Exploring the Value of Double Marking in Dissertation Assessments[*]

Classical Test Theory and Item Response Theory Approaches

James Steele[†] Matthew Shaw[‡]

29/09/2022

Abstract

Extended constructed response assessment methods such as the dissertation are common in higher education assessment and are typically afforded considerable weight in overall student assessment. Thus, quality assurance processes such as double marking are commonly implemented. However, the value of such processes has been questioned. Further, the measurement properties of the dissertation assessment method have received little attention. As such, we explored the dissertation assessment method through both Classical Test Theory (CTT) and Item Response Theory (IRT) approaches using a historical dataset of first and second marker grades. Under CTT we found poor agreement between markers which could threaten the validity of the true grades assigned to students. However, under IRT models we found that markers showed greater agreement regarding the underlying latent abilities thought to give rise to the extended constructed response that is the dissertation. We conclude by questioning the value of double marking processes. Grades *qua* grades (i.e., true grades) typically show poor agreement between markers suggesting double marking may be a waste of resource. Instead, the determination of grades from latent ability score using an IRT measurement model, which showed greater agreement between markers, might enable a single marker to provide a valid grade to students.

1 Introduction

In his chapter in the **Handbook on Measurement, Assessment, and Evaluation in Higher Education**, Mislevy (2017) differentiates between *assessment*, *examination*, *test*, and *measurement* as follows:

> "Assessment, conceived broadly, is gathering information about what students know and can do for some educative purpose. Examinations and tests, as the terms will be used here, are particular ways of doing this. Measurement is different. Measurement is situating data from an assessment in a quantitative framework, to characterize the evidence the observations provide for the interpretations and inferences the assessment is meant to support."
>
> — Mislevy (2017), p. 37

A students responses to different assessment methods such as examinations and tests are thought to be representative of their ability or proficiency (Brookhart et al., 2016; Edgeworth, 1888). Grades are symbols assigned to individual assessments methods (e.g. examination, test, essay) and composite assessments of a students performance, such as degree classification (Brookhart et al., 2016). That we, in our approach to assessment of students, assign grades or scores to the examinations or tests employed seems to imply that

[*] Preprint Please cite as: Steele, J., & Shaw, M. (2022). Exploring the Value of Double Marking in Dissertation Assessments: Classical Test Theory and Item Response Theory Approaches. edarxiv. DOI: 10.35542/osf.io/ug7yb

[†] Faculty of Sport, Health, and Social Sciences, Solent University, United Kingdom, james.steele@solent.ac.uk

[‡] Department of Sport, Food and Natural Sciences, Faculty of Education, Arts and Sports, Western Norway University of Applied Sciences, Bergen, Norway, Matthew.Shaw@hvl.no

we are indeed interested in trying to measure something i.e., their ability or proficiency. Yet, while we often spend time thinking about and discussing these grades, in addition to applying approaches we hope assure the quality of their provision, rarely do we take stock to consider what exactly it is we are trying to measure and how valuable these methods and approaches are. In fact, despite many standards regarding assessment methodology and measurement theory being developed from within higher education institutions, in many regards these institutions have failed to actually follow such standards (Scriven, 2017).

Extended constructed response assessment methods (i.e., essay type methods whether in exams or as coursework; Almond (2014)) are common in higher education assessment. This seems primarily due to the belief that they do a better job of assessing higher level thinking skills/abilities; the student cannot merely select a response from those available (as in a multiple choice test), they must recall or construct it using these abilities. Within most higher education undergraduate degrees, the culminating assessment is an extended constructed response method: the dissertation. It has been noted as having:

> "...a privileged place within many degree programmes. Viewed as the culmination of the degree, the dissertation is seen as the mechanism through which students construct a synthesis of theory, published studies, methodological understanding, the selection, and application of appropriate research methods, analysis, and decision."
>
> — Hemmings (2001), p. 241

This occurs under supervision and some guided instruction, with the supervisor also involved in the assessment of the dissertation (Nyamapfene, 2012). The dissertation is also typically marked across various sections (though dependent on the topic/discipline) and then from this an overall composite grade is awarded. Given the importance of the dissertation to assessment, evidenced by the weighting it is usually afforded toward the overall degree classification, a quality assurance process is often incorporated to these types of assessment method. Typically, a dissertation goes through some process of grade moderation (a sample of grades are checked by another assessor) or some form of double marking and grade agreement. Double marking is where a grade and feedback is provided by a first marker and then by a second marker who has access to the first marker's grade when completing their assessment. However, a blinded double marking process is sometimes used where the first marker's grade is not known by the second marker (Bloxham, 2009). Both markers are in essence supposed to be attempting to perform the same assessment process and thus be measuring the same abilities of the students being assessed.

It seems reasonable to propose that we should be concerned with the measurement properties of this method of assessment, and the process of double marking, given that there is historical contention regarding them (Bloxham, 2009; Hornby, 2003). As noted, by assigning grades in our assessment methods we appear to be claiming to be measuring something and ideally we would like that method to display both validity and reliability. These two measurement concepts are inherently intertwined and, given the typical process of double marking used, the dissertation assessment method presents an interesting opportunity to explore them. Should there be present disagreement between first and second markers, this issue of interrater reliability also poses a threat to the validity of the grades assigned leading to the question of *"What exactly are we measuring?"*.

However, if there is good agreement between first and second markers then, independent of a moderation process (i.e. double or double blind marking), there is potentially an argument in terms of resource constraints to remove a second marker from the assessment process. Moderation consumes time (Bloxham, 2009; Winstone & Boud, 2022) in multiple ways, such as delaying feedback to students, and competing with research time to the extent where some academics neglect grading procedures in order to conduct research activity (McIntosh et al., 2022; Pan et al., 2014). The false assumptions of laborious moderation processes highlighted by Bloxham (2009) likely continue to prevail today, with many academics contending that this represents *'wasted resources'*. Rigorous bureaucracy and implementation of quality assurance processes with untested assumptions may simply be a mechanism of protection from student complaints (Winstone & Boud, 2022) as opposed to actually providing some demonstrable benefits regarding assessment from a measurement perspective. As such, exploration of measurement properties of dissertations under double marking processes should be a priority. Results of such investigations have considerable practical implications. For example, we argue that high levels of agreement in a double blinded marking process across students and markers might yield a second marker unnecessary thus freeing academic staff to spend more time on other teaching-related or research-related activities. Alternatively, poor agreement in a present double blinded marking process might suggest the need for additional approaches to enhance the measurement properties of this assessment method.

There are different paradigms typically applied regarding measurement in educational settings that can be applied to evaluating assessment and quality assurance approaches, two of which we will consider in this report: Classical Test Theory (CTT) and Item Response Theory (IRT). While these are more closely aligned than many realise (Raykov et al., 2019; Raykov & Marcoulides, 2016) and indeed we will move from one to the other and back again in our empirical exploration of dissertation assessment data, we will first introduce the theoretical foundations of these two paradigms for this context. Then, we shall apply approaches from these paradigms in the context of exploring the value of current double marking practices for final year dissertation projects.

1.1 Classical Test Theory

The CTT paradigm regarding measurement focuses on assessment grades *qua* grades and relies on a form of operationalism at least with respect to the concepts purported to be measured (Bridgman, 1927); though, strictly speaking CTT (or IRT) is not an approach to 'measurement' in the fundamental sense of classical scientific measurement (i.e., *"the estimation or discovery of the ratio of some magnitude of a quantitative attribute to a unit of the same attribute"*; -Michell Michell (1997)) instead falling more within representational theories of measurement. CTT, or 'true score theory' as it is also known, starts from the simple assumption that each individual has a *true score* for any given measurement operation which would be observed if there were no errors in measurement. CTT can thus be illustrated with the following simple formalisation:

$$y_p \sim t_p + e_p \tag{1}$$

Where for the pth person y_p is the observed measure, t_p is the true score or grade, and e_p is the error of the measurement, assumed to be independent of t_p and described by a normal distribution with mean i.e., $Normal(0, \sigma)$. Under this, the true score can be considered the underlying latent variable we are interested in but can only observe imperfectly.

Under the assumptions of CTT regarding sampling error, the validity of a given measure is inherently linked to its reliability through the square root law; that is to say that the validity coefficient of a given test cannot exceed the square root of the reliability coefficient. Given that ideally we want our assessment methods to be *valid* measures, that is to say the the grades assigned actually provide us with *"...information about what students know and can do..."* (i.e., the students ability; Mislevy (2017)) that has acceptable verisimilitude, we should therefore take an interest in their *reliability* also. If reliability is poor then this brings into question the extent to which any grade awarded is really a valid reflection of the true grade that would be obtained. Agreement between observers/raters (i.e., interrater agreement) is inherently related to reliability though involves the assumption that measurement error occurs due to inconclusive observations between raters.

The primary statistics used to estimate interrater agreement for categorical data under CTT are Cohen's κ like indexes. These are coefficients which typically range from 0 to a theoretical maximum of 1. When $\kappa = 0$ this means that agreement between raters is no different than what would be expected given random guessing of grades[1]. Conversely, when $\kappa = 1$ this means that there is perfect agreement between raters even accounting for chance agreement. Of course, random guessing probabilities (i.e., the chance that two independent raters would obtain the same grademark if they merely guessed randomly) need to be accounted for, which it is in Cohen's κ like indexes.

While an individual's observed grade on an assessment method might be considered as indicative of the underlying ability that the assessment is intended to measure, within CTT it is specific to that method and

[1] Though notably can obtain values $\kappa < 0$ which would imply that agreement was actually less than random chance.

the specific demands it presents. That is to say there is a dependency between the characteristics of the assessment method (e.g., its difficulty) that can be estimated under CTT for a given assessment method and the ability of the group taking it. Further, both the assessment characteristics and student ability cannot be placed on the same scale within CTT[2].

