Routledge
Taylor & Francis Group

Language Assessment Quarterly

ISSN: 1543-4303 (Print) 1543-4311 (Online) Journal homepage: http://www.tandfonline.com/loi/hlaq20

The B2 Level and the Dream of a Common Standard

Bart Deygers, Koen Van Gorp & Thomas Demeester

To cite this article: Bart Deygers, Koen Van Gorp & Thomas Demeester (2018): The B2 Level and the Dream of a Common Standard, Language Assessment Quarterly, DOI: 10.1080/15434303.2017.1421955

To link to this article: https://doi.org/10.1080/15434303.2017.1421955

Published online: 16 Jan 2018.

LANGUAGE ASSESSMENT QUARTERLY
https://doi.org/10.1080/15434303.2017.1421955

The B2 Level and the Dream of a Common Standard

Bart Deygers[a], Koen Van Gorp[b], and Thomas Demeester[c]

[a]KULeuven, CTO, Leuven, Belgium; [b]Michigan State University, East Lansing, Michigan, USA; [c]Ghent University, iMinds, Ghent, Belgium

ABSTRACT
In Flanders, Belgium, university admission of undergraduate international L2 students requires a certificate of an accredited test of Dutch. The two main university entrance tests used for certification share highly comparable oral components and CEFR-based oral rating criteria. This article discusses to what extent ratings on the oral components of these tests can be compared. The data used are the ratings of the oral performances of the same 82 candidates on both oral test components, which were administered within the same week. The correlation on the overall scores is high, but lower on the oral test component. Further analyses, including linear regression and multifaceted Rasch analysis, indicate that the B2 level was interpreted differently in the two tests. The results show that using the same language proficiency scales as the basis for rating scale criteria may lead to superficial correspondences or a perceived equivalence but does not necessarily lead to greater comparability of shared criteria. The findings of this study are especially useful for contexts in which different tests use similar criteria that are based on the same descriptors, and comparability is only assumed.

Introduction

Prior to 1800, when Henry Maudslay developed the first standardized screw thread, nuts and bolts were not easily interchangeable. Maudslay introduced a common standard and changed the life of every plumber to this day. Standards help to achieve transparency, uniformity, and interchangeability—at least in hardware. Language performance, in all its idiosyncratic and contextual variation, does not easily permit such standardization; nevertheless, fair and valid testing hinges on score comparability and score transparency (Kane, 2013). The first concept implies that scores on tests that target the same audience and share the same purpose can be meaningfully compared. The second concept—score transparency—entails that test scores have meaning to users. Candidates and admission officers need to be able to meaningfully interpret test scores (Alderson, 1991), and raters need to have the same conception of the same level. Still, in language testing standards are constant in one sense: they vary.

Currently, numerous language performance standards exist alongside each other. The ACTFL standards resulted from national efforts (ACTFL, 2012), the STANAG 6001 standards are used within supranational organizations (NATO, 2014), and others have been developed by testing organizations in the form of rating scales. Because different organizations use different scales, it is not easy for test takers or test users to interpret scores and compare them with other tests (Gomez, Noah, Schedl, Wright, & Yolkut, 2007). In Messick's (1989) approach to validity, which considers score use essential to a validity argument, a lack of score transparency presents a problem, which could potentially be resolved if all tests were linked to the same universally accepted levels and standards of performance. In Europe and beyond, the CEFR (Council or Europe, 2001) is often

CONTACT Bart Deygers ✉ bart.deygers@arts.kuleuven.be KULeuven, CTO, Blijde Inkomststraat 7, Leuven, 3000 Belgium.

considered such a standard (Deygers, Zeidler, Vilcu, & Carlsen, 2017). The CEFR's uptake has been widespread, but it has not been empirically validated in every intended or unintended context of use.

This article examines the use and usefulness of the CEFR in the context of rating scale design. More specifically, it explores to what extent the same CEFR descriptors have been similarly operationalized and interpreted in the oral components of two university entrance language tests. The research was conducted in Flanders, Belgium, but the implications are relevant beyond that context.

