

# Investigating level descriptors and their impact on validity

Craig Grocott <sup>1</sup>, Eli Moe <sup>1</sup> & Norman Verhelst <sup>2</sup>

<sup>1</sup>University of Bergen, Norway

<sup>2</sup>Eurometrics, The Netherlands

This article presents a validation study which investigates the relationship between mastery level descriptors and item difficulty in the National Tests of English (NTE) in Norway. The aim is to establish the extent to which the descriptors indicate item difficulty and thus support the argument that the mastery levels are a reflection of the framework of the NTE. This argument strength has direct implications for validity and for the defensibility of moving from criterion-based descriptors to norm-based national data. The study involved a panel of 10 raters assigning level descriptors from 7 content categories to 80 test items, giving a total of 5,600 individual judgements. These judgements are compared to the real test scores of around 46,000 pupils to establish if the level descriptors assigned by raters can predict pupil performance on the tests. The results show a strong correlation between rater judgements and real test scores, meaning that the descriptors offer an indication of item difficulty. However, some individual descriptor categories contain deviations from the expected order. We conclude that, while the level descriptors reflect the test's framework, and thus support the validity argument, the argument could be strengthened with the revision of some individual descriptors.

**Keywords:** construct validity, national tests, validity, structural validity, Messick

---

Email address for correspondence: [craig.grocott@uib.no](mailto:craig.grocott@uib.no)

© The Author(s) 2024. This is an open access article distributed under the terms of the Creative Commons Attribution 4.0 International License, which permits the user to copy, distribute, and transmit the work provided that the original authors and source are credited. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

## Introduction

All language tests have some form of scoring system against which they are graded, which must be appropriate to the tests to support arguments made for the uses and interpretations that occur on the back of them (Kane, 2013). The way a test score is interpreted can broadly be divided into norm-based and criterion-based systems; the former involves a pre-determined, but flexible, number of test-takers being assigned each grade, and the latter involves test-takers being judged against a set of absolute standards or criteria (Lok et al., 2015). Sometimes, elements of both can be integrated, as is the case with the tests discussed in the present study: the National Tests of English (NTE), which are taken by virtually all fifth and eighth-grade pupils in Norway. This integration must serve as a 'bridge' between norm-based and criterion-based scoring systems, and its effectiveness has implications for validity. The present study seeks to examine this connection from a validity perspective by comparing the judgements of a panel of 10 expert raters, based on level descriptors, to the real test scores of over 45,000 eighth-grade pupils in Norway.

As is expanded upon in the theoretical background section, this study takes as its starting point Messick's (1996) presentation of the structural aspect of construct validity (referred to in the present article as 'structural validity'). Structural validity can be at risk if the scoring system used by a test is not an accurate reflection of the test's specified model(s). The study seeks to establish if this is the case for the NTE. The panel of raters was used to investigate the strength of the relationship between the descriptors which accompany the mastery levels used in the scoring system and the national results which are used for assigning mastery levels, and thus difficulty, to items.

The NTE are the subject matter for the present study. They are graded across three levels, known as mastery levels, (M1-M3) at fifth grade and five mastery levels (M1-M5) at eighth grade, with levels M1 to M3 common to both. This study used the eighth-grade NTE as its subject matter, thereby including all five mastery levels. The tests are digital tests of receptive skills, with a range of item formats which include text and images. The tests were exclusively reading tests up until 2021, at which point listening tests were introduced.

Upon creation, each item is assigned a mastery level according to its perceived difficulty by test developers at the University of Bergen (UiB). This perception of difficulty is mostly based on the descriptors which accompany the mastery levels. Test-makers assess what the item requires of the test-taker, and the mastery level whose descriptors are perceived to match closest to what the item requires is then assigned to the item.

The mastery levels, which were created by test developers at UiB on behalf of the Norwegian Directorate for Education and Training (*Utdanningsdirektoratet*, henceforth Udir), are intended to reflect what is expected of pupils in order to place them at a certain level within the field of English reading. The mastery levels have been slightly updated recently to reflect the fact that texts can be both written and spoken, although there is a large degree of overlap between listening and reading comprehension (Wolf et al., 2018). This applies especially to the NTE as the listening and reading aspects share some item formats (e.g., picture tasks), so the changes to the mastery levels were minor.

This article attempts to investigate the structural validity of the NTE and the scoring system by answering the following research question:

*To what extent do the mastery level descriptors used by the National Tests of English indicate the difficulty of test items and thus support the validity argument for the tests?*

The article addresses the question by first examining the theoretical background of validity theory and its relationship to scoring systems for language tests, as well as the relationship between criterion-based and norm-based assessment scores. It goes on to outline the method of the quantitative panel study. The results are then presented and discussed with relation to theory to offer a picture of the structural validity of the scoring system. The article concludes by summarising the findings and making recommendations for further research and potential changes to the mastery level descriptors.

The present study is relevant for all stakeholders in the NTE, because no such study has previously been conducted on the mastery levels, and it is relevant for the validity of the NTE as a whole. Given that the mastery levels which make up the scoring system

are accompanied by level descriptors, it is important that these descriptors can be empirically argued to be valid, as they form a key part of the information teachers receive about pupils' English receptive skills (Udir, 2022a). This information is intended to be a basis for the consequences of the NTE, namely contributing to formative assessment and quality development in the English school subject, thus necessitating a high level of accuracy and validity.

The study can also be of wider relevance to a variety of interested parties engaged in the testing of English or other languages as an L2: (1) those using scoring systems made up of levels with associated descriptors, such as those based on the Common European Framework of Reference (CEFR) (Council of Europe, 2020); (2) those concerned with the impact that a scoring system, and its associated descriptors, can have on the validity of a language assessment; and (3) those looking to investigate the validity of assessments which attempt to combine criterion-based and norm-based assessment, a practice for which models have been produced in recent years (e.g. Lok et al., 2015).

## **Theoretical background and previous research**

### **Validity background**

The present study is part of a wider project concerned with the validity of the NTE, which is largely based on three of Messick's (1996) six aspects of construct validity: the structural, the substantive, and the consequential aspects. Messick describes validity as "the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions on the basis of test scores or other modes of assessment" (1989, p. 13). His other three aspects of construct validity are content, generalizability, and external aspects (Messick, 1996). This study takes as a starting point the structural aspect of validity. This refers to the relationship between the scoring system used in a test and the specified model, along with evidence-supported constructs (Hasselgreen, 2004). Messick (1989) argues that, for a scoring system to be structurally valid, it needs to interact with what is already known about the construct of the test itself, namely what the individual items require of a test-taker. The measurement of structural validity is then the degree to which scores on a scale can be described as reflecting the construct being measured (Brown & Bonsaksen, 2018; Rios & Wells, 2014) and the strength of the argument for this being the case

(Kane, 2006). The scoring system and band scores therefore need to reflect the content of items. This means that multiple criteria are usually required to offer a full picture of the test's construct and to justify decisions made on the back of the scores (Bachman & Palmer, 1996). This can sometimes necessitate band scores with many descriptors attached to them, especially in cases of tests which seek to test multiple skills at multiple levels. Given that these descriptors can act as the starting point for the interpretations of a test-taker's ability, and thus intended or actual consequences, they clearly need to be featured in the validation process as a key part of an AUA – assessment use argument (Bachman & Palmer, 2010).

There is an argument that structural validity as employed in the present study overlaps with Weir's (2005) presentation of scoring validity, given that it is made up of aspects concerned with the scoring system. For instance, Weir argues that a key consideration for a scoring system in terms of validity is its clarity in specifying performance criteria to reduce the chances of subjectivity in marking. This is relevant for the present study as it seeks to establish the effectiveness of the mastery level descriptors in predicting item difficulty. However, Knoch and Chapelle (2017) point out that Weir's presentation of scoring validity as a *type* of validity runs contrary to Messick's (1989) presentation of construct validity as a unitary concept, with structural validity acting as a component of this larger concept. As the wider project uses Messick's concept of construct validity as its base, the use of structural validity is appropriate for the present study. Additionally, Messick's presentation of validity as a unitary concept has been largely embraced by the field (AERA, 2014; Knoch & Chapelle, 2017).

### **Scoring frameworks**

In the specific context of the NTE, the mastery levels were originally created based on knowledge of the tests' construct and previous results (NTE development team, personal communication, October 2021). The levels were accompanied by can-do descriptors, reflecting in part the CEFR, albeit tailored to reflect the specifications of the eighth-grade tests and what is expected of pupils after the seventh grade, according to the curriculum. The resemblance to an international framework such as the CEFR can act as a measure to ensure criterion-referenced validity (De Jong & Zheng, 2016), as well as making the band scores more understandable to external observers (Kane, 2012). The CEFR was intended to be a framework, or a model (Fulcher, 2016), but it

has increasingly been used as a standard upon which other scoring frameworks are based (Milanovic & Weir, 2010; Papp, 2018). However, some argue that the CEFR itself has been the subject of surprisingly few empirical studies which can underpin its validity (Alderson, 2007; Carlsen, 2014), and associated frameworks, such as the levels used in the NTE, could therefore be argued to have an increased need for empirical validation. This is especially true given that the CEFR descriptors have been subject to criticism for, among other things, gaps and inconsistencies and their lack of clear definition (Alderson et al., 2004; Figueras, 2012; Fulcher, 2003). Such criticisms can therefore extend to scoring systems and descriptors that resemble the CEFR. This is especially true for assessments such as the NTE, which are taken by young learners, as the CEFR was not designed with young learners in mind (Papp, 2018), and there have been calls for the CEFR to be adapted for local contexts (Harsch, 2019).