Although CTT has been incredibly valuable in ascertaining parameters of different methods such as reliability and validity under a form of operationalism, some argue that, although discussion of student assessment often revolves around grades, it is not grades *qua* grades that we are interested in particularly given the foibles of the assessors involved in producing them. The thing that we are really interested in is measuring the underlying latent *ability* that gives rise to response to the assessment method that we actually observe that grades are awarded for by assessors; the students *understanding* or *knowledge* demonstrated through the response to the assessment.

1.2 Item Response Theory

The IRT paradigm to measurement involves a probabilistic framework for a family of measurement models. IRT, also known as 'latent trait theory', consists of a mathematical model that relates an individual's unobserved (i.e., latent, or not directly measured) ability and the observed performance on a particular assessment method (i.e., test or examination). These models express the probability of a particular response to an assessment method item as a function of both an individual's ability, and one or more item parameters including its difficulty.

For example, if for each person we assess $p(p = 1, ..., P)$, and each item in an assessment method they complete $i(i = 1, ..., I)$, we have a binary response y_{pi} which is coded [=1] for a correct answer (i.e., success), and [=0] for an incorrect answer (i.e., failure) such as in a multiple choice exam, a binary IRT model aims to model $p_{pi} = P(y_{pi} = 1)$; in essence the probability that a person p correctly responds to an item i which is assumed to follow a Bernoulli distribution ($y_{pi} \sim Bernoulli(p_{pi})$).

Different models imply different assumptions one is willing to make about the data being examined given the nature of the assessment method conducted. For example, a recently popular model due to its flexibility is the four-parameter logistic model (4PL) where $P(y_{pi} = 1)$ is expressed through the equation:

$$P(y_{pi} = 1) = \gamma + (1 - \gamma_i - \psi_i)\frac{1}{1 + exp(-(\alpha_i\theta_p - \beta_i))} \tag{2}$$

In this model there are four key parameters as the name suggests, which reflect the assumptions about the data. The β_i parameter describes the item location which, depending on the sign direction people prefer, can refer to either the 'difficulty' or the 'easiness' of the item. The α_i parameter refers to how well an item discriminates abilities or how strongly an item is related to the latent ability θ_p which is typically positive (i.e., that responding correctly typically implies higher ability than if responding incorrectly).The parameters γ_i and ψ_i refer to the probability of random correct response or guessing probability (i.e., that the correct response on an item could be guessed and not due to ability), and a lapse probability respectively (i.e., that a person could make a mistake etc. despite having the ability to make a correct response).

An item characteristic curve is usually used to visualise the relationship between ability and the probability of a correct response to items. So for example, a 4PL model might look something like figure (1). Changes to the model imply different assumptions about these parameters. For example, the simplest one-parameter logistic model (1PL or Rasch model) assumes that $\alpha = 1$ and both $\gamma = 0$ and $\psi = 0$ in equation (2); that is to say, items discriminate between higher and lower abilities equally well and there is no guessing or lapses occurring. A model of that kind might look like figure (2).

[2] Difficulty under CTT is defined as the proportion of individuals obtaining a particular score, whereas ability is the total score. As a students ability is defined only in terms of a particular assessment method, when it is "hard" they will appear to have lower ability, and when it is "easy" they will appear to have higher ability despite their ability being the same in both instances.

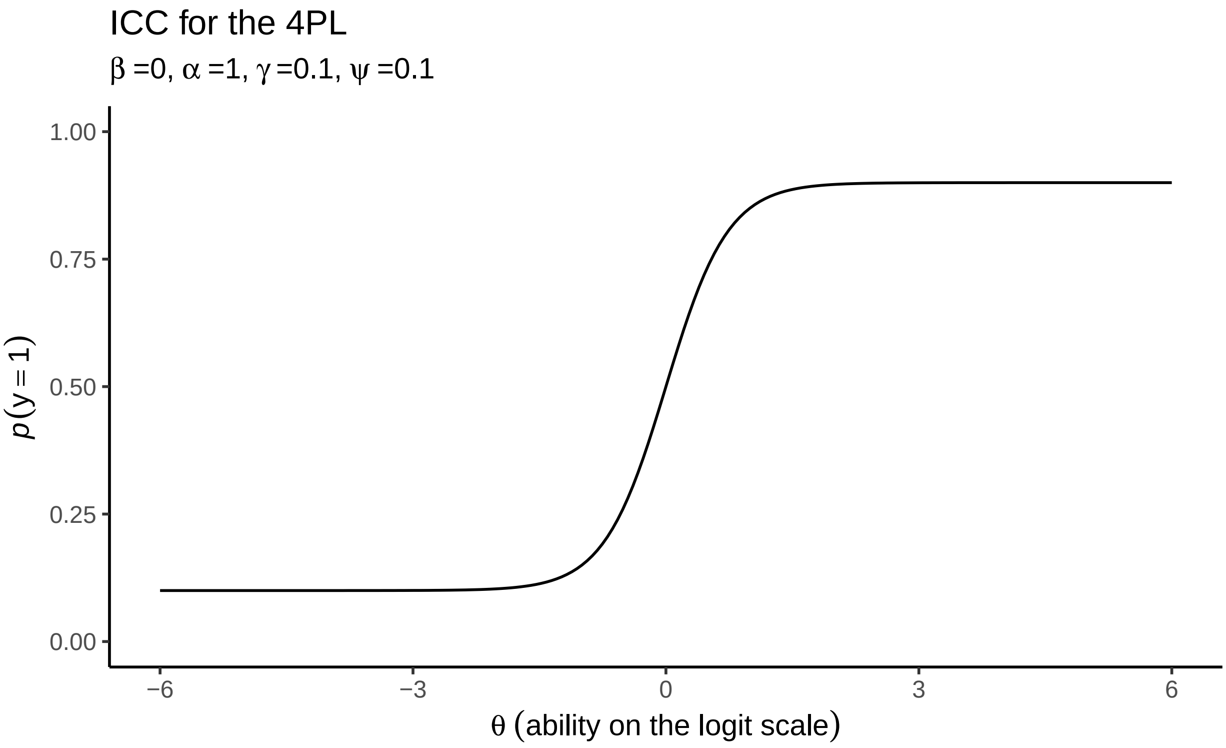


Figure 1: Example item characteristic curve for the 4PL model.

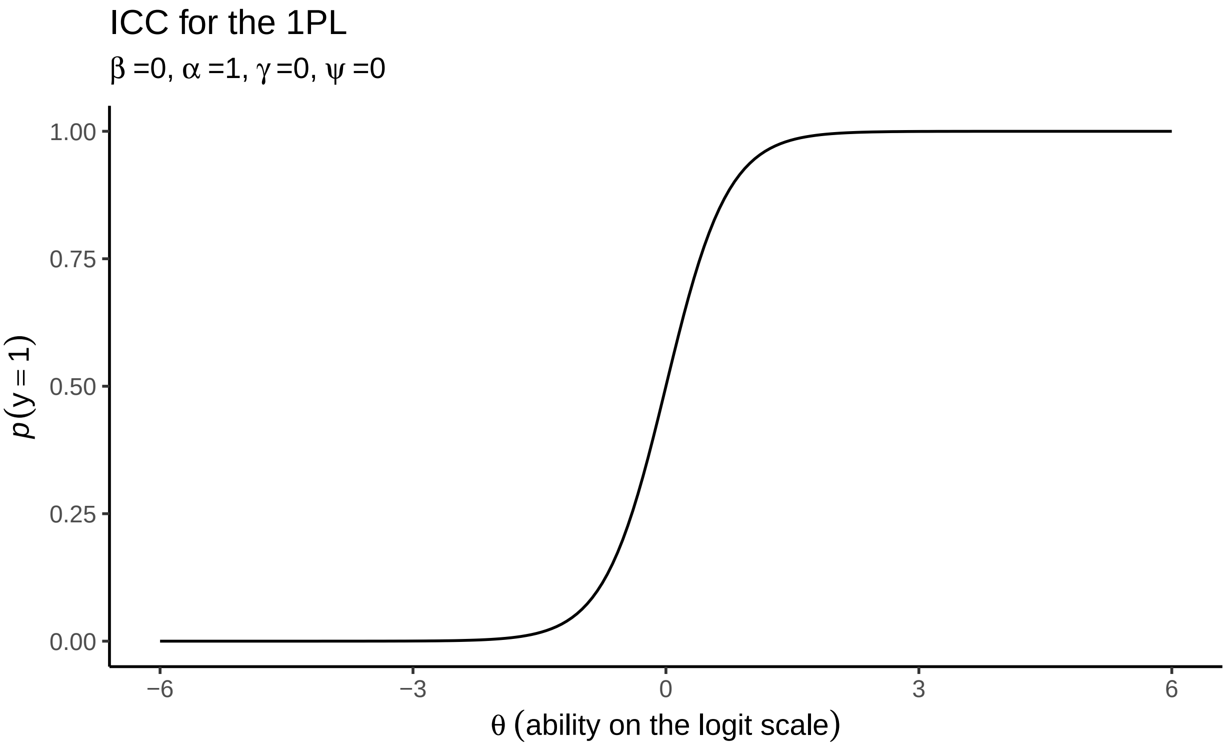


Figure 2: Example item characteristic curve for the 1PL model.