Literature review

Since its publication in 2001, the CEFR (Council of Europe, 2001) has become widely used and adopted by test developers, policy makers, teachers, publishers, and candidates alike. It has come to be seen as a common currency in language performance levels (Figueras, 2012), and in Europe it is now the leading framework in language testing (Figueras, 2012; Little, 2007; Papageorgiou, Xi, Morgan, & So, 2015). The CEFR is so influential that it has become necessary for tests to link to it to gain recognition within Europe (Deygers, Zeidler, Vilcu, & Carlsen, 2017; Fulcher, 2004). Outside of Europe too, many scoring systems and performance standards have been mapped onto the CEFR (e.g., Bärenfänger & Tschirner, 2012; Baztán, 2008; Tschirner, Bärenfänger, & Wisniewski, 2015; for ACTFL; Tannenbaum & Wylie, 2008; for TOEFL iBT; Swender, 2010; for STANAG 6001; Zheng & De Jong, 2011; for PTE Academic; also see Green, this issue). Theoretically, a framework that has received such a wide uptake by all parties involved could address the score transparency and comparability concerns Kane (2013) and Alderson (1991) warned about. In practice, however, there are issues.

Even though the goals of the CEFR in its current form are descriptive, not normative (North, 2014a), achieving score comparability across tests was one of the primary goals of its earliest drafts (van Ek, 1975, p. 8). Today too, in many European contexts, the CEFR level descriptors are used in a normative way, as performance standards, or as labels to facilitate score transparency (Fulcher, 2012; O'Sullivan & Weir, 2011; Roever & McNamara, 2006). With score transparency in mind, many test developers are using CEFR descriptors as the basis for rating scale development, but even though treating the CEFR as a heuristic is common practice (North, 2014a, 2014b; Weir, 2005b), it is not unproblematic. First, the CEFR offers guidance on essential test development matters, such as test purpose, response format, time constraints, and topic (Weir, 2005b). Two tests could have the same CEFR level, but very different specifications, and it would be wrong to consider them equivalent simply because they share a CEFR label (Green, this issue; Taylor, 2004). Second, because the CEFR is context and language-independent, test developers need to add specific details to the descriptors when using it in a rating context (Harsch & Martin, 2012). This necessary step may cause two tests to deviate in their interpretation of the CEFR levels, blurring the comparability as a result. In fact, CEFR descriptors have been criticized for their vagueness and inconsistencies, both within and across levels (Alderson, 2007; Harsch & Rupp, 2011; Papageorgiou, 2010) and may suffer from "descriptional inadequacy" (Fulcher, Davidson, & Kemp, 2011, p. 8), leaving room for dissimilar interpretations. Because there is ample evidence that even trained raters interpret the same test-specific criteria differently (Deygers & Gorp, 2015; Lumley, 2002) and that also trained raters' experience and background may influence the score that is assigned (Barkaoui, 2011), there can be no guarantee that different test developers interpret the same CEFR descriptors in the same way. To the best of our knowledge, no study has yet compared the ratings of two high-stakes tests using corresponding CEFR-based criteria.

Nonetheless, quite a few studies have discussed rating scale construction in relationship to the CEFR (Galaczi, Ffrench, Hubbard, & Green, 2011; Harsch & Martin, 2012; Papageorgiou, 2015). These studies typically discuss fitting CEFR descriptors to rating scale logic by rectifying descriptor vagueness and by straightening blurred lines between levels (Alderson, 2007; Papageorgiou, 2010). In addition, Galaczi *et al.* (2011) have highlighted the positive wording of CEFR descriptors and the brevity of certain CEFR scales as matters of ongoing concern during rating scale construction and rater training. Deygers and Gorp (2015) showed in their study that a CEFR-based rating scale that was iteratively constructed together with raters did not guarantee a uniform interpretation of the descriptors, despite high inter-rater reliability indices. In this regard Harsch and Rupp (2011) rightfully stressed the need for a high level of analytic detail in CEFR-based scales to compensate for the broadness of the initial descriptors.