There has been a larger amount of research into linking external frameworks and scoring systems to the CEFR, a good overview of which can be found in Green (2018). Green notes that, when linking a test to a model such as the CEFR, different processes can produce different results. He therefore points to Kaftandjieva's (2004) call to use multiple methods to ensure that links would be as defensible as possible. In the case of the present study, it can be described as another method of validating the scoring system of the NTE in addition to an earlier standard setting procedure (Moe, 2008).

The standard setting for the NTE pre-dated the creation of the mastery levels and was used to create the original CEFR-based cut scores, using the Kaftandjieva and Takala Compound Cumulative Method (Kaftandjieva & Takala, 2002), which is a variation of the Angoff method (Angoff, 1971; Moe, 2008). A panel of test developers and teachers familiar with the NTE was used. This familiarity is vital for standard setting (Moe & Verhelst, 2017) or indeed for participation in a study such as the present one. The panel's judgements on the difficulty of individual items were compared to observed pupil performances during the piloting stage of the items, as well as in the 'real' tests, indicating a satisfactory degree of correspondence. This standard setting can be described as a precursor to the present study in that it used test data to help determine grade boundaries. However, there is a clear need for validation of the newer scoring system, namely the mastery levels. This is especially true given that the mastery levels consist of descriptors tailored to Norwegian eighth-grade pupils, rather than using the

more general CEFR descriptors. Conclusions from standard setting with CEFR-based band scores for the NTE are therefore less defensible.

### **The NTE**

While the NTE have been the subject of some validation studies (Grocott, 2022; Pižorn & Moe, 2012; Sibbern, 2013), none of these specifically focus on the mastery levels, with some studies pre-dating the creation of the current mastery levels. Pižorn and Moe (2012) do however point out that the mandate given to test developers upon the conception of the NTE was to create a test which can discriminate between pupils at a *range of levels* of competence. The test developers were asked to aim at an average p-value of 0.5 for the whole test, meaning the average student would answer approximately 50% of the items correctly. Initially, the test results were in the form of raw scores, i.e., the number of correct answers. From 2014, Udir wanted to measure trends, linking results from one year to another to see whether students' ability improved. This led to a change in the way the results were reported. Udir decided to report the results in five norm-referenced groups using the percentages 10-20-40-20-10 the first year and calling the groups M (mastery level) 1, 2, 3, 4 and 5. Since the same cut scores (between the levels) were used the following years, the percentages of students' results assigned to the different levels would vary slightly each year.

Another challenge the test developers received was giving the mastery levels content, i.e., describing what students assigned to different levels could do. Trying to describe key competences, the test developers studied what characterised items assigned to different mastery levels as well as using theories underlying frameworks such as the CEFR.

Although the mastery levels started out as a means of characterising the statistical groupings of pupils' results, the aim of this study is to examine the degree to which the individual characterisations, i.e., the descriptors, can predict the statistical difficulty of an item. A close resemblance between raters' judgements, based on descriptors, and the distribution of results may strengthen the argument that the descriptors reflect the framework of the NTE (Hasselgreen, 2004).

The relationship between the level descriptors and the levels assigned after the tests are taken raises questions of the relationship between norm-referenced and criterion-

referenced testing. While the two usually have different methods of interpreting test scores (Burkett, 2018), and thus different means of item construction (Brown, 2014), the NTE combines elements of both. While one can argue that the tests are criterion-based assessments in that they “describe the performance of examinees in terms of the amount they know of a specific domain of knowledge or set of objectives” (Brown & Hudson, 2002, p.5; Sawaki, 2016), this is only half of the story. Items are indeed classified into mastery levels using criteria (the mastery level descriptors) and test developers’ knowledge and experience *a priori*, but they are re-classified using norm referencing by means of an IRT analysis afterwards. While criterion-referenced and norm-referenced testing are sometimes perceived as opposing practices that require a binary choice (Frisbie, 2005), they can be effectively combined and can complement one another (Brown, 1989; Lok et al., 2015), which is an aim of the NTE through the process of re-classifying results using norm-referencing. Sadler (2005) notes that, although criterion-referenced assessment has become desirable in recent years, its implementation has been highly inconsistent, making ‘good’ criterion-referenced assessment hard to characterise, especially in relation to norm-referenced assessment.

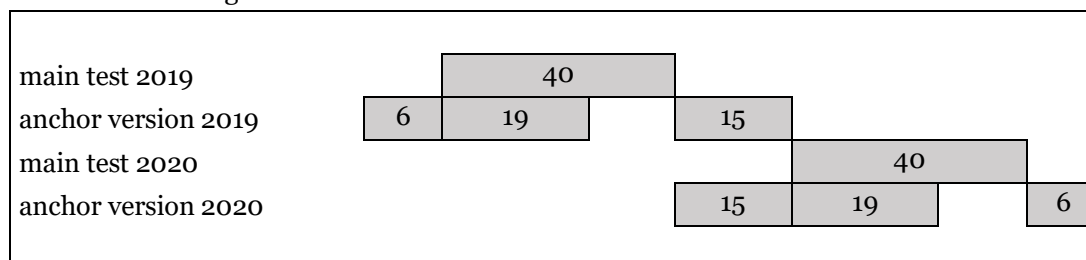
In the case of the NTE, the move from criterion-based assessment to norm-based assessment is followed by a move back to the criterion-based approach again. This is because, once the results are subjected to IRT analysis (Udir, 2022b), the year’s mastery level boundaries are re-calculated, and pupils’ results are assigned to a corresponding mastery level. These figures are distributed to schools and relevant stakeholders. They are accompanied by explanations of the levels, in the form of the descriptors (see Appendix A) and expanded upon in a teacher’s guide document distributed by Udir. Teachers and schools therefore have access to both numerical norm-referenced data, where they can compare performances to the rest of the country, and criterion-referenced data, where they can see what the figures mean for individual test-takers. This underlines the importance of a valid, defensible set of level descriptors, as providing schools with information into pupils’ ability in English receptive skills is the main intended consequence of the NTE (Grocott, 2022; Udir, 2022a).

## Methodology

### Overview of data collection

This study builds on two types of data: data from approximately 46,000 students taking the NTE in either 2019 or 2020, and data from 10 raters whose competencies include English teaching, language test development and researching language testing. The raters assigned descriptors from the mastery level scale to items. Since the NTE tries to measure trends, 21 anchor items are included each year for linking purposes, making it possible to analyse the 2019 and the 2020 items on the same scale. Each year, about 84% of the students take the main test, which consists of 40 items. The main tests of 2019 and 2020 have no items in common. The other 16% of the students take the so-called anchor version of the test. This version has (for the two years considered) 19 items in common with the main test of the current year, 15 items in common with each other, and 6 unique items; these unique items form part of the anchor with 2018 or 2021, but do not concern us here. The design is displayed graphically in Table 1. The numbers in the shaded bars represent the number of items.

**Table 1.** Test design for main and anchor tests



Each of the four data sets available for this study were randomly split in two, one part assigned to sample 1, the other to sample 2. All analyses are repeated independently for the two samples, allowing to account for the stability of the results. The samples both consist of 22,893 students. Each sample includes responses to 107 items. Most are dichotomous items, but some are partial credit items with a maximum score of three or four.

The raters did not assess the anchor items, only the 80 main test items from 2019 and 2020, 40 items from each year.

**Table 2.** Number of students in the two samples

Sample set	Description	Sample 1	Sample 2
1	NTE 2019 main test	9630	9631
2	Anchor version 2019	1810	1810
3	NTE 2020 main test	9633	9633
4	Anchor version 2020	1820	1819
Total		22893	22893

### Raters' assignment of items to mastery levels

Each rater assigned 80 items to mastery level descriptors from the scale developed for the NTE (Appendix A). The descriptors were divided into seven groups, based on their *content*, mirroring a mixture of linguistic and operational competences. For example, descriptors reflecting linking information in a text were grouped together, as were those describing understanding grammatical constructions. These groups are referred to here as boxes. The boxes contain the following categories:

- 1) Vocabulary – recognition of an individual word or phrase integral to completion of the task.
- 2) Sentences – understanding of a sentence or adjacent sentences integral to completion of the task.
- 3) Understanding of texts – whole-text level understanding, such as understanding the main point.
- 4) Connecting information – make one or multiple links of information from different parts of a text.
- 5) Grammar – understand and/or use a grammatical structure, such as verb tenses or passive voice.
- 6) Drawing conclusions – making a conclusion or inference, often from information not explicitly stated.
- 7) Finding information – Locating a key piece of information explicitly stated in the text.

Like the mastery level scale, the boxes contained descriptors that are assumed to progress in an ascending order of difficulty. Recognising vocabulary, for example, would ascend from “can understand some simple, concrete words and expressions” to “has a broad and nuanced vocabulary”. Some boxes did not contain descriptors for every mastery level. Some of the descriptors in the framework are ‘stand-alone’ descriptors, i.e., they do not belong to a sequence of related descriptors that form a

progression along a particular dimension. They were not included in the present study, as they can apply to all items, such as “can follow short, clear, and simple instructions” at M1. Table 2 shows the levels at which each box is represented; a shaded cell indicates that this level has no descriptor. In addition, not all boxes are relevant for all items, therefore a “not relevant” option was added. Each of the 10 raters assigned 80 items to a separate level on seven boxes, a total of 560 level assessments per rater, yielding 5,600 level assignments.