For constructed response assessments often a polytomous grading system is used. Most modern assessments also have a marking rubric which splits into sections that are each awarded a grademark. So in this case we have for each person we assess $p(p = 1, ..., P)$, and each item (i.e., section) in an assessment method they complete $i(i = 1, ..., I)$, a categorical response y_{pi} awarded which is coded with a grade ranging C categories. Where y is a categorical response with $C > 1$ unordered categories, the categorical distribution is employed ($y \sim categorical(\psi_1, ..., \psi_C)$). Most grading systems however have ordered categories (e.g., A, B, C etc.) and the possible models build upon the categorical distribution in how they define the category probabilities $P = (y = c)$. A common ordinal family applied is the cumulative model which assumes:

$$P(y = c) = F(\tau_c - \psi) - F(\tau_{c-1} - \psi) \tag{3}$$

where F is the cumulative distribution function of a continuous unbounded distribution and τ is a vector of $C - 1$ ordered thresholds. Where F is the standard logistic distribution the resulting IRT model is known as the graded response model (Samejima, 1969).

Irrespective of the specific model applied, an IRT model allows for data from a given assessment method to be decomposed into an estimate of the characteristic of the individual (i.e., their *ability*, θ_p) that is invariant to the specific characteristics of the assessment method employed. Under this model, students' total scores or grades are a sufficient statistic for inference about θ_p. That is, the total scores or grades contain all the information there is about θ_p. Despite this, inference through IRT models regarding the θ_p space differs from inference in the total scores or grades space as under CTT. If the model holds for a given collection of students and items, θ_p are measures for invariant comparisons of persons on the logit scale with regard to expected performance in the logit scale. Where there is variance between either populations or items with respect to the probabilities of a particular response at a given level of θ_p then it is that differential item functioning (DIF; or for the whole test differential test functioning [DTF]) is present. DIF/DTF is a statistical characteristic of an item or test that shows the extent to which the item/test might be measuring different abilities for members of separate subgroups. In the case of assessments involving multiple markers, DIF/DTF can be used to examine agreement and measurement equivalence (i.e., whether both markers are measuring the same thing).

1.3 Aim of the Present Work

Given the two measurement paradigms outlined, CTT and IRT, the aim of this work is to examine the blind double marking approaches as currently employed in final year dissertation projects. We will first begin by exploring the agreement between first and second markers for the actual grades awarded under the CTT paradigm. Following this, we will explore the use of IRT models for such data and examine the impact of first and second markers upon the underlying latent ability estimates for students in addition to DTF between markers. Finally, we will offer some recommendations for practices in assessment through dissertation projects moving forwards.

2 Data Description and Preparation

We manually extracted data from available first and second marker feedback sheets across the academic years 2019-2020 to 2021-2022 from a range of courses broadly falling under the disciplines of the *sport and exercise sciences* at the lead authors institution where they lead the dissertation module for these courses[3]. A 'grademark' system is employed at the lead authors institution for marking assessment methods. This is

[3] Years prior to this were not available due to online management systems having been changed. Further, not all sheets were available for all students presumably due to staff not uploading them to the relevant folders for external examiners (which is where they should be uploaded and we accessed them from). In some cases it was not clear from the sheets who the markers were (i.e., names were omitted), or the marker had not awarded grades by section, and so where this was the case they were excluded.

in essence an ordinal scale comprising 18 categories ranging (see https://osf.io/v87de)[4]. We transformed the grademarks to their numeric grades so that the software used for analysis (R, version 4.2.1, "Funny-Looking Kid", The R Foundation for Statistical Computing, 2022) automatically recognised the ordering of categories. To maximise the amount of data available we have extracted the grades by each section of the assessment (i.e., Abstract, Introduction, Methods, Results, Discussion, and Structure/Presentation etc.) covered by the marking rubric (see https://osf.io/2xmtc). We then structured the data in long format meaning each row contained a single observation; in this case, a grade for a single section of the assessment for a single student provided by a single marker. Each student and marker were given an independent numerical identifying code treated as a factor. This resulted in a long dataset containing 3456 grades in total from 288 students and 36 markers. The anonymised dataset used for analysis is available on the Open Science Framework project page for this article (see https://osf.io/pj9sg) in addition to the analysis script (see https://osf.io/vzy8n). Note, the nature of this work meant that it was granted exemption from institutional ethics committee approval according to the lead authors institutions guidelines.

3 Absolute Agreement under Classical Test Theory Between Markers for Observed Grades

We begin by first examining the absolute agreement between first and second markers for the observed (that is to say, awarded) grades. As noted below, given the assumptions implied in the structure of the data generating process, traditional approaches to calculation of Cohen's κ would not be appropriate. Thus an alternative formulation was employed.

3.1 Model-based Cohen's κ like index

Given the hierarchical structure of the dataset (i.e., that there were multiple grades per student due to the multiple sections according to the rubric, nested within multiple markers due to the first and second marker), and that the number of markers in total exceeded two (not every student was marked by the same two markers), we used a model-based approach to obtain a Cohen's κ like index using the framework of ordinal generalised linear mixed models. We analysed the response variable as grademark using the **ordinal** package with the probit link function, with random intercepts for both student and for marker and also a fixed effect for section. Following the methods of Nelson and Edwards (2015) we calculated both ρ and their κ like index (κ_m) and also report the observed probability of agreement (p_o).

[4] Note, we assume an ordinal scale here though it is in fact not entirely clear how best to actually categorise the grademarking system this institution employs. To the best of our knowledge it is *intended* to be used as an ordinal scale by markers, yet is overtly linked to an underlying bounded scale $[0, 100]$ with known thresholds (i.e., the numerical mark that each grademark is transformed to). Typically an ordinal scale is assumed to reflect some latent underlying continuous variable with a number of ordered categories with unknown threshold values. However, the thresholds are known by the markers using the grademarking system (i.e., the numeric equivalents). It is unclear the extent to which knowledge of the thresholds impact the use of the ordinal categories by the marker. We would note that it seems strange from an ontological perspective to have a grade for an essay based assessment as part of a bounded underlying continuous distribution. Part of the reason that staff seemingly rarely award A1 grades in the system employed here is because of the knowledge that this amounts to a numerical grade of 100 out of a possible 100 (i.e., a perfect grade). This suggests that they actually acknowledge that there is no upper bound, at least practically speaking, to the underlying latent ability that is being measured through the operationalisation of the grademark awarded to an essay assessment. So, for the purposes of this exploration we have assumed the scale to be truly ordinal. Some assumptions were made for data imputation in certain cases. For example, some staff used slightly different grades e.g., low B, B, and high B. Across staff members using this approach we assumed that this was meant to equate to [grademark]3, [grademark]2, and [grademark]1 respectively across grademark boundaries D to B. Given that from experience most staff seem to state that they do not award A1 grades (given they equate to a numeric 100) we assumed that for A grades they equated to A4, A3, and A2 respectively. For F grades we took the opposite symmetrical assumption i.e., F3, F2, F1. At least one member of staff also gave percentage grades weighted by section which were converted (rounding up) to the nearest grademark category. Data was imputed in this manner in <5% of grade cases.

3.2 Correlation and Agreement Between First and Second Markers

Figure (3) below presents a heat map across the first and second markers and the frequencies with which certain grade pairings were awarded. As can be seen, visually at least there is some degree of rank correlation between the first and second markers awarded grades; that is to say that typically where the first marker awarded a higher grade so too did the second marker, and vice versa. The estimated value for ρ gives the ratio of the random effects variances; a natural measure of the variability between students relative to the variability between markers. Given ρ is close to 1 (in this case 0.803 with a standard error of 0.013) this means that the majority of variability in the model comes from between students which is what might be naturally expected.

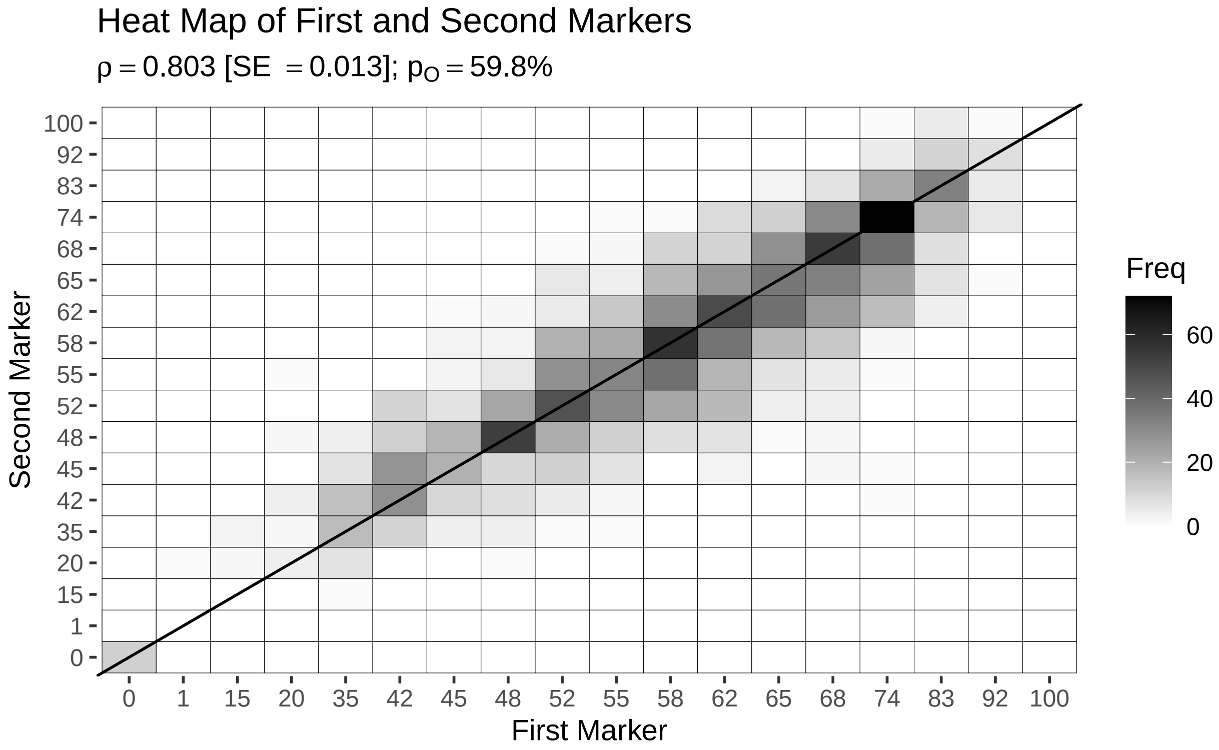


Figure 3: Heat map of first and second marker grademark pairings.