The abovementioned studies show how individual test developers have operationalized CEFR descriptors in their rating scales to fit the purpose of a test. Other documents describe how some tests have been aligned with the CEFR (e.g., Tannenbaum & Wylie, 2008; Khalifa & Ffrench, 2009; De Jong, Becker, Bolt, & Goodman, 2014 for TOEFL iBT, IELTS, and Pearson PTE, respectively). Green (this issue) has scrutinized some of these reports in an effort to understand the varying ways in which the major English tests have established their CEFR links.

To date, little if any CEFR research has been comparative. Concurrent analyses themselves are not alien to the language testing endeavour, however, and have even been central to test validation since its earliest days (Moss, 2007). Concurrent validation is about determining the empirical relationship between two measures (Weir, 2005a) by estimating the strength of the relationship between two sets of scores, taken from the same candidates during two administrations not too far apart. (Bachman and Palmer (1996) comparison between the content and the operationalization of TOEFL and Cambridge's FCE is an important milestone in this regard.) Determining an acceptable correlation coefficient for concurrent validity is somewhat arbitrary, so it is useful to compare findings of the results of similar studies. Davies (1984) proposed setting the minimally acceptable correlation in concurrent validity studies at .45 because he correlated the ELTS with the ELBA (.78) and with the EPTB (.84). In a TOEFL iBT – IELTS concurrent validity study (ETS, 2010) the total correlation was .73, and the correlation between the oral components was .57. Riazi (2013) conducted a similar concurrent validity study with 60 students who self-reported their IELTS scores prior to taking the PTE Academic. Here, the overall correlation was .82, and the correlation between the oral components was .72.

To the best of our knowledge, no studies, concurrent or otherwise, have analysed empirical data to compare the interpretation and operationalization of CEFR descriptors across tests. Nevertheless, it could be argued that exactly this kind of research determines whether the CEFR can act as a catalyst for increased score transparency and score comparability, which was one of its original goals (Van Ek, 1975). The current study thus addresses an important gap in the literature by examining the potential of the CEFR in facilitating score transparency in tests that share the same purpose, the same population, and use corresponding rating criteria that are based on the same CEFR descriptors.

University admission tests in Flanders, Belgium

In Flanders, international L2 students must obtain a certificate of either the ITNA (*Inter University test of L2 Dutch*) or the STRT (*Ready to start higher education*) to be admitted to university. Both tests have been linked to the B2 level of the CEFR following the *familiarisation, specification, standard setting and validation* procedures described in Figueras, North, Takala, Van Avermaet, and Verhelst (2009), but their operationalization is quite dissimilar. A computer-based and face-to-face test, the ITNA is owned by a consortium of Flemish university language centres (Interuniversity Testing Consortium [IUTC]). Candidates need to pass the computer test to gain access to the oral exam. The STRT, on the other hand, is a task-based, integrated language skills test, consisting of four written and two oral tasks. The STRT is the largest international Dutch language test and is owned by the Dutch Language Union, an international, intergovernmental organization overseeing the Dutch language policy in the Netherlands, Belgium, and Suriname.

Typically, the oral STRT or ITNA components do not take more than 25 minutes, including preparation time. In both tests, candidates interact with a trained examiner during the oral component, which consists of a presentation and an argumentation task. The argumentation task invites the candidates to weigh a number of alternative solutions to a problem and to argue why their choice is the better one. In the presentation task the candidates briefly present a study by using input material such as graphs and tables. Oral ITNA performances are scored in situ by the examiner and an additional rater, who come to a joint overall score for five criteria: vocabulary, grammar, coherence, pronunciation, and fluency. The oral STRT component is administered by a trained examiner, recorded and centrally scored by two independent trained raters, by using a rating scale that includes five criteria corresponding to those used in the ITNA, plus register, taking initiative, and content (i.e., whether the candidate mentions all salient points asked for in the prompt).