**Table 3.** Overview of boxes and mastery levels raters had to consider for each item

	Not relevant	M1	M2	M3	M4	M5
<b>Box 1: Vocabulary</b>						
<b>Box 2: Sentences</b>						
<b>Box 3: Understanding of texts</b>						
<b>Box 4: Connecting information</b>						
<b>Box 5: Grammar</b>						
<b>Box 6: Drawing conclusions</b>						
<b>Box 7: Finding information</b>						

Before starting, raters were given instructions to “select the level descriptors you believe are the *minimum* requirements for an eighth-grade pupil to successfully answer the item”.

### **Relationship between item difficulty and raters’ assignment of items to mastery levels**

The general assumption is that the raters’ assignments have some relationship with the difficulty of the items as derived from the answers to the items by a representative sample of the two populations (Grade 8 in 2019 and 2020). The relationship will be expressed as a correlation coefficient. Two aspects will be discussed in turn: the definition and estimation of the difficulty of the items on the one hand and the estimation of the correlation between the ratings and the difficulty.

## Item difficulty

Because the data were collected in an incomplete design, the difficulty index used in Classical Test Theory, the proportion of correct answers or p-value, is not appropriate. For the analysis of the item answers, a member of the IRT family of models was used: a restricted version of the two-parameter logistic model (2PLM), known as the one-parameter logistic model (OPLM) (Verhelst & Glas, 1995; Verhelst et al., 1993). In IRT models, the answer to the items is controlled in a probabilistic way by a unidimensional variable  $\theta$ , which is not directly observable, and is therefore called the latent variable. In the 2PLM, each item  $i$  is characterised by two parameters: a discrimination parameter  $a_i$  and a difficulty parameter  $\beta_i$ . For dichotomous items, we can define the difficulty of an item  $i$  as its difficulty parameter  $b_i$ .

If the value of the latent variable of a test taker equals  $\beta_i$ , then, according to the model, this test taker has a probability of 0.5 to give a correct answer, i.e., the expected value of the response variable  $X_i = 0.5$ , the average of 0 and 1 if both outcomes have the same probability. This expected value is also half of the maximum value of the score. For partial credit items, there is no difficulty parameter<sup>i</sup>. To find a reasonable difficulty value for these items, one can choose a value with the same interpretation as the meaning of the difficulty parameter for dichotomous items: it is the value of the latent variable for which the expected value of the score is just half of the maximum score<sup>ii</sup>.

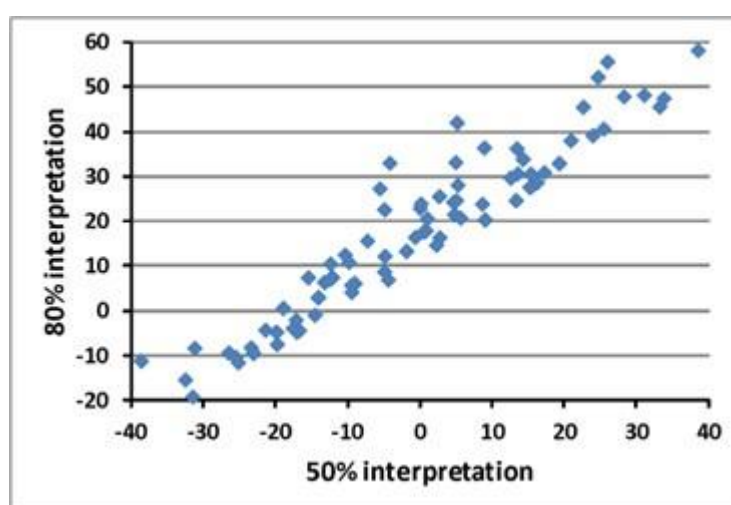
Given the instructions to the raters, however, one could object to the use of 50% of the maximum score as too lenient a criterion for mastery of an item and thus set a stricter requirement, where mastery means answering correctly in the clear majority of cases. Therefore, we have used two definitions of difficulty, one with a 50% interpretation and one with an 80% interpretation. For dichotomous items, the latter means that the value of the latent variable  $\theta$  must be chosen such that

$$P(X_i = 1 | \theta) = \frac{\exp[a_i(\theta - \beta_i)]}{1 + \exp[a_i(\theta - \beta_i)]} = 0.8$$

where  $a_i$  is the discrimination index of item  $i$  and  $\beta_i$  its difficulty parameter. The solution is

$$\theta = \frac{1}{a_i} \ln \frac{0.8}{1-0.8} + \beta_i = \frac{1.3863}{a_i} + \beta_i$$

For partial credit items, an iterative procedure is needed. Item parameters were estimated using the OPLM software package (Verhelst et al., 1993). To avoid too many decimal places, the difficulty values in the two definitions have been multiplied by 100. In Figure 1, the relation between the two definitions is displayed graphically for sample 1; the results for sample 2 are very similar. We see that the relation between the two choices for the difficulty is not linear. This is because the discrimination parameters are not constant across items. Therefore, it is a legitimate question to ask if the result (the correlation between ratings and difficulties) will vary, and to what extent, using different definitions of difficulty.



**Figure 1.** Relation between the difficulty values with two interpretations

### **The quantitative analysis with qualitative data**

To appreciate the justification for the analysis used, it is important to understand the status of the collected observations. The difficulty of the items, the dependent variable in a prediction, is clearly quantitative and, given the uncertainty about the exact nature of this variable, two different definitions have been selected. As seen in Figure 1, the relationship between these two quantifications is not linear, but the correlation (for the 80 items that are used) is high: 0.95. The main problem, however, concerns the status of the independent variables. The observations are not quantitative: they just are the assignment of one out of several descriptors to an item (and this is repeated across items, across raters, and across boxes). Thus, to apply a quantitative prediction, with a technique like regression analysis, one must convert

the observations into a quantitative variable – the quantification of qualitative observations (Coombs, 1964).

The general hypothesis, or rather hope, in the research is that the highest possible correlation between difficulty and quantifications will show that the values of the quantifications will be in increasing rank order within each box. The aim of the analysis, therefore, is twofold: (1) find the value of the highest possible correlation between the difficulty and the independent variables, properly quantified and (2) find the quantification of the qualitative variables that yields this correlation. A related, but important, problem is to find out whether the quantification is unique or if multiple solutions exist. The quantifications that lead to the highest correlation are called the optimal quantifications (Gifi, 1990; Young et al., 1976). The general procedure, known as an Alternating Least Squares procedure (ALS) can be described in the following steps:

1. Choose some arbitrary quantification of the independent variables.
2. Fix the quantifications to their most recently found values and find the regression coefficients.
3. Fix the regression coefficients to their most recently found values and find the best fitting quantifications.

If the largest change in one of the regression parameters or one of the quantifications is less than some predetermined criterion, stop; otherwise go to Step 2. The criterion, called the convergence criterion, is some small value such as 10<sup>-8</sup>.

### Quantification of level descriptors

If we knew the optimal quantifications to be given to the level descriptors, we could write a simple model for multiple regression:

$$y_{ri} = \alpha_r + \beta_1 X_{1ri} + \beta_2 X_{2ri} + \dots + \beta_7 X_{7ri} + \varepsilon_{ri}, \quad (1)$$

where  $r$  is the index for the rater and  $i$  the index for the item;  $y_{ri}$  is the difficulty value of the item, it holds that this value is the same for all raters. The quantity  $\alpha_r$  is the rater effect for rater  $r$  and is a regression parameter that must be estimated along with the regression parameters  $\beta_1$  to  $\beta_7$ , one for each of the seven boxes. The rater effects can be

understood as a harshness characteristic of the rater: a high value indicates that the rater tends to assign relatively low levels to the items (thus considering items relatively easy), which corresponds to a harsh attitude towards the students. The residual is indicated by  $\varepsilon_{ri}$ .

It is important to understand the meaning of the  $X$ -variables. For example, box 4 has five level descriptors, and we supposedly know the values of them. Then  $X_{5ri}$  means the value given to the descriptor that rater  $r$  has chosen for item  $i$ . If they have judged that box 5 is not relevant for this item, this implies that (according to this rater) the dimension represented by box 5 cannot contribute to the item's difficulty. Thus, whatever the value of  $\beta_5$ , the value of  $X_{5ri}$  must be zero. As this reasoning applies to all items and boxes, the value of 'not relevant' must be zero.

The computation of the only remaining unknowns in equation (1) are the regression coefficients  $\alpha$  and  $\beta$ . It amounts to finding the values of the regression parameters such that the sum (across raters and items) of the squared residuals is as small as possible, making the correlation between the difficulties and the predicted difficulties as high as possible. As we supposedly have the optimal quantifications, the problem is seemingly solved. But here it is worthwhile to find out whether the solution is unique, and we can see that it is not: for example, box 4 allows raters to choose one of five descriptors, or the non-applicable option. In Table 3, three different equivalent solutions are presented: the leftmost one is the one allegedly found (fictitious), and the two others can be found through a simple procedure: the 'new'  $\beta_4$ -value is given by dividing the 'found' value by a non-zero constant  $c$  (the bottom row of the table), and the 'new' quantifications are given by multiplying the 'found' values by the same constant  $c$ .

