational Measurement*, *25*(1), 47–55. Retrieved from http://www.jstor.org/stable/1435023

Toulmin, S. (1958). *The uses of argument*. Cambridge, UK: Cambridge University Press.

Tufte, E. R., Goeler, N. H., & Benson, R. (1990). *Envisioning information* (Vol. 126). Graphics press Cheshire, CT.

Tufte, E. R., & Robins, D. (1997). *Visual explanations*. Graphics Cheshire, CT.

von Davier, A. A., Deonovic, B., Yudelson, M., Polyak, S. T., & Woo, A. (2019). Computational psychometrics approach to holistic learning and assessment systems. *Frontiers in Education*, *4*, 1–12.

Wan, L., Brennan, R. L., & Lee, W.-C. (2007). *Estimating classification consistency for complex assessments* (p. 61). Iowa City: Center for Advanced Studies in Measurement; Assessment, University of Iowa.

Welsh, M. (2018). *Bias in science and communication*. IOP Publishing.

Wheadon, C. (2014). Classification accuracy and consistency under item response theory models using the package **classify**. *Journal of Statistical Software*, *56*(10). Retrieved from http://www.jstatsoft.org/v56/i10/

Woodruff, D. J., & Sawyer, R. L. (1989). Estimating measures of pass-fail reliability from parallel half-tests. *Applied Psychological Measurement*, *13*(1), 33–43.

Wyse, A. E., & Babcock, B. (2016). Does maximizing information at the cut score always maximize classification accuracy and consistency? *Journal of Educational Measurement*, *53*(1), 23–44.

Zapata-Rivera, D., & Katz, I. R. (2014). Keeping your audience in mind: Applying audience analysis to the design of interactive score reports. *Assessment in Education: Principles, Policy & Practice*, *21*(4), 442–463.

Zapata-Rivera, D., & VanWinkle, W. (2010). *A research-based approach to designing and evaluating score reports for teachers*. Princeton, NJ: ETS.

Zenisky, A. L., & Hambleton, R. K. (2016). A model and good practices for score reporting. In L. S., M. R. Raymond, & T. M. Haladyna (Eds.), *Handbook of test development* (Second, pp. 585–602). Routledge.