Five oral rating criteria are included in both tests: vocabulary, grammar, coherence, pronunciation, and fluency. For scoring these criteria, both tests use analytic band descriptors that are based on the A2, B1, B2, and C1 levels in the corresponding CEFR scales. In both tests the cutoff level for each criterion is B2, except for pronunciation and grammar, where the ITNA uses B1 and B2+, respectively. Both tests developed their rating scales from the original descriptors, but both made choices based on their interpretations of the CEFR descriptors. ITNA rating scale designers often copied the original CEFR text and supplemented it with language-specific examples, identifying typical errors of users at a given level. The STRT criterion descriptors deviated from the original wording more often, to make the original descriptors more concrete and easier to grasp for the novice raters they often employ (see Deygers & Gorp, 2015 for a discussion).

The oral ITNA performances are scored immediately after the test by two trained raters who come to one composite score for each of the five criteria. ITNA examiners and raters tend to be experienced L2 teachers of Dutch who typically attend training at least once a year and score oral tests at different times throughout the year. The STRT performances are recorded and subsequently each task is separately scored by two independent trained raters. STRT raters are usually novice raters with a background in linguistics or communication. They undergo a two-day training and take part in a trial rating session to establish their consistency and reliability.

The ITNA and STRT certificates grant access to university and in that context they are considered to be equivalent measures of B2 ability by policy makers and admission officers. To date, their actual equivalence has not been verified.

Research questions

The present study investigates to what extent the STRT and ITNA tests score the same candidates in the same way. It also takes a detailed look at the way in which these tests score the same test takers for corresponding criteria, and questions whether using the CEFR to create corresponding criteria also leads to equivalent scores.

The first research question of the study deals with the overall correlation—the starting point for a more fine-grained analysis.

RQ1: How strong is the relationship between the ITNA and STRT test results?

The second research question focuses on the relationship between the rating criteria in the two tests and borrows from scoring validation studies by determining to what extent the same level descriptors were uniformly interpreted (Messick, 1989; Weir, 2005a). Determining how tests with a similar goal and similar task types use and interpret the same CEFR descriptors in their rating scales is a valuable check on the potential of the CEFR as a common currency in language proficiency levels (Figueras, 2012).

RQ2: Do the two tests apply the same CEFR-based level descriptors in the same way for similar task types?

To systematically answer this RQ, four subquestions were identified:

(a) How much do the STRT and ITNA criteria deviate from the original CEFR descriptors?
(b) Can corresponding CEFR-based levels in both tests be considered equivalent?

(c) Are corresponding CEFR-based criteria likely to measure the same construct?
(d) Are corresponding CEFR-based criteria equally difficult in both tests?

If both tests interpret the same CEFR descriptors in the same way, high overall correlations should carry through down to the criterion level. If the same candidates are rated highly dissimilarly on corresponding criteria, using the CEFR as a standard for score transparency may not be warranted, because it may create a false sense of uniformity.

Method

Study participants

During the data collection period, 485 candidates registered for the ITNA, 138 of whom agreed to first take the STRT test. After the candidates with incomplete data sets on one or both tests were removed, the overall scores of 118 candidates could be meaningfully compared. Because taking part in the ITNA oral test was conditional on gaining a minimum score of 54% on the computer test, the comparison of oral performance data relied on a subset of 82 candidates. This sample size is not exceptional in concurrent validation.

We applied *t*-tests to examine if the distribution of scores in this subset of the population was similar to that in the entire population. The validity of these *t*-tests was verified by using Levene's test for the equality of variances. The results showed that the distribution of scores found in the respondent population corresponds to the score distribution in the overall test populations. No significant differences were found between the final scores of the respondent population and those of the total ITNA population who took the test in the same period ($t = -.493$, $df = 485$, $p = .622$). For the STRT, this study was the first administration of a new test version, and apart from pilot data, no other scores were available. Levene's test for the equality of variances (Field, Miles, & Field, 2012) confirmed the variance comparability of the final scores of the sample population to the regular STRT population in the last administration of the previous STRT test ($F = 0.014$, $p = .907$).

The background variables of the participants corresponded with those of the actual populations of both tests (see Table 1). In terms of L1, the actual STRT population had a slightly different distribution due to the large number of candidates from Belgium's neighbouring countries, expat communities, and countries with historical ties to the Dutch language.