**Table 3.** Illustration of an infinite number of solutions

	'found values'	Equivalent values-1	Equivalent values-2
$\beta_4$	0.7	0.35	1.00
Not relevant	0.0	0.0	0.00
Level M1 <sup>a)</sup>	0.6	1.2	0.42
Level M2	1.0	2.0	0.70
Level M3	1.4	2.8	0.98
Level M4	1.6	3.2	1.12
Level M5	2.2	4.4	1.54
$c$		2	$\beta_4 (= 0.7)$

a) Note that the expressions 'Level M1', 'Level M2', etc., are short for 'the descriptor available for Level M1, M2, etc. in Box 4'

From the illustration in Table 3, we immediately see that the value of  $c$  is arbitrary, and moreover, it may differ from box to box: all the transformations shown in Table 3 show that the product of the regression coefficient and the quantification does not change, so that the value of the predicted difficulty<sup>iii</sup> does not change.

We chose the solution where all regression coefficients of the boxes (the  $\beta$ -parameters) equal one. Then, referring to equation (1), the prediction of the difficulty (according to rater  $r$ ) is just the *sum* of the values of the level descriptors selected by this rater plus the rater effect  $\alpha_r$ . Thus, in the ALS-procedure, the estimation of the regression coefficients  $\beta$  can be skipped as their value is fixed to one. The method to find the optimal quantifications and at the same time the rater effects is then simple: we define a dummy variable for the raters and a dummy variable for each descriptor. We have ten raters and 30 level descriptors, so in total we have 40 dummy variables and  $80 \times 10 = 800$  difficulties<sup>iii</sup>. We write the regression equation as

$$y_{ri} = \beta_1 X_{1r} + \beta_{10} X_{10,r} + \beta_{11} X_{11,ri} + \beta_{40} X_{40,ri} + \varepsilon_{ri} \quad (2)$$

The  $X$ -variables in (2) can only take two values, 0 or 1. The first ten indicate the rater, and the others refer to a level descriptor in an arbitrary but fixed order. For example, if the second level descriptor in the third box has number 25, then  $X_{25,ri} = 1$  if rater  $r$  has chosen this descriptor in box three when judging the requirements of item  $i$ , otherwise  $X_{25,ri} = 0$ . Note that there is no  $X$ -variable corresponding to a ‘not relevant’ answer in any of the seven boxes.

With a regression analysis, the 40  $\beta$ -parameters are easily calculated, and this finishes the analysis: the estimated  $\beta$ -parameters  $\beta_1$  to  $\beta_{10}$  are the rater effects and the parameters  $\beta_{11}$  to  $\beta_{40}$  are the optimal quantifications, i.e., the values that make the correlation between the difficulty and its prediction as large as possible. The proof is simple: the product of a  $\beta$ -parameter and an  $X$ -variable equals zero if  $X=0$  and equals the  $\beta$ -parameter if  $X=1$ . Thus, the predicted difficulty is the sum of the  $\beta$ -parameters for which the  $X$ -variables take the value 1.

For each item, the right-hand member of equation (2), ignoring the residual  $\varepsilon_{ri}$ , is the prediction of the difficulty of item  $i$  by rater  $r$ . As an interesting result, one can compute the correlation between these predictions and the actual difficulty for each rater

separately and then probably compute the average correlation across raters, but it should be noted that this correlation is attenuated by the unreliability of both the difficulty estimates of the items and of the quantified judgments of the raters, such that the average correlation somewhat blurs the question of the extent to which item difficulty is predictable. The unreliability of the difficulties is negligible given the huge number of observations in both samples, but the unreliability of the quantified judgments may be substantial. To get a view of this theoretical predictability we can simply average the individual predictions. This is discussed further in the Results section.

### **Quality of research and ethical considerations**

The present study was preceded by a pilot study to ensure clarity of the process and to ensure that the survey format functioned technically. The pilot involved three participants, all of whom were researchers or test developers (or both) with experience of the NTE and scoring frameworks. None of the participants were on the panel for the main study. The pilot study raised no issues with the format.

In terms of ethical considerations, the panel members are not identified in the reporting of the study. For the data samples, no personal information about the pupils whose scores were used was available to the researchers. Permission to use the data was granted by Udir. The study was confirmed as being compliant with data protection laws by *Rette*, a system for research compliance.

## **Results**

### **Inter-rater comparisons**

As stated in the methodology section, we have a prediction of the difficulty of the item for each item-rater combination. So we can answer two questions: (1) what is the relationship between the predictions in all pairs of raters; and (2) what is the relationship between the predictions of the raters and the (estimated) difficulty of the items?

The first question refers to inter-rater reliability. There are 45 pairs of raters in total. The 45 correlations were computed in the four cases: two interpretations of difficulty

and two samples. The results in all four cases were identical up to two decimal places. They are displayed in Table 4. The smallest correlations (less than 0.5) all occurred when rater number four was involved.

The correlation between the predictions per rater (across items) and the estimated difficulties has a more direct relation to the validity of the criterion referenced interpretation of the scores. The results are displayed in Table 5.

**Table 4.** Inter-rater correlations of predicted difficulties

	Difficulty (50%)	Difficulty 80%
minimum	0.41	0.45
maximum	0.78	0.79
average	0.63	0.65
#(less than 0.5)	4	2

**Table 5.** Correlation between predicted difficulty and difficulty parameters (per rater)

	r-1 <sup>a)</sup>	r-2	r-3	r-4	r-5	r-6	r-7	r-8	r-9	r-10	average
50%-s1 <sup>b)</sup>	0.71	0.64	0.65	0.51	0.66	0.73	0.72	0.79	0.72	0.72	0.685
50%-s2	0.71	0.65	0.65	0.51	0.66	0.73	0.72	0.79	0.72	0.73	0.687
80%-s1	0.74	0.65	0.66	0.57	0.68	0.74	0.72	0.78	0.75	0.76	0.705
80%-s2	0.74	0.65	0.65	0.57	0.68	0.74	0.72	0.78	0.75	0.76	0.704

a) r-1,...,r10: rater 1, etc.

b) 50%-80%: difficulty interpretation; s1, s2: sample 1, sample 2

The correlations in samples 1 and 2 are virtually identical. The correlations in the 80% cases tend to be a little higher than the ones in the 50% interpretation. The lowest correlations are found for rater 4, as with the inter-rater correlations.

The low correlations for rater 4 might be considered reason to exclude this rater's judgments from the data, but such decisions should be made carefully. The raters were chosen based on their familiarity with the NTE and the Norwegian education system. Thus, one must have good reasons to exclude a selected rater from the study solely on the ground of results which may not please the researchers. Therefore, it was decided to not exclude rater 4 from the sample.

In Table 6, the correlations between the average difficulty predictions (across raters) and the estimated difficulty are given, along with the average correlations (rightmost column of Table 5). The correlations with the average predictions are substantially higher than the average correlations (across raters), showing that the predictability of

the difficulties from the raters' judgments is better than suggested from the average of the ten individual correlations.

**Table 6.** Correlation of difficulties with average predictions

	Correlation with averages	Average correlation
50%-s1 <sup>a)</sup>	0.836	0.685
50%-s2	0.839	0.689
80%-s1	0.848	0.705
80%-s2	0.851	0.704

a) 50%-80%: difficulty interpretation; s1, s2: sample 1, sample 2

## Rater assignments

The data collected from the panel members was divided by the individual boxes, and the frequencies of each mastery level descriptor chosen, indicated by their corresponding mastery level, were recorded. These frequencies are displayed in Table 7. The shaded cells indicate mastery levels not represented in that individual box.

**Table 7.** Frequency of descriptors chosen by raters

	Not relevant	M1	M2	M3	M4	M5	Total
Box 1: Vocabulary	4	34	226	313	181	42	800
Box 2: Sentences	55	0	218	323	165	39	800
Box 3: Understanding of texts	347	0	121	206	112	14	800
Box 4: Connecting information	451	42	138	96	65	8	800
Box 5: Grammar	418	24	81	121	132	24	800
Box 6: Drawing conclusions	479	0	0	132	170	19	800
Box 7: Finding information	304	66	140	206	84	0	800
<b>Total</b>	2058	166	924	1397	909	146	5600

The table indicates that box 1, understanding vocabulary, (and to a lesser extent box 2, understanding sentences) was deemed relevant in almost all cases; it was only deemed not relevant four times out of 800. Aside from 'not relevant', the totals indicate that the descriptors associated with M3 had the highest frequency of being chosen by raters.