However, while evidently there is a relationship between first and second markers grades and less variation between markers compared to between students, it is the actual degree of agreement in grades awarded that we are interested in for this analysis. The directly observed probability of agreement between first and second markers was 59.8%. But, accounting for the probability of chance agreement, the κ_m was only 0.114 [95% Confidence Interval: 0.033 to 0.195]. Figure (4) presents the κ_m point estimate and 95% confidence interval alongside visualization of the qualitative thresholds suggested by Landis and Kock (1977) for interpretation of Cohen's κ like indexes.

Given that κ_m is in essence a form of intraclass correlation coefficient (ICC) for categorical data, we can use the Spearman-Brown prophecy formula to determine what number of markers we would estimate are needed to boost the current agreement to more acceptable levels (Warrens, 2017). Figure (5) presents the number of markers needed from between target values of $\kappa_m = 0.2$ and $\kappa_m = 0.8$ indicating "Fair", "Moderate", and "Substantial" agreement bands respectively according to Landis and Kock's (1977) thresholds. To achieve the minimum threshold for "Fair" agreement we would need to increase to an estimated 5 markers [95% Confidence Interval: 3 to 20 markers], for "Moderate" 12 markers [95% Confidence Interval: 6 to 45 markers], and for "Substantial" 31 markers [95% Confidence Interval: 16 to 119 markers].

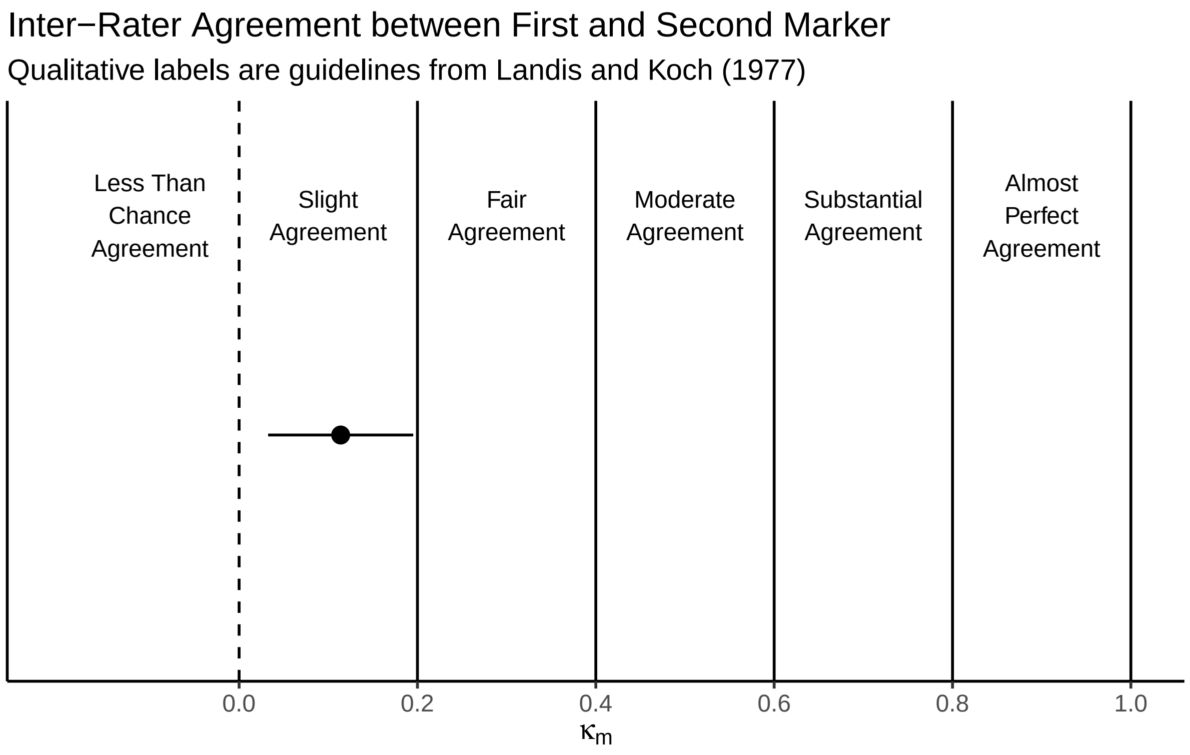


Figure 4: Inter-rater agreement between the first and second marker.

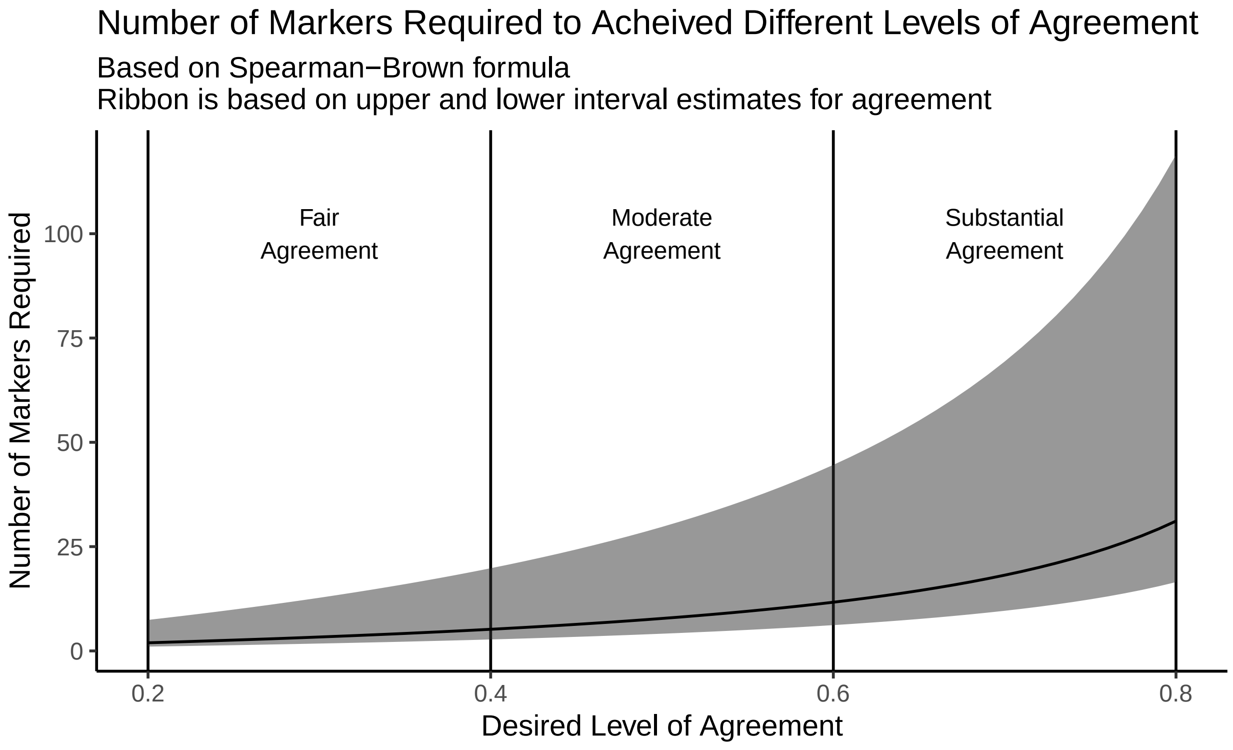


Figure 5: Number of marker needed for different levels of agreement calculated using the Spearman-Brown prophecy formula.

3.3 Summary of Absolute Agreement under Classical Test Theory Between Markers for Observed Grades

Whilst the data examined appear to indicate a reasonably strong relationship between first and second markers grades for dissertation projects, with most variance due to between student heterogeneity, the level of absolute agreement between markers whilst accounting for chance agreement appears to be unacceptably low. It should also be noted that this is likely an *overestimate* of the true agreement. Whilst marking is supposed to be independent, and indeed we have assumed this to be the case for this analysis, it is known at least anecdotally from colleagues that second markers attempt to anticipate the grade awarded by the first marker to ensure discussion for overall grade is simpler. Further, it is also known that during discussion markers physically alter their grades on marking sheets to fall closer in agreement with one another with respect to the overall grademark awarded; as such, even if they did initially mark independently, the data extracted from the marking sheets may have been altered towards agreement prior to the uploading of the marking sheets. Considering this we find it hard to believe that any reasonable academic educator, let alone student, would feel that the agreement reported here was acceptable. The poor agreement between markers identified under the CTT paradigm of course raises concerns regarding the validity of grading dissertations or other extended constructed response based assessments. However, this validity refers to the concept of the *true grade* under CTT. If instead it is not grades *qua* grades that we are interested in then agreement between markers may hold more strongly where the underlying latent *ability* is concerned and examined through IRT models which we next explore.