Data collection

Between June 2014 and September 2014 all ITNA candidates were invited to take the STRT free of charge one week before the ITNA administration, which granted them an extra opportunity to gain

Table 1. Research Population Variables vs. Regular STRT and ITNA Populations.

		Sample		ITNA		STRT	
Age	Mean (SD)	27 (7)		28 (8)		26 (7)	
	Min - Max	16 – 50		15 – 61		14 – 60	
Gender	Female	70%		67%		65%	
	Male	30%		33%		35%	
Goal	Educational	66%		68%		58%	
	Professional	25%		24%		30%	
	Proficiency	8%		7%		7%	
L1		French	17%	French	16%	French	29%
		Spanish	8%	Spanish	11%	German	25%
		Arabic	7%	Arabic	9%	Papiamento	7%
		Russian	7%	Russian	8%	Dutch	6%
		German	6%	German	5%	Russian	4%
N		138		485		521	

access to university. The predetermined stopping criterion for data collection was the start of the 2014–2015 academic year. The exams were administered at the largest Flemish universities (37% at the University of Antwerp, 34% at Ghent University, and 29% at the University of Leuven) by trained examiners. The first author was always on site to ensure the consistency of the test administration.

Rating the STRT test takes a few weeks because all performances are scored centrally. However, the ITNA scores in the current administration were available the day of testing. Because it was assumed that respondents would be disinclined to take a second test after being informed they had passed the ITNA, the STRT administration was conducted first. The candidates received no further formalised instruction between the tests, and given the one-week time span between the two administrations, it was assumed that their language skills remained constant. The respondents were rated anonymously under normal rating conditions.

Data analysis

Concurrent validation studies that focus on the correspondence of scores between two tests are often limited to correlation coefficients or *t*-tests, because they are often based on candidate self-reporting, and overall scores are all the researchers have at their disposal (Davies, 1984; Educational Testing Service, 2010; Riazi, 2013; Zheng & De Jong, 2011). An important difference between this study and other concurrent validity studies is that the data were collected from the same respondents in a controlled test/retest setting, and the researcher had access to all scores, allowing for a more detailed comparison than overall correlations allow.

RQ1: The strength of the relationship between the ITNA and STRT results was examined by recoding the scores onto a common percentile scale. Based on the descriptive statistics, a *t*-test was used to determine whether the mean scores were significantly different. Cohen's *d* served to quantify the magnitude of the difference. The correlations between the overall results and between the results on the oral component also were analyzed. Because of the sample size (Howell, 1997) and because it is generally more robust (Field *et al.*, 2012), Kendall's Tau was chosen over Spearman's correlation to correlate the ordinal scores on the oral tests. Independent *t*-tests were conducted to verify whether the overall mean scores on the tests differed significantly.

RQ2: This research questions concerns the scores on corresponding criteria. At issue here is whether CEFR-based criteria that recur in both tests yield comparable scores when applied to performances of the same participants on highly similar tasks. Consequently, criteria that only occur on the STRT rating scale (e.g., content, register) but are not part of the ITNA scale will not be considered here.

Determining whether both tests apply the same CEFR-based rating criteria in the same way required recoding certain scoring categories. The STRT rating scale distinguished four proficiency levels (A2, B1, B2, and C1), but the ITNA had six or seven, because it includes the plus levels of B1, B2, and occasionally A2. After consulting with the ITNA coordinators, these double bands were merged to come to a four-band scale, which facilitates direct comparisons across scales. The analysis of RQ2 will be discussed by subquestion:

How much do the STRT and ITNA criteria deviate from the original CEFR descriptors?

The STRT and ITNA rating criteria were compared to each other and to the Dutch translation of the CEFR, by using the Jaccard similarity index. The index provides a very simple quantification of the similarity between descriptors. It has been applied under different forms in the field of information retrieval and text comparison (Manning, Raghavan, & Schütze, 2008). The Jaccard index expresses the similarity between two descriptions as their overlap in terms, more specifically, the ratio of the number of unique terms present in both texts and the number of unique terms in either of the texts.