With reference to the original 10-20-40-20-10 distribution, the distribution of mastery levels chosen by raters through their associated descriptors by percentage is shown in Table 8:

**Table 8.** Distribution of raters' mastery level allocations by percentage

Mastery level	Percentage assigned by raters
1	4.7%
2	26.1%
3	39.4%
4	25.7%
5	4.1%

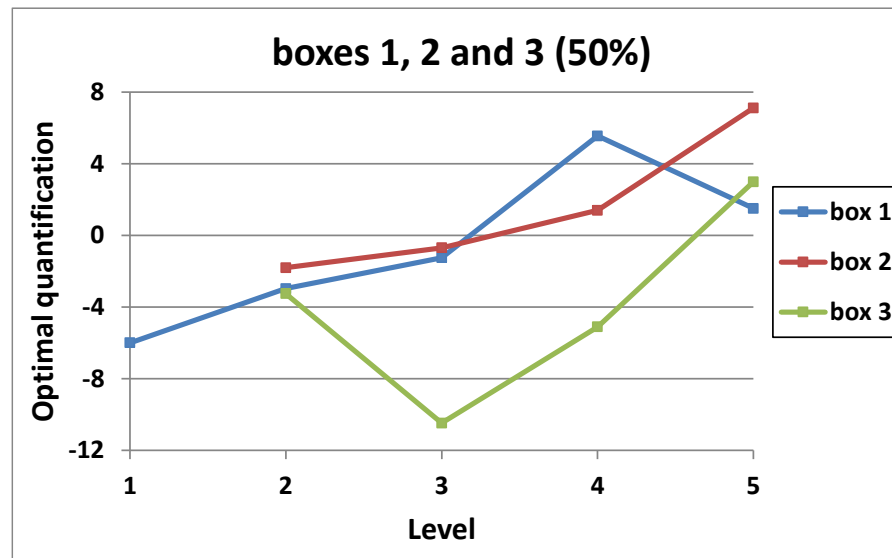
### **Relationship between item difficulty and raters' assignment of items to mastery levels**

Once the raters' assignments were compared to the item difficulty at the two interpretation levels, 50% and 80%, comparable results were obtained. For both interpretations of item difficulty, there is a strong correlation, indicating that the mastery level descriptors, and thus the levels to which they belong, appear to broadly be a reliable indicator of item difficulty. However, this does not account for individual differences and discrepancies at both inter- and intra-mastery level. Moreover, a lot of freedom was left to the analysis method: the hypothesis was that the quantifications would be monotonically increasing with the levels (the higher the allocated level, the more difficult the item). The results for sample 1, for both difficulty interpretations, are presented in Figures 2-5. The results for sample 2 are virtually identical to the ones for sample 1 and are not displayed.

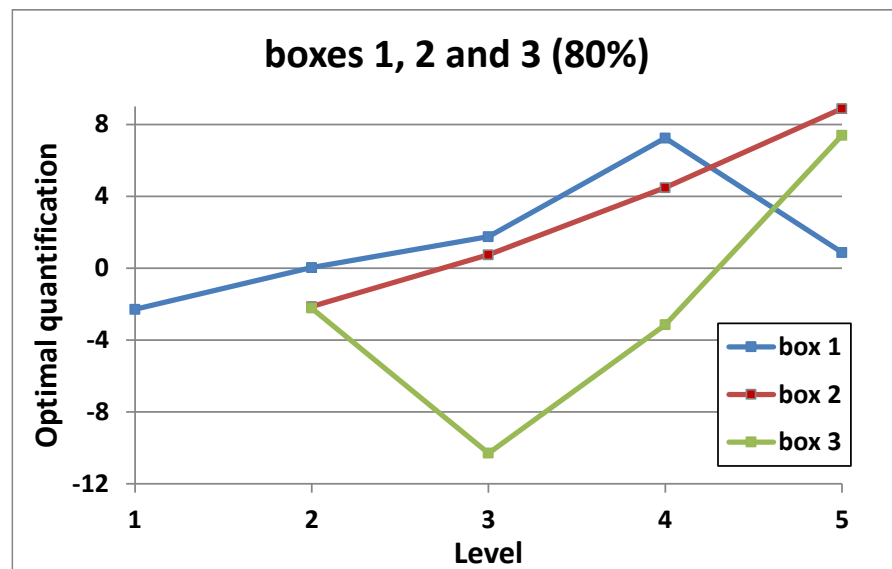
### **Boxes 1-3**

Figures 2 and 3 show the relationship between item difficulty and the raters' assignments of items to mastery levels regarding box 1 (vocabulary), box 2 (sentences) and box 3 (understanding of texts), Figure 2 for the 50% interpretation of difficulty and Figure 3 for the 80% interpretation. The levels without descriptors (see the greyed cells in Table 7) are not represented in the figures.

The first thing worth noticing is that the difficulty interpretations do not seem to significantly influence the result. This is because the curves representing the three boxes for both interpretations, shown in Figures 2 and 3, are very similar.



**Figure 2.** Boxes 1-3 (vocabulary, sentences, and understanding of texts) for sample 1 for the 50% interpretation

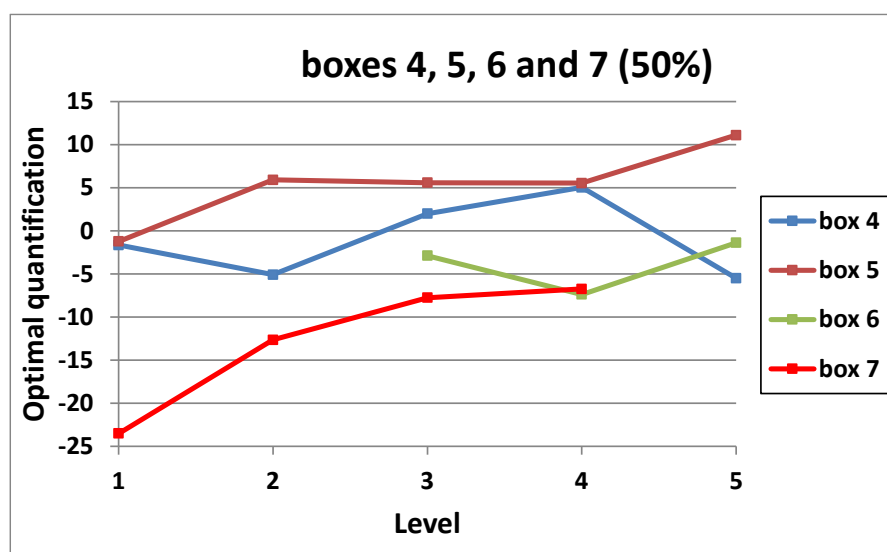


**Figure 3.** Boxes 1-3 (vocabulary, sentences, and understanding of texts) for sample 1 for the 80% interpretation

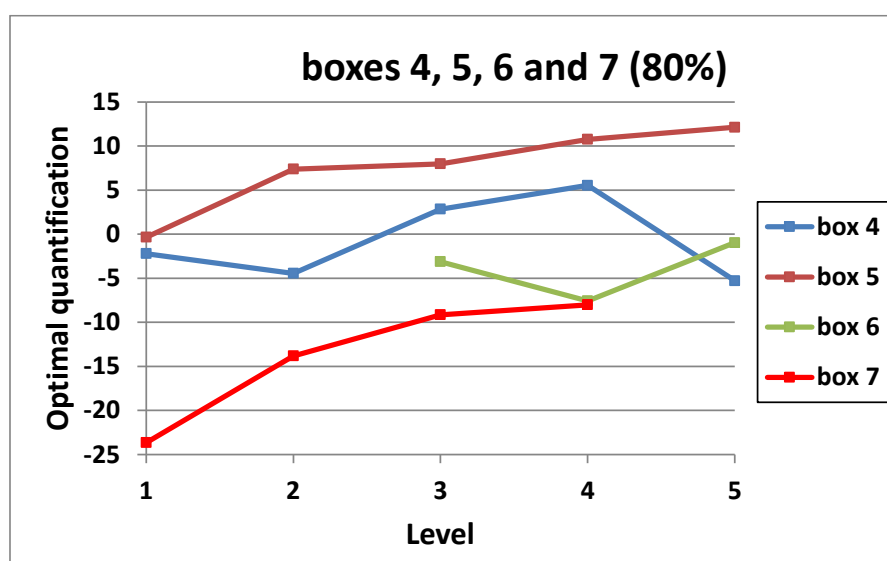
The most interesting point is that the curves indicating the relationship between item difficulty and raters' assignments behave differently for the three boxes. Box 2 (sentences) shows the rater assignments matching the expected ascending order of difficulty, meaning that the higher the level descriptors chosen, the higher the corresponding difficulty level based on pupil data. Box 1 (vocabulary) also displays this pattern until it comes to M5, where a decline is noticeable. This decline is slightly sharper for the 80% interpretation of difficulty, but the pattern is the same. Like box 2, box 3 (understanding of texts) does not contain a descriptor for M1, but the line begins with a decline from M2 to M3, before following the expected ascending pattern.

### Boxes 4-7

Figures 4 and 5 are almost identical, meaning that, like boxes 1-3, the choice of difficulty interpretation has little impact, except possibly for box 5 (grammar), where there is an increase in the optimal quantification across the whole level range in the 80% interpretation, while in the 50% there is slight decrease from M2 to M4. Again, we see differences between the behaviour of each box in terms of raters' assignments and item difficulty. Box 5 (grammar) and box 7 (finding information) both follow the expected order of ascending difficulty, although box 7 is not represented by any descriptors at M5.