4 The Effect of First and Second Markers on Student Latent Ability Scores

Before we begin to explore the effects of markers on students' latent abilities through IRT models, as with any mathematical model applied to data there are a set of assumptions which need to be explored. One key assumption of most IRT models is that the assessment method employed measures only one underlying dominant latent ability i.e., *unidimensionality* [5].

4.1 Assumption of Unidimensionality

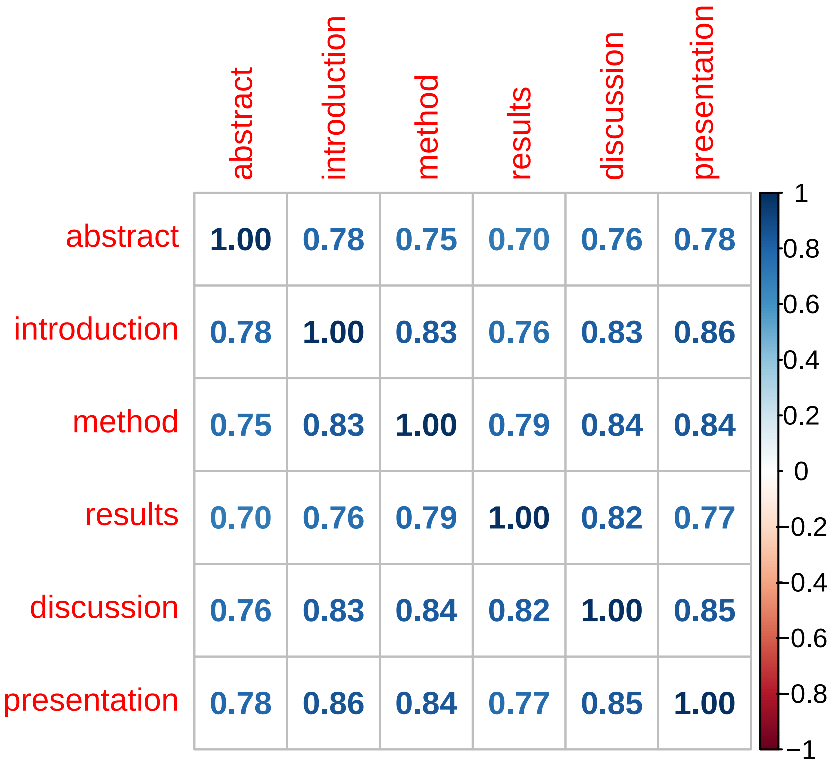
In the case of our grademark data for sections in the dissertation assessment we might refer to this single underlying latent variable broadly speaking as some underlying "dissertation ability", or perhaps "independent research ability". We can explore this assumption through the use of a classical item analysis such as exploratory factor analysis. First we examine the inter-correlations among the grademark numeric equivalents and see strong correlations between all sections (see figure (6).

A Kaiser-Meyer-Olkin (KMO) statistic was also used to examine sampling adequacy of each variable and determine factorability (Kaiser & Rice, 1974). In our case it is 0.94 which is deemed to be 'excellent'[6]. So we proceeded to fit an exploratory factor analysis to determine the number of factors needed to explain our variables. Figure (7) plots their eigenvalues to see how many exceed a value of 1 (Cattell, 1966; Kaiser, 1960). The largest eigenvalue for the first factor is over five times larger than the second indicating its clear dominance[7]. This supports our assumption of unidimensionality. The next step then is to fit the IRT

[5] Strictly speaking this is not an assumption of all IRT models as there are methods that can be employed to handle multidimensional data where assessment methods are found to estimate a range of separate abilities.

[6] Kaiser proposed with more than a little flair that a KMO > 0.9 was "marvelous", in the 0.80s, "meritorious", in the 0.70s, "middling", in the 0.60s, "mediocre", in the 0.50s, "miserable", and less than 0.5 would be "unacceptable".

[7] Of course, there is some conceptual assumption here too; we could argue that, whilst there are likely to be associations between each section of the dissertation, each section measures a different underlying ability. However, we ultimately award a single overarching grade anyway and so the behaviour of assessment in this fashion at least to some extent implies we believe some dominant underlying ability might be overarching which other more nuanced abilities nest within (perhaps akin to Spearman's g factor, or general intelligence). Had there been evidence though of additional factors then we might still be able to apply multidimensional IRT models designed for such instances.



	abstract	introduction	method	results	discussion	presentation
abstract	1.00	0.78	0.75	0.70	0.76	0.78
introduction	0.78	1.00	0.83	0.76	0.83	0.86
method	0.75	0.83	1.00	0.79	0.84	0.84
results	0.70	0.76	0.79	1.00	0.82	0.77
discussion	0.76	0.83	0.84	0.82	1.00	0.85
presentation	0.78	0.86	0.84	0.77	0.85	1.00

Figure 6: Inter-correlations between different section numeric equivalent grades.

model(s) and examine how well they explain the data, in addition to exploring the impact of the first and second markers upon student ability estimates and agreement through DTF.

4.2 Fitting the Graded Response Model(s)

The graded response model (GRM) introduced earlier (see equation (3)) seems to be an appropriate choice for this dataset given the nature of the ordinal response variable. Thus we fit a series of GRMs including the 1PL form (GRM-1PL) where a single item parameter for difficulty (β) is included, the same model but with ordinal threshold locations allowed to vary across items (GRM-VAR), and the 2PL model (GRM-2PL) which also includes the item parameter for discrimination (α). Further, as the dataset included a number of different first and second markers varying across students, an additional random intercept was included for the marker thus incorporating variance due to this and enhancing generalisability of the model across markers. This initial modelling was performed without the covariate of the first or second marker to determine which model type would best fit. Analysis was conducted in the Bayesian hierarchical regression framework (Bürkner, 2020) using the **brms** package and the probabilistic programming language **Stan**. Weakly regularising priors were applied to aid model convergence and restrict certain parameters to sensible values, with ability scores scaled to a mean of 0 and standard deviation of 1. The models used four Monte Carlo Markov Chains with 1000 warmup and 3000 sampling iterations. The three models were compared formally using approximate leave-one-out cross-validation (LOO-CV) and the differences in the expected log pointwise predictive probabilities for the discrete models (ELPD; Vehtari et al. (2017)). Table 1 shows this comparison.

The GRM-2PL model is the best fit to the data, and the difference in ELPD between it and the next best fitting model (GRM-1PL) is over three times the standard error. So for the remainder of exploration we used the GRM-2PL model. To give an impression of what exactly the model is estimating from this dataset, figure (8) below shows the distribution of ability estimates in addition to the latent ordinal thresholds for

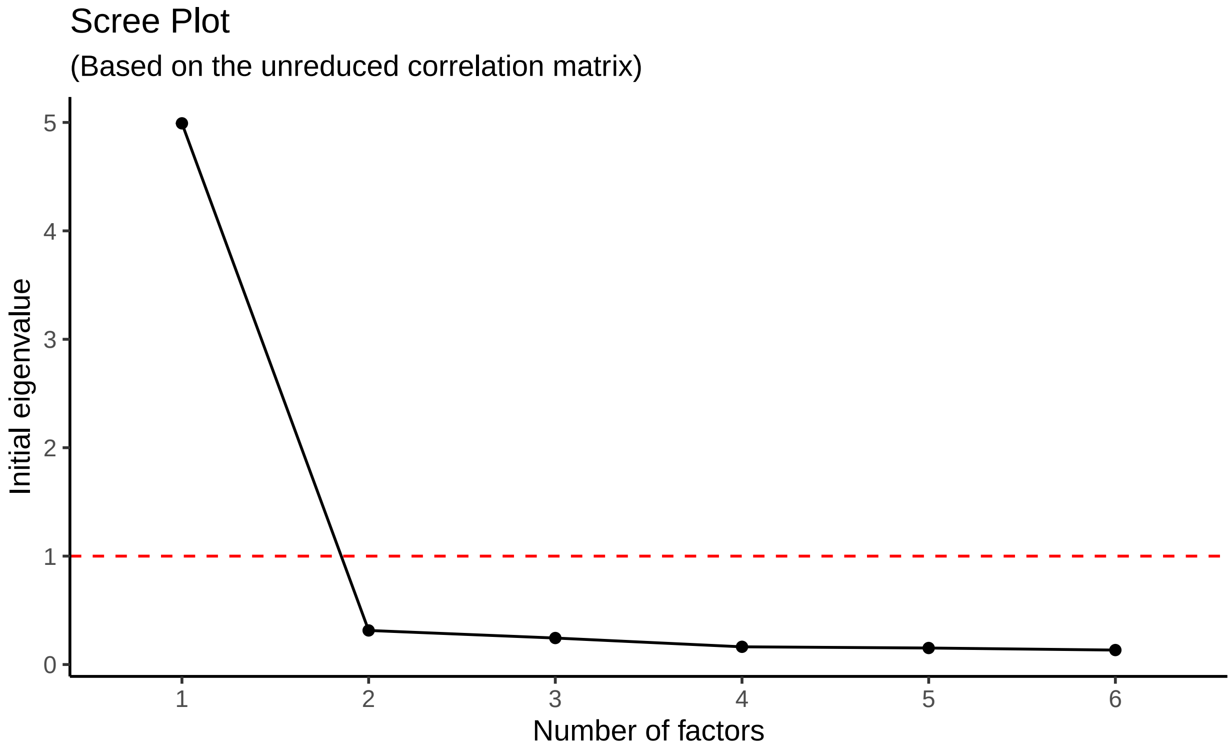


Figure 7: Scree plot of eigenvalues from exploratory factor analysis of grade data.