**Figure 4.** Boxes 4-7 (connecting information, grammar, drawing conclusions, and finding information) for sample 1 for the 50% interpretation



**Figure 5.** Boxes 4-7 (connecting information, grammar, drawing conclusions, and finding information) for sample 1 for the 80% interpretation

Box 4 (connecting information) does not follow the expected pattern as there is a slight drop between M1 and M2, and a more severe drop between M4 and M5. However, the M5 descriptor was only chosen eight times in total, as per Table 3, so conclusions taken from this should be treated with caution. Box 6 (drawing conclusions) arguably demonstrates the most pronounced departure from the expected order, as it is only represented by levels M3 to M5, and there is no evidence of an ascending order. There is a drop from M3 to M4, and even M5, while higher than M4, is only minimally higher than M3, indicating that the raters' assignments of the levels do not show a monotone correspondence to the difficulties. There can be several reasons for this departure, which are explored in the Discussion section.

### Discrimination

An important question remains: can one give an indication of the relative importance of the boxes for the predictability of the item difficulty? One cannot use any reference to the magnitude of the regression coefficients, as they are arbitrary in the search for an optimal quantification. A useful indication could be the frequency of 'not relevant' choices per box, but this may rather reflect the choices made by the test construction team. Thus, we should rather ignore the 'not relevant' cases and look at the other cases, separately per box. A good candidate might be the 'steepness' of the lines in Figures 2 to 5. An example is the lines for boxes 5 and 7 in Figure 4: the line representing box 7 is steeper than the one for box 5. But this graphical information alone may be misleading, as one cannot infer anything from it with respect to the frequency of the choices. An index which combines the steepness and the frequencies is the standard deviation of the quantifications<sup>iv</sup>. They are given for the 50% difficulty interpretation in Table 9; the results for the 80% interpretation are very similar.

**Table 9.** Standard deviations of the quantifications – sample 1, 50% interpretation

	box 1	box 2	box 3	box 4	box 5	box 6	box 7
n	796	745	453	349	382	321	496
naïve <sup>a</sup>	5.275	1.823	0.783	0.385	3.898	0.792	1.489
optimal	3.423	2.055	3.629	4.046	2.205	2.367	5.388
monotonic	3.319	2.055	3.113	2.468	2.200	0.951	5.388

a) when descriptors in the boxes are given consecutive values, e.g., 1, 2, 3,....

The numbers in the row labelled 'n' are the frequencies (out of 800) where 'not relevant' is not given as an answer. The standard deviations in the naïve case are quite

different from the other two cases. Another reason to not use the naïve case is that three of the seven regression coefficients were negative. The SD in the monotonic case cannot be higher than in the optimal case; they are lower because at least two different values from the optimal case have been given the same value. They are the same in boxes 2 and 7 where the optimal quantifications are already monotonic with the levels. The most drastic change is in box 6, where two out of three quantifications (M<sub>3</sub> and M<sub>4</sub>) have been set equal to each other in the monotonic case. Another substantial drop occurred in box 4 where two pairs of values (at M<sub>1</sub> and M<sub>2</sub> and at M<sub>4</sub> and M<sub>5</sub>) had to be set equal. The three best discriminating cases are box 7, box 1 and box 3.

## Discussion

The first point apparent from the results is that they appear to support the rough distribution (Pižorn & Moe, 2012) of mastery levels intended by Udir (2022b), at least in terms of the frequency of choosing descriptors from the corresponding levels. This shows at least that raters consider the items to most frequently fall under the umbrella of M<sub>3</sub>, which the supporting documentation of the NTE suggests should be the case. One can further argue that this frequency distribution supports the argument for structural validity (Messick, 1996). This is because, if we take the raters' judgements as representative of the descriptors, their correspondence with the intended distribution indicates that the descriptors do indeed represent that which is required of test-takers and interact well with the scoring system. This would indicate consistency with the framework, which Messick (1989) describes as a key to structural validity. This is supported by the strong correlation between raters' judgements and difficulty level at both the 50% and 80% interpretations. This further supports the argument that the mastery level descriptors are reliable indicators of item difficulty.

The NTE descriptors seek to characterise the norm-based mastery levels for key stakeholders, primarily English teachers and pupils, and also test developers, who need to have the criteria in mind when developing items for specific levels. The validity evidence from the present study, specifically the strong correlation between the judgements and empirical difficulty, would appear to support the notion that the descriptors do indeed characterise the varying difficulties found across the mastery levels without major issue.

In addition to the overall correlation between the raters' judgements and item difficulty, the results also indicated differences between the individual boxes of descriptors. In some instances, these differences could indicate issues with individual descriptors, or sets of descriptors, which in turn could represent threats to structural validity. This is because there is a clearly intended order of difficulty the descriptors should follow, based on the levels they represent. To exemplify, if the raters had assigned descriptors which mostly fall under M5 to a group of items, yet those items had proved less difficult than a set of items to which the raters had mostly assigned descriptors from M4, we could reasonably argue that there is an issue with the descriptors at M4, M5, or both. This would not reflect the construct being measured and would thus weaken the validity argument (Brown & Bonsaksen, 2018; Rios & Wells, 2014).

As it transpired, there were no extreme, whole-level discrepancies, but some boxes certainly indicated issues in that they did not follow a monotonic correspondence to item difficulty. There can be several reasons for this, and to choose one of them requires further empirical research:

1. The wording of the descriptors may be unclear, meaning that the reference to an increasing competence (in the interpretation of the raters) is lost.
2. The general hypothesis that a higher level implies a more difficult item may be too strong; it may be possible that there are items which require a high processing level, but which are nevertheless relatively easy.
3. A statistical reason: the discriminatory power of the descriptors between adjacent levels may be such that an inversion of the quantification relative to the levels can easily occur due to sampling error.

Box 1 (vocabulary), as an example, followed an ascending order until it came to M5, at which point the graphs indicate a fall. The M5 descriptor describes a pupil at this level as having a "broad and nuanced vocabulary", whereas M4 describes a pupil as having a "quite broad vocabulary". This clearly demonstrates a difference in ability with regards to this aspect of reading, so it would not immediately suggest that the M5 descriptor should behave the way it does in the results. However, the answer may lie in the instructions given to the raters: to select the descriptors they feel describe the *minimum* requirements to answer an item correctly. Based on the details in the

accompanying documentation, recognition of vocabulary is not necessarily a skill associated with M5, but with lower levels. Even the items which require the identification of a word often test the ability to obtain the meaning of a word from the context, rather than pure vocabulary recognition – there is no real need for the vocabulary to be nuanced.

Another example of the results not demonstrating an ascending order of perceived difficulty is box 4 (connecting information). Here, the results also show a sharp drop at M5 level. The difference in the descriptors for M4 and M5 merely lies in the complexity of the text in which information is being connected; M4 describes connecting information from a “quite complex text” whereas M5 describes a “complex text”. The rater’s own interpretations of complexity may contribute here, raising questions of rater reliability (Deygers & Van Gorp, 2015; Knoch & Chapelle, 2017). The difference between “quite complex” and “complex” is not clearly defined, much like differences between adjacent levels in some of the scales of the CEFR (Alderson et al., 2004; Figueras, 2012; Fulcher, 2003), so one cannot definitively say how the raters defined complexity. For example, complexity can be viewed in relation to other items in the study or in the wider context of texts one could expect an eighth-grade pupil to read outside of the NTE. Another explanation may be found in the fact that, by design, there are fewer items intended to be at M5 in the NTE, as M1 and M5 are at the two ends of the distribution curve. The data for the M5 descriptor are therefore limited, thus accentuating discrepancies.

The final box which did not display the expected ascending order of item difficulty was box 6 (drawing conclusions), demonstrated in Figures 5 and 6. There may be an explanation, albeit a speculative one, for this departure from the expected order. The M3 descriptor says that a candidate “can interpret/have an overall understanding of a text and find the answer even if it is not explicitly stated in the text”, whereas M4 describes a candidate who “can interpret/have an overall understanding of a text and draw conclusions”. One can argue that the M3 descriptor sounds like a more advanced ability than that at M4, as it describes the answer being arrived at even if it is not explicitly in the text, whereas the M4 descriptor merely describes drawing a conclusion – this could be a conclusion based on something explicitly stated in the text. The M5 descriptor, which describes drawing “advanced conclusions”, differentiates itself from the M4 descriptor with the use of the adjective “advanced”. This of course raises the

same issue as descriptors in other boxes, that it is not easy to define what constitutes an advanced conclusion. Indeed, the CEFR has been criticised on the same grounds (Alderson et al., 2004; Figueras, 2012; Fulcher, 2003). Still, even the M5 descriptor does not mention these advanced conclusions being made even if the information is not explicitly in the text, which may explain why it was considered by the panel to only represent a slightly higher level of difficulty than the M3 descriptor.

The lack of the intended ascending order evident in box 6 may indicate a need for the descriptors concerning drawing conclusions to be revised. Revised descriptors could better reflect the framework of the NTE and the construct being measured, thus contributing to the strength of the validity argument (Brown & Bonsaksen, 2018; Rios & Wells, 2014). Boxes 1 and 4 do not seem to display such a severe departure from the expected pattern, but their drop-offs between M4 and M5, while possibly explainable by a relative lack of data, may nevertheless require attention.

While some of the discrepancies in individual boxes may be remedied by revisions of descriptors, the main focus of the present study is the structural validity of the NTE and the strength of the argument for it (Kane, 2006). The strongest contribution to the validity argument is the high degree of correlation between the descriptors chosen by raters and item difficulty based on national data. A correlation of around 0.84 suggests that the descriptors offer an accurate picture of the difficulty of items, and thus of what pupils can do at the specific mastery level. Given that these are the descriptors fed back to teachers and schools, the intended consequences of the tests, based on the documentation produced by Udir, may be more likely to come to fruition. This is because accurate information about pupils' ability can directly contribute to a formative assessment process and quality development in schools (Udir, 2022a). The descriptors offering an accurate picture of what pupils can do and, crucially, what they need to work on also supports the argument that the moves between criterion-based and norm-based scoring are defensible and thus support the argument that the scoring system is a strong reflection of the framework (Messick, 1996).

The results in the present study must be viewed with a degree of caution, despite the strong correlations that were evident. This is because the study applied an exploratory, experimental method. While it resonates well with Kaftandjieva's (2004) call to use multiple methods to ensure the defensibility of inferences and links made from a

scoring system, it must be viewed *alongside* studies such as standard setting events. It is therefore important for the NTE, and other comparable assessments, that such studies are not one-time events, but rather part of a continuous improvement and validation effort. This would also contribute to the quality development aspect of the main purpose of the NTE (Udir, 2022a), thus supporting the inferences made on the back of the test scores and, ultimately, the validity argument (Messick, 1989).

### **Limitations of study**

Given their similarities, the present study shares several limitations with standard setting procedures in general. For example, the wording of the instructions given to participants is of paramount importance (Moe & Verhelst, 2017), to avoid any ambiguity in what is expected of them. The issue of ambiguity could have been exacerbated in that the mastery level descriptors are, like those in the CEFR, a broad description of what a language learner can do at that stage of their development. This meant that a degree of subjectivity was inevitable when judging, for example, text complexity and length. This was explained to raters beforehand, but it may still have been the case that these judgements were made relative to other texts and items in the tests, as opposed to a general assessment of a reading level. It is arguably unclear which approach is more appropriate, given that the descriptors were created with the NTE in mind, but their ultimate purpose is as a referential tool to contextualise the level pupils are at, with the NTE results acting as a form of supporting evidence. Nevertheless, potential ambiguity was mentioned during the piloting stage, meaning it was necessary to ensure participants were fully aware of the arguable dual roles played by the level descriptors.

It is important to stress that the present study does not provide a permanent solution to the challenge of matching criterion-based descriptors to empirical item difficulty, since it is only based on items from two consecutive years. If the study were to take place using data from two or more other years, the results would likely be different.

The present study applied an experimental method, which, much like other validity studies, can only offer a picture of the strength and plausibility of the validity argument (Kane, 2006). It is not intended to offer a ‘final answer’ as to whether the NTE and its level descriptors are valid; validity is not considered a binary concept (Green, 2014).

With the study being exploratory, there is a necessary degree of speculation, as seen in the discussion, to support assessments of the strength of the validity argument.

## Conclusion

The present study involved 5,600 individual judgements being compared to real test data from around 46,000 pupils. One can then argue that the conclusions made are well-grounded, despite the limitations of the study's experimental design. The results indicate that the panel reached a distribution of mastery levels in their assessments of the items comparable to that intended by Udir, suggesting that the descriptors are an accurate reflection of expected mastery levels. This was one of two strong factors supporting the argument for structural validity. The second factor was the strong correlation between the raters' judgements and the *actual* difficulty of the items based on the test data. It seems reasonable to say that these two factors are enough to argue that the mastery level descriptors, and thus the scoring system, are an accurate reflection of the framework of the NTE and therefore provide valid information to teachers and schools – one of the intended purposes of the NTE (Grocott, 2022; Udir, 2022a).

Despite the findings that the descriptors are, in general, a good indication of item difficulty, some of the individual descriptor categories raised issues. The upper levels of box 1 (vocabulary) and box 4 (connecting information) did not show that the M5 descriptors were indicative of increased item difficulty, meaning that they may need revision. The fact that Box 6 (conclusions) deviated more severely from the expected order suggests that all the descriptors for this category are in more pressing need of revision. If a similar study were to take place after such revisions, and demonstrate a closer resemblance to the intended order, the validity argument would be strengthened further. We argue that, along with standard setting, such studies should be part of a continuous validation process, working to strengthen the validity arguments for the NTE.

## Notes

- i. For a partial credit item with a maximum score of  $m$ , there are  $m$  location parameters to be estimated.

- ii. This value can be computed from the values of the location parameters, but a simple closed form formula does not exist for it; one needs an iterative procedure which has been implemented in a short computer program.
- iii. The predicted difficulty of the item is given by the right-hand side of Equation 1, ignoring the residual  $\varepsilon_{ri}$ .
- iv. In the naïve case, the quantifications have to be understood as the product of the regression coefficients and the fixed values of the  $X$ -variables (1-5), so that in this case the predicted difficulties are just the sum of seven quantifications (see Equation 1).


### Author disclosures


The authors reported no conflict of interest in conducting the study.

The research formed part of the first author's PhD and did not involve any external funding.

The authors had the following roles respectively in conducting the research and writing the article: Craig Grocott – conceptualization, planning, data collection and recording, writing original draft, revising, proofreading; Eli Moe – planning, recruitment of panel members, writing original draft, revising, proofreading; Norman Verhelst – data analysis and visual presentation, writing original draft (data analysis and results), revising.

### ORCID iDs

Craig Grocott  <https://orcid.org/0000-0003-4482-926X>

Eli Moe  <https://orcid.org/0000-0001-6774-0706>

Norman Verhelst  <https://orcid.org/0009-0007-6537-8257>

### References

Alderson, J.C. (2007). The CEFR and the need for more research. *The Modern Language Journal*, 91(4), 659-663. [https://doi.org/10.1111/j.1540-4781.2007.00627\\_4.x](https://doi.org/10.1111/j.1540-4781.2007.00627_4.x)

- Alderson, J. C., Figueras, N., Kuijper, H., Nold, G., Takala, S., & Tardieu, C. (2004). *The development of specifications for item development and classification within The Common European Framework of Reference for Languages: Learning, Teaching, Assessment: Reading and Listening: Final report of The Dutch CEF Construct Project*. <https://eprints.lancs.ac.uk/id/eprint/44/>
- American Education Research Association, American Psychological Association, & the National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. American Psychological Association.
- Angoff, W.H. (1971). Scales, norms and equivalent scores. In R.L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 508-600). American Council on Education.
- Bachman, L.F., & Palmer, A.S. (1996). *Language testing in practice: Designing and developing useful language tests*. Oxford University Press.
- Bachman, L.F., & Palmer, A.S. (2010). *Language assessment in practice*. Oxford University Press.
- Brown, J.D. (1989). Improving ESL placement tests using two perspectives. *TESOL Quarterly*, 23(1), 65-83. <https://doi.org/10.2307/3587508>
- Brown, J.D. (2014, May 23). *Differences in how norm-referenced and criterion-referenced tests are developed and validated?* [Conference paper]. Kuroshio Seminar, Tokyo. <https://hosted.jalt.org/sites/jalt.org.teval/files/18-1-29%20Brown%20Statistics%20Corner.pdf>
- Brown, J.D. & Hudson, T. (2002). *Criterion-referenced language testing*. Cambridge University Press. <https://doi.org/10.1017/CBO9781139524803>
- Brown, T. & Bonsaksen, T. (2018). An examination of the structural validity of the Physical Self-Description Questionnaire-Short Form (PSDQ-S) using the Rasch measurement model. *Cogent Education*, 6(1). <https://doi.org/10.1080/2331186X.2019.1571146>
- Burkett, T. (2018). Norm-referenced testing and criterion-referenced testing. In J.I. Lontas & M. DelliCarpini (Eds.), *The TESOL encyclopedia of English language teaching*. Wiley. <https://doi.org/10.1002/9781118784235.eelt0351>
- Carlsen, C.H. (2014, April 9-10). *How valid is the CEFR as a construct for language tests?* [Conference presentation]. 5<sup>th</sup> ALTE Conference, Paris.

- <http://events.cambridgeenglish.org/alte-2014/docs/presentations/alte2014-cecile-hamnes-carlsen.pdf>
- Coombs, C.H. (1964). *A theory of data*. Wiley.
- Council of Europe. (2020). *Common European Framework of Reference for Languages: Learning, teaching, assessment. Companion volume*. Council of Europe Publishing. <https://rm.coe.int/common-european-framework-of-reference-for-languages-learning-teaching/16809ea0d4>
- De Jong, J.H.A.L., & Zheng, Y. (2016). Linking to the CEFR: Validation using a priori and a posteriori evidence. In J. Banerjee & D. Tsagari (Eds.), *Contemporary second language assessment* (pp. 83-100). Bloomsbury. <https://doi.org/10.5040/9781474295055.ch-004>
- Deygers, B., & Van Gorp, K. (2015). Determining the scoring validity of a co-constructed CEFR-based rating scale. *Language Testing*, 32(4), 521-541. <https://doi.org/10.1177/0265532215575626>
- Figueras, N. (2012). The impact of the CEFR. *ELT Journal*, 66(4), 477-485. <https://doi.org/10.1093/elt/ccs037>
- Frisbie, D.A. (2005). Measurement 101: Some fundamentals revisited. *Educational Measurement: Issues and Practice*, 24(3), 21-28. <https://doi.org/10.1111/j.1745-3992.2005.00016.x>
- Fulcher, G. (2003). *Testing second language speaking*. Longman/Pearson.
- Fulcher, G. (2016). Standards and frameworks. In D. Tsagari & J. Banerjee (Eds.), *Handbook of second language assessment* (pp. 29-44). De Gruyter Inc. <https://doi.org/10.1515/9781614513827-005>
- Gifi, A. (1990). *Nonlinear multivariate analysis*. Wiley.
- Green, A. (2014). *Exploring language assessment and testing: Language in action*. Routledge. <https://doi.org/10.4324/9781315889627>
- Green, A. (2018). Linking tests of English for academic purposes to the CEFR: The score user's perspective. *Language Assessment Quarterly*, 15(1), 59-74. <https://doi.org/10.1080/15434303.2017.1350685>
- Grocott, C. (2022). Intended or unintended consequences? Assessing the ways in which the National Tests of English are implemented and acted on in