Table 1: Results of leave-one-out cross-validation

Model	**ELPD Difference**	**SE of Difference**
GRM-2PL	0.00	0.00
GRM-1PL	-23.99	7.58
GRM-VAR	-49.02	13.21

Note:

ELPD = Expected log pointwise predictive probabilities;
SE = Standard error

the model for each grademark boundary.

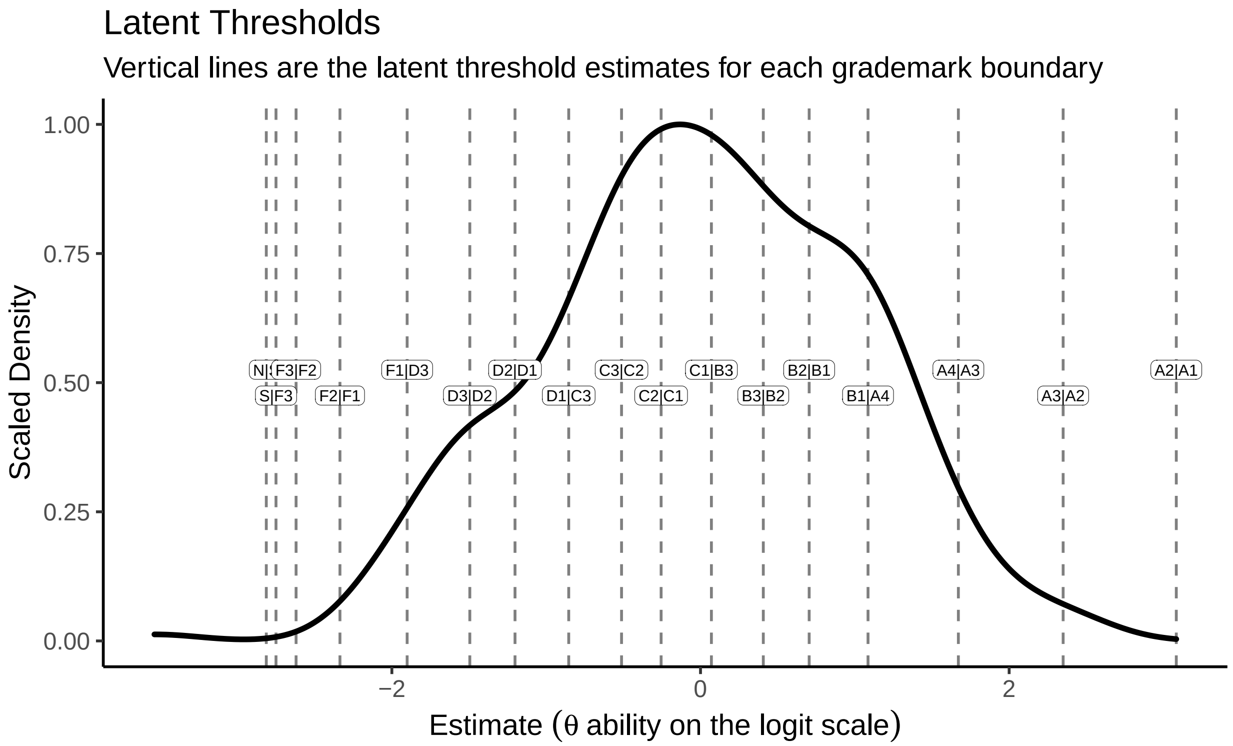


Figure 8: Latent thresholds for each grade boundary in the grademark scale from the GRM-2PL model with ability estimates overlaid.

4.3 Exploring First and Second Marker Impact on Model Estimates

Next we proceeded to employ a DIF/DTF approach to examine agreement and measurement equivalence across grades using the GRM-2PL model (i.e., whether both markers are measuring the same thing). A categorical covariate for whether the grades were awarded by the first or second marker was added to the GRM-2PL. This was allowed to vary (i.e., random slopes were included) over students to allow determination of the extent to which first or second markers differed in their estimation of each students latent ability estimates. In essence this meant the model yielded two ability estimates, one from each marker such that the intercorrelation between the estimates can be thought of as reflecting the degree to which their conception of ability converged whilst allowing for measurement error. The covariate for first or second marker was also allowed to vary over section characteristics including difficulty and discrimination. This facilitated exploration of DTF (in this case we focus on the overall assessment) and whether measurement equivalence could be determined between first and second markers.

The mean of each student's ability estimate and their 95% quantile intervals were extracted for both the first and second marker and visualised to explore their degree of agreement (see figure (9)). In addition an IRT based empirical reliability coefficient (essentially similar to the ICC) for each of the markers (first and second) was calculated from model point estimates and errors of ability.

Visually there appeared to be a fairly strong relationship between both the first and second marker (the Pearson's correlation is 0.93). The reliability of individual estimates from each marker are both fairly high and depending on what guidelines for ICCs are used are either both 'excellent' (Cicchetti, 1994), or 'excellent' and 'good' respectively (Koo & Li, 2016).

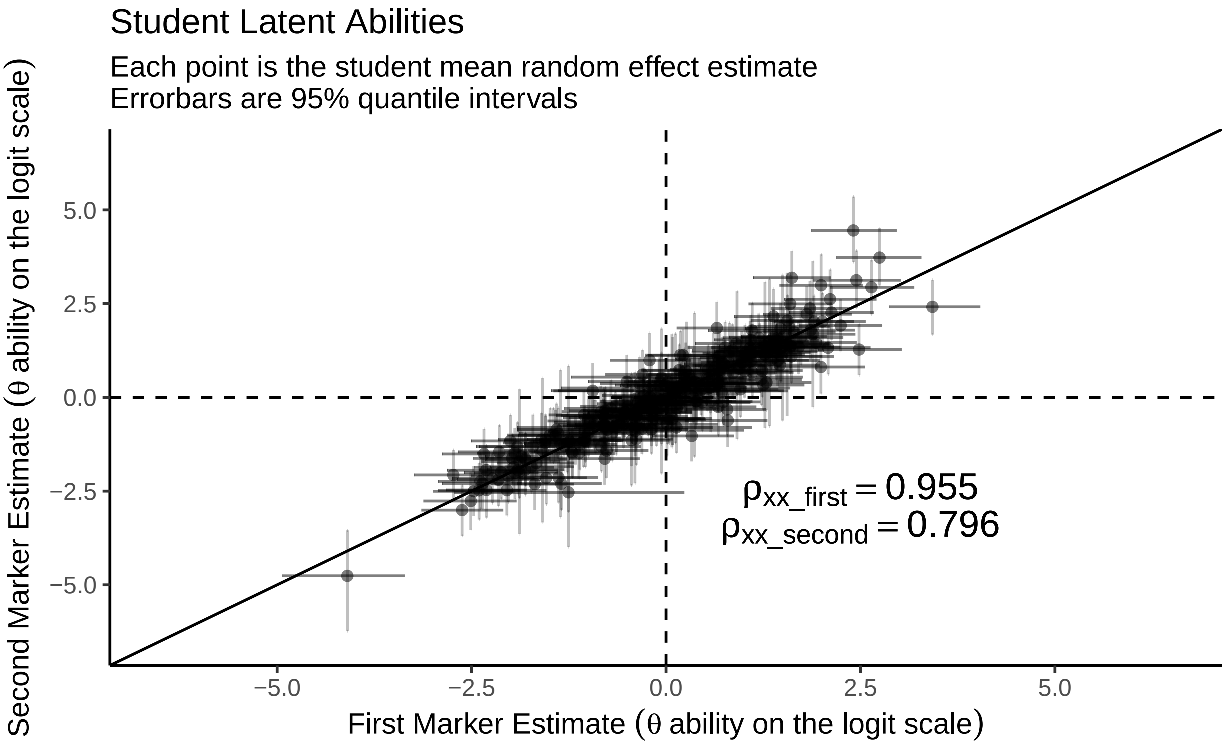


Figure 9: Relationship between first and second marker ability point and interval estimates from the GRM-2PL model with marker as covariate.

However, whether or not the first and second markers do indeed *agree* with one another, that is to say they are measuring the same thing, is explored through DIF/DTF. Conversion of ability estimates to true score estimates (i.e., the true grademark that would be predicted given the underlying ability of the student) based upon the different model parameter estimates for each item (section) between first and second marker allows for exploration of measurement equivalence (Barr & Raju, 2003). According to Raju et al. (1995) the central theme of the DIF/DFT framework in this regard is to find a true score for a person when rated by one source and to compare that to their true score when rated by another source. Where true score estimates are equal for all levels of ability it can be said that there is measurement equivalence across first and second markers.

After fitting a GRM to a given dataset point estimates of individual true scores t_p can be obtained as follows per item (section):

$$t_{pi} = 1 + P(y_{pi} \geq 2) + P(y_{pi} \geq 3) + \cdots + P(y_{pi} \geq C) \tag{4}$$

This is in essence a summing of the probabilities according to each category threshold estimated from the model. In this case we are not interested in each section *per se* but instead the overall assessment. So to obtain t_p for the assessment overall according to ability estimates we can merely sum t_{pi} for each section. Given we have a Bayesian model we took samples from the posterior distribution for each parameter, calculated the relevant threshold probabilities given the GRM-2PL model, then we took a mean and 95% quantile interval for the probabilities and calculated the corresponding true score estimates and intervals (see figure (10)).

As can be seen from visual comparison of the true score estimates in figure (see figure 10), there appears to be measurement equivalence between first and second markers. We can explore this also by examining the differences in true score estimates across the range of ability estimates. The majority of grademark categories in the system used for this dataset are 3 numeric equivalents apart (e.g., C3 = 52, C2 = 55, and C1 = 58) and indeed this is the smallest difference between grades. As such, we would accept measurement

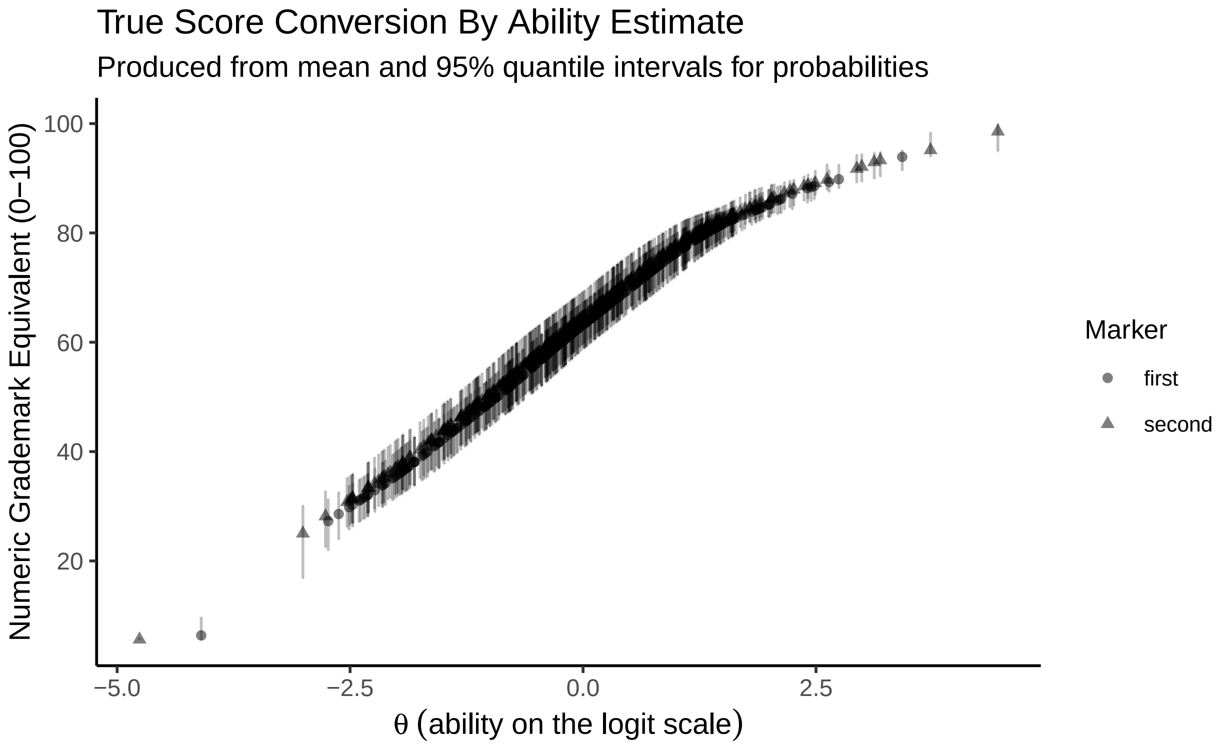


Figure 10: True score estimates across first and second marker ability estimates from the GRM-2PL with marker as covariate.

equivalence where the true score estimated did not exceed this range (i.e., the difference should be <3) at any ability level. Figure (11) shows the differences between the first and second marker are within this measurement equivalence interval at all levels of ability. Given this comparison our model implies that, while first markers tend to assess ability to be slightly lower on average across all levels of ability, true score estimates at a given ability level from this model would never be more than a single grademark category apart. Thus, from the perspective of an IRT paradigm we appear to have measurement equivalence between first and second markers.

4.4 Summary of The Effect of First and Second Marker on Student Latent Ability Scores

Considering the current dataset under the IRT paradigm offers a different insight compared with the admittedly bleak one yielded from a solely CTT based examination of agreement. The grades awarded to dissertations across sections seems to be indicative of a dominant latent ability; likely the higher level thinking skills required of independent research projects. Further, a GRM-2PL was a good fit overall for the data. Whilst raw grademarks showed poor agreement between first and second markers when examined through CTT derived κ_m, there appeared to be a strong relationship between estimates of latent ability between first and second markers. Further, examination of DIF/DFT revealed that true score estimates are equal (or at least within a practically equal interval) for all levels of ability and thus it can be said that there is measurement equivalence across first and second markers. Given that the source of grademark, either first or second marker, had negligible impact on the measurement of the underlying latent ability score it seems plausible that the GRM-2PL parameter estimates produced from this historical data could be used to provide a measurement model for future dissertation grading without the need for multiple markers. Given the

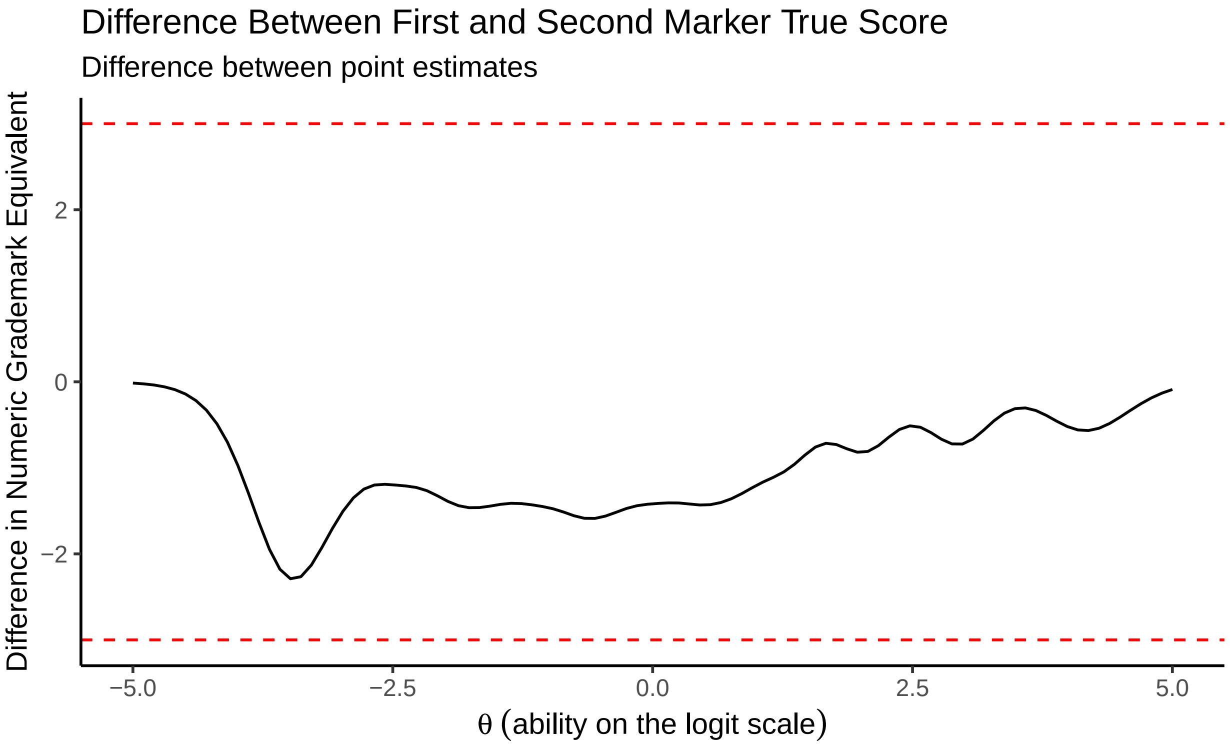


Figure 11: Difference in true score estimates between first and second markers across a range of ability (-5,5) from the GRM-2PL with marker as covariate.

Table 2: Example of true score/grade determination from ability estimates using the GRM-2PL model

Student	Abstract	Introduction	Methods	Results	Discussion	Presentation	Theta	True Score	Grademark
High Ability	65	74	83	68	74	83	1.91	95	A1
Mid Ability	55	52	62	52	55	58	-0.04	67	B1
Low Ability	0	42	35	20	35	42	-1.66	37	F1

connection between CTT and IRT, and thus true scores and ability, regarding polytomous items (Raykov et al., 2019) first marker grades alone could be used to estimate latent ability scores via maximum likelihood or other estimators which could subsequently be converted to true scores (i.e., numeric grademark equivalents) and thus grademark categories (rounding up or down appropriately). For example (see table (2)), we could imagine a pattern of grademarks awarded by a marker by section for a 'high', 'mid', and 'low' ability student respectively. Each grade awarded is only an indirect and error laden observation of the constructed response that the students underlying ability gives rise to. From these responses however we can employ our IRT measurement model (in this case the GRM-2PL) and using the item (section) parameters from the model estimate θ_p for each student. We can then convert their ability estimates to the corresponding true score estimates and award these as the students final grade.

5 Conclusion

Within this work we have considered the nature of measurement that takes place through the application of an extended constructed response assessment method, namely the dissertation, and the process of double blind marking that occurs with it. In particular, questions relating to reliability/agreement and implications for measurement validity have been explored through both CTT and IRT approaches. Examining grade data through these two paradigms reveals different conclusions and implications regarding measurement inferences about what students know and can do. Within the operationalist framework of CTT there appears to be poor agreement between markers regarding grades *qua* grades implying that validity regarding the latent true grades is likely to be threatened. However, when considering marker agreement regarding the underlying latent ability of students that gives rise to extended constructed responses, such as dissertations, through an IRT framework there appeared to be little evidence of DFT with practical measurement equivalence between first and second marker. As such, a GRM-2PL model estimated from historical data could be used going forward to generate true score estimates from ability estimates for the purposes of more valid grading of student dissertations. Given the lack of DFT it further suggests that we might be able to do away with the laborious and unnecessary process of double marking as it currently stands.