- Norwegian schools. *Nordic Journal of Language Teaching and Learning* 10(1). <https://doi.org/10.46364/njltl.v10i1.1017>
- Harsch, C. (2019). What it means to be at a CEFR level – or why my mojito is not your mojito – on the significance of sharing mojito recipes. In A. Huhta, G. Erickson, & N. Figueras (Eds.), *Developments in language education: A memorial volume in honour of Sauli Takala*. European Association for Language Testing and Assessment (EALTA) and University of Jyväskylä. <https://jyx.jyu.fi/handle/123456789/65608>
- Hasselgreen, A. (2004). *Testing the spoken language of young Norwegians*. Cambridge University Press.
- Kaftandjieva, F. (2004). Section B: Standard setting. *Council of Europe. Reference Supplement to the Preliminary Pilot Version of the Manual for Relating Language Examinations to the CEFR for Languages: Learning, Teaching, Assessment*. Language Policy Division.
- Kaftandjieva, F. & Takala, S. (2002). *Relating the Finnish matriculation examination in English test results to the CEF Scales* [Conference presentation]. Helsinki Seminar on Linking Language Examinations to Common European Framework of Reference for Languages: Learning, Teaching, Assessment. <https://rm.coe.int/16806a6d1a>
- Kane, M.T. (2006). Validation. In R. Brennan. (Ed.), *Educational measurement* (4th ed., pp. 17–64). American Council on Education and Praeger.
- Kane, M.T. (2012). Validating score interpretations and uses. *Language Testing*, 29(1), 3-17. <https://doi.org/10.1177/0265532211417210>
- Kane, M.T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, 50(1), 1-73. <https://doi.org/10.1111/jedm.12000>
- Knoch, U. & Chapelle, C.A. (2017). Validation of rating processes within an argument-based framework. *Language Testing*, 35(4). <https://doi.org/10.1177/0265532217710049>
- Lok, B., McNaught, C., & Young, K. (2015). Criterion-referenced and norm-referenced assessments: Compatibility and complementarity. *Assessment & Evaluation in Higher Education*, 41(3) 450-465. <https://doi.org/10.1080/02602938.2015.1022136>

- Messick, S. (1989). Validity. In R.L. Linn (Ed.), *Educational Measurement* (3rd ed., pp. 13-103). Macmillan.
- Messick, S. (1996). Validity and washback in language testing. *Language Testing*, 13(3), 241-256 <https://doi.org/10.1177/026553229601300302>
- Milanovic, M. & Weir, C. (2010). Series editors' note. In W. Martyniuk (Ed.) *Aligning tests with the CEFR: Reflections on using the Council of Europe's draft manual* (pp. 8-20). Cambridge University Press.
- Moe, E. (2008). Juggling numbers and opinions: An attempt to set CEFR standards in Norway for a test of reading in English. In L. Taylor & C. Weir (Eds.), *Multilingualism and assessment* (pp. 67-79). *Studies in Language Testing 2*. Cambridge University Press.
- Moe, E., & Verhelst, N. (2017). Setting standards for multistage tests of Norwegian for adult immigrants. In S. Blömeke & J.E. Gustafsson (Eds.), *Standard setting in education* (pp. 181-204). Springer. [https://doi.org/10.1007/978-3-319-50856-6\\_11](https://doi.org/10.1007/978-3-319-50856-6_11)
- Papp, S. (2018). Criterion-related validity of tests of English for young learners. In S. Papp & S. Rixon (Eds.), *Examining young learners: Research and practice in assessing the English of school-age learners* (pp. 510-546). Cambridge University Press.
- Pižorn, K. & Moe, E. (2012). A validation study of the national assessment instruments for young English language learners in Norway and Slovenia. *Center for Educational Policy Studies Journal*, 2, 75-96. <https://doi.org/10.26529/cepsj.348>
- Rios, J., & Wells, C. (2014). Validity evidence based on internal structure. *Psicothema* 26 (1), 108-116. <https://doi.org/10.7334/psicothema2013.260>
- Sadler, D.R. (2005). Interpretations of criteria-based assessment and grading in higher education. *Assessment & Evaluation in Higher Education*, 30 (2), 175-194. <https://doi.org/10.1080/0260293042000264262>
- Sawaki, Y. (2016). Norm-referenced vs. criterion-referenced approach to assessment. In D. Tsagari & J. Banerjee (Eds.), *Handbook of second language assessment* (pp. 45-60). De Gruyter Inc. <https://doi.org/10.1515/9781614513827-006>

- Sibbern, M. (2013). *The National Test in English: Why it is important and why it is not enough*. [Master's thesis]. University of Oslo.
- Udir [Utdanningsdirektoratet]. (2022a). *Hva er nasjonale prøver?* [What are National Tests?] Norwegian Directorate for Education and Training. <https://www.udir.no/eksamen-og-prover/prover/rammeverk-for-nasjonale-prover2/hva-er-nasjonale-prover/>
- Udir [Utdanningsdirektoratet]. (2022b). *Rammeverk for nasjonale prøver* [Framework for National Tests]. Norwegian Directorate for Education and Training. <https://www.udir.no/eksamen-og-prover/prover/rammeverk-for-nasjonale-prover2/>
- Verhelst, N.D., & Glas, C.A.W. (1995). The one parameter logistic model. In G.H. Fischer & J.W. Molenaar (Eds.), *Rasch models: Their foundations, recent developments and applications* (pp. 215-238). Springer.
- Verhelst, N.D., Glas, C.A.W., & Verstralen, H.H.F.M. (1993). *OPLM: One parameter logistic model. Computer program and manual*. Cito.
- Weir, C.J. (2005). *Language testing and validation*. Palgrave Macmillan. <https://doi.org/10.1057/9780230514577>
- Wolf, M.C., Muijselaar, M.M.L., Boonstra, A.M., & De Bree, E.H. (2018). The relationship between reading and listening comprehension: shared and modality-specific components. *Reading and Writing* 32, 1747-1767. <https://doi.org/10.1007/s11145-018-9924-8>
- Young, F.W., de Leeuw, J., & Takane, Y. (1976). Regression with qualitative and quantitative variables: An alternating least squares method with optimal scaling features. *Psychometrika*, 41(4), 505–529. <https://doi.org/10.1007/BF02296972>

## Appendix A

N.B. The descriptors below are translations from the original Norwegian versions, which can be found at <https://www.udir.no/eksamen-og-prover/prover/nasjonale-prover/mestringsbeskrivelser-og-hva-provene-maler/engelsk-mestringsbeskrivelse/#a127083>

Given that some of the level descriptors involve degrees and approximations, the use of adverbs is necessary, and some of the translations are as close as possible without being direct translations.

### Mastery level 1

Pupils:

can understand some concrete and simple words and expressions

can follow short, clear and simple instructions

can find/recognise familiar and concrete words in a simple text

can find/recognise simple information in pictures and simple texts

can connect familiar and concrete words in the same area, for example homework and school

can recognise some familiar grammatical constructions in a context

### Mastery level 2

Pupils:

can understand some everyday words and expressions

can understand simple sentences

can find/recognise information in pictures and simple texts

can find/recognise specific details in a longer text

can understand the main content of a simple text

can find/recognise simple information in shorter, adapted texts, even when there is competing information

can connect information from different parts of a simple text

can place information from a simple text in the correct order

can recognise and use some simple forms of verbs and function words in a context

### **Mastery level 3**

Pupils:

can understand some less normal words and expressions

can find/recognise information in some complex sentences

can find/recognise information in adapted texts of varying lengths, even when there is competing information

can understand the main content of an adapted text

can connect information from different parts of adapted texts of varying lengths

can use the context to understand difficult parts of a text

can interpret/have a broad understanding of a text and find the answer even if it is not explicitly stated in the text

can place information from a text in the correct order

can recognise and use basic grammatical structures and function words in a context

### **Mastery level 4**

Pupils:

can understand what unfamiliar words mean from the context

have a quite broad vocabulary

can understand quite complex sentences

can find/recognise information in quite long and complex texts

can understand long and in parts complex texts

can connect information from different parts of a quite complex text

can interpret/have a broad understanding of a text and draw conclusions

can recognise and use normal grammatical structures and function words in a context

### **Mastery level 5**

Pupils:

can use reading and listening strategies which are appropriate for the purpose

have a broad and nuanced vocabulary

can understand complex sentences

can understand long and complex texts

can connect information from different parts of a complex text

can interpret/have a broad understanding of a text and draw advanced conclusions

can recognise and use quite advanced grammatical structures and function words in a context