In the future we may also be able to use IRT models to empirically evaluate the applications of other methods of assessment. For example, as dissertations are essentially research projects similar to those that would be reported in academic outlets such as peer reviewed journals, checklist-based approaches to grading might be appropriate. At least in the disciplines explored here (i.e., sport and exercise sciences), if students have completed a research project of a particular kind then it would be fairly simple to build marking criteria around typical reporting guidelines/templates used by most academic journals[8]. These can guide what aspects should be included in a given assessment dependent on the type of study for which they are given a grade or mark accordingly. These checklist items would likely also help to callibrate markers grading whereby they essentially grade on whether students have, or have not, included relevant aspects for a given type of research project. That is to say they form binary response items. IRT models can combine and calibrate tests that include both objective selected response items, such as from a checklist, and constructed response items (Ercikan et al., 1998; Thissen et al., 1995). Thus, we could have a final section of grading that gives over to the general qualitative impression from the marker of the quality of developed rationale, interpretation of findings etc. to accompany the basic reporting required of the kind of study design conducted.

A final consideration, and one aspect we have not considered in the present exploration, is the weighting of different items/sections in a given assessment method. For most assessments in higher education it is typical to weight sections differently and this is indeed the case for the dissertation example explored here ("Abstract" = 5%, "Introduction" = 20%, "Methods" = 15%, "Results" = 20%, "Discussion" = 30%, "Presentation" = 10%). IRT models can also incorporate true score estimates for the whole assessment method weighted by item/section (Stucky, 2009), though their inclusion may or may not necessarily improve precision of ability estimation (Gordon et al., 2012). It is perhaps also worth noting thought that by weighting sections in terms of their contribution to the overall grademark there appears to be an implicit Calvanist work ethic valued here; that is to say that the weighting reflects the *amount* of work done, and not the extent to which a given section provides information regarding the underlying latent ability of the student. As such, assessors should also reflect on whether they merely wish to measure a students underlying ability, or whether they also wish to determine to some extent the desert of a student in awarding a particular grade to their response.

6 References

Almond, R. G. (2014). Using Automated Essay Scores as an Anchor When Equating Constructed Response Writing Tests. *International Journal of Testing*, *14*(1), 73–91. https://doi.org/10.1080/15305058.2013.816309

Barr, M. A., & Raju, N. S. (2003). IRT-based assessments of rater effects in multiple-source feedback instruments. *Organizational Research Methods*, *6*(1), 15–43. https://doi.org/10.1177/1094428102239424

Bloxham, S. (2009). Marking and moderation in the UK: False assumptions and wasted resources. *Assessment & Evaluation in Higher Education*, *34*(2), 209–220. https://doi.org/10.1080/02602930801955978

Bridgman, P. (1927). *The Logic of Modern Physics.* Arno Press.

Brookhart, S. M., Guskey, T. R., Bowers, A. J., McMillan, J. H., Smith, J. K., Smith, L. F., Stevens, M. T., & Welsh, M. E. (2016). A Century of Grading Research: Meaning and Value in the Most Common Educational Measure. *Review of Educational Research.* https://doi.org/10.3102/0034654316672069

[8] We would argue most staff supervising research projects *should* be already largely aware of such reporting guidelines as CONSORT, COSMOS, COREQ, SRQR etc. (see https://www.equator-network.org/) and as such the use of them in structuring assessment should be a relatively easy ammendment to current processes.

Bürkner, P.-C. (2020). *Bayesian Item Response Modeling in R with brms and Stan* (No. arXiv:1905.09501). arXiv. http://arxiv.org/abs/1905.09501

Cattell, R. B. (1966). The Scree Test For The Number Of Factors. *Multivariate Behavioral Research*, *1*(2), 245–276. https://doi.org/10.1207/s15327906mbr0102_10

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment*, *6*(4), 284–290. https://doi.org/10.1037/1040-3590.6.4.284

Edgeworth, F. Y. (1888). The Statistics of Examinations. *Journal of the Royal Statistical Society*, *51*(3), 599–635. https://www.jstor.org/stable/2339898

Ercikan, K., Sehwarz, R. D., Julian, M. W., Burket, G. R., Weber, M. M., & Link, V. (1998). Calibration and Scoring of Tests With Multiple-Choice and Constructed-Response Item Types. *Journal of Educational Measurement*, *35*(2), 137–154. https://doi.org/10.1111/j.1745-3984.1998.tb00531.x

Gordon, D., Howe, L. D., Galobardes, B., Matijasevich, A., Johnston, D., Onwujekwe, O., Patel, R., Webb, E. A., Lawlor, D. A., & Hargreaves, J. R. (2012). Authors' Response to: Alternatives to principal components analysis to derive asset-based indices to measure socio-economic position in low- and middle-income countries: The case for multiple correspondence analysis. *International Journal of Epidemiology*, *41*(4), 1209–1210. https://doi.org/10.1093/ije/dys120

Hemmings, S. (2001). The place of the dissertation in learning to research. In R. Humprey, C. Middleton, R. Finnegan, S. Hemmings, & D. Phillips (Eds.), *Learning to Research: Resources for Learning and Teaching in Sociology and Social Policy* (Vol. 1&2). University of Sheffield.

Hornby, W. (2003). Assessing Using Grade-related Criteria: A single currency for universities? *Assessment & Evaluation in Higher Education*, *28*(4), 435–454. https://doi.org/10.1080/0260293032000066254

Kaiser, H. F. (1960). The Application of Electronic Computers to Factor Analysis. *Educational and Psychological Measurement*, *20*(1), 141–151. https://doi.org/10.1177/001316446002000116

Kaiser, H. F., & Rice, J. (1974). Little Jiffy, Mark Iv. *Educational and Psychological Measurement*, *34*(1), 111–117. https://doi.org/10.1177/001316447403400115

Koo, T. K., & Li, M. Y. (2016). A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *Journal of Chiropractic Medicine*, *15*(2), 155–163. https://doi.org/10.1016/j.jcm.2016.02.012

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159–174.

McIntosh, S., McKinley, J., Milligan, L. O., & Mikolajewska, A. (2022). Issues of (in)visibility and compromise in academic work in UK universities. *Studies in Higher Education*, *47*(6), 1057–1068. https://doi.org/10.1080/03075079.2019.1637846

Michell, J. (1997). Quantitative science and the definition of measurement in psychology. *British Journal of Psychology*, *88*(3), 355–383. https://doi.org/10.1111/j.2044-8295.1997.tb02641.x

Mislevy, R. (2017). On Measurement in Educational Assessment. In *Handbook on Measurement, Assessment, and Evaluation in Higher Education* (pp. 37–64). Routledge.

Nelson, K. P., & Edwards, D. (2015). Measures of agreement between many raters for ordinal classifications. *Statistics in Medicine*, *34*(23), 3116–3132. https://doi.org/10.1002/sim.6546

Nyamapfene, A. (2012). Involving supervisors in assessing undergraduate student projects: Is double marking robust? *Engineering Education*, *7*(1), 40–47. https://doi.org/10.11120/ened.2012.07010040

Pan, W., Cotton, D., & Murray, P. (2014). Linking research and teaching: Context, conflict and complementarity. *Innovations in Education and Teaching International*, *51*(1), 3–14. https://doi.org/10.1080/14703297.2013.847794

Raju, N. S., Linden, W. J. van der, & Fleer, P. F. (1995). IRT-Based Internal Measures of Differential Functioning of Items and Tests. *Applied Psychological Measurement*, *19*(4), 353–368. https://doi.org/10.1177/014662169501900405

Raykov, T., Dimitrov, D. M., Marcoulides, G. A., & Harrison, M. (2019). On the Connections Between Item Response Theory and Classical Test Theory: A Note on True Score Evaluation for Polytomous Items via Item Response Modeling. *Educational and Psychological Measurement*, *79*(6), 1198–1209. https://doi.org/10.1177/0013164417745949

Raykov, T., & Marcoulides, G. A. (2016). On the Relationship Between Classical Test Theory and Item Response Theory. *Educational and Psychological Measurement*, *76*(2), 325–338. https://doi.org/10.

1177/0013164415576958

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika*, *34*(1), 1–97. https://doi.org/10.1007/BF03372160

Scriven, M. (2017). The failure of higher education to follow the standards it has established in methodology and evaluation. In *Handbook on Measurement, Assessment, and Evaluation in Higher Education* (pp. 27–26). Routledge.

Stucky, B. D. (2009). *Item response theory for weighted sum scores* [PhD thesis]. University of North Carolina at Chapel Hill.

Thissen, D., Pommerich, M., Billeaud, K., & Williams, V. S. L. (1995). Item Response Theory for Scores on Tests Including Polytomous Items with Ordered Responses. *Applied Psychological Measurement*, *19*(1), 39–49. https://doi.org/10.1177/014662169501900105

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, *27*(5), 1413–1432. https://doi.org/10.1007/s11222-016-9696-4

Warrens, M. J. (2017). Transforming intraclass correlation coefficients with the Spearman-Brown formula. *Journal of Clinical Epidemiology*, *85*, 14–16. https://doi.org/10.1016/j.jclinepi.2017.03.005

Winstone, N. E., & Boud, D. (2022). The need to disentangle assessment and feedback in higher education. *Studies in Higher Education*, *47*(3), 656–667. https://doi.org/10.1080/03075079.2020.1779687