The Jaccard index becomes 1 if both texts use the exact same set of words, independent of how often these terms are repeated, and it decreases the more the terms used in both texts diverge.

To make the comparison of the descriptors used in this article more robust, the texts were automatically pre-processed by using a standard stemming algorithm for Dutch. This means that all words were stemmed (e.g., plural endings removed) and all non-informative words (such as "of" and "with," as defined by the Python NLTK Dutch stopword list) were removed.

Can corresponding CEFR-based levels in both tests be considered equivalent?

To determine the equivalence of the same levels in corresponding criteria, frequency distributions were supplemented with probability estimates. For every criterion the probability of attaining a score of B2 or higher was estimated. The strength of the relationship between corresponding criteria was calculated by using Kendall's Tau. To determine the level of agreement between the two tests, linear weighted kappa (K_w) was used. K_w is a variation on Cohen's kappa, which measures the level of agreement between ordinal data sets (Sim & Wright, 2005), whereby 0 indicates no agreement except one stemming from chance, and values above .8 can be read as almost perfect agreement (Landis & Koch, 1977; Vanbelle & Albert, 2009). Usually, weighted kappa is used to determine rater agreement, but in this study it served as an additional metric to determine whether the STRT and ITNA raters scored the same candidates in the same way for corresponding criteria.

Are corresponding CEFR-based criteria likely to measure the same construct?

The ITNA rating scale consists of five criteria: vocabulary, grammar, coherence, pronunciation, and fluency. These criteria also occur in the STRT rating scale, in addition to others (e.g., content, register). If the STRT and ITNA raters interpreted the same criteria in the same way, the STRT scores on the shared criteria would explain a large proportion of the total ITNA score variance. To investigate this, three multivariate linear regression models were constructed. Each model took the following general form:

$$\text{ITNA}_{\text{Total}} = \ (b_0 + b_1\,\text{criterion}_{1i} + b_2\,\text{criterion}_{2i} + \ldots + b_n\,\text{criterion}_{ni}) + \varepsilon_i$$

Three regression models were run and compared in terms of R^2 using an Anova. The first regression model included the five criteria from the two STRT tasks that significantly correlated with the ITNA criteria at $\tau > .3$. The second model included the seven criteria that significantly correlated, regardless of the strength of the correlation. The final model included all the STRT criteria. Prior to running the regression analyses, the assumptions were checked: The proportion of cases with large residuals was acceptable (4% in the oral component after removal of two outliers), Cook's distance was <1, no cases were larger than three times the average leverage, the covariance ratio was satisfactory, and the multicollinearity and independence assumptions were supported (Norris 2015; Purpura, Brown, & Schoonen, 2015).

Next, to determine the relationship between individual criteria, a linear regression model was constructed for every ITNA criterion as a function of the same STRT criterion.

Are corresponding CEFR-based criteria equally difficult in both tests?

In a multifaceted Rasch (MFR) measurement analysis a test score is seen as the result of an interaction between different facets, such as test-taker ability, task difficulty, rater severity, and criterion difficulty (McNamara, 1996). In MFR the effect of all these variables on the score is taken into consideration and mapped onto the same logit scale. In this study all comparable STRT and ITNA ratings for the same candidates were combined in the same MFR model, and the Facets program was used to estimate criterion difficulty. Of interest are the difficulty measures (a higher measure indicates a more difficult criterion), the strata index (which shows whether different measures also translate into different levels of difficulty that can be separated reliably), and the fit statistics (InfitMnSq). The closer the value of these fit statistics is to 1, the better the observed data fit the Rasch model. A criterion that has fit statistics in the range between .50 and 1.5 is considered to have an acceptable model fit. Lower values indicate overfit (i.e., redundancy), and higher values indicate misfit (Barkaoui, 2014; Linacre, 2012).

The analyses were conducted by using R (*psych, irr, Hmisc, QuantPsyc, car*, and *ggplot2* packages), Facets (Linacre, 2015), and Python (with the NLTK library).

Results

1. How strong is the relationship between the ITNA and STRT test results?

Table 2 shows the descriptive statistics of the total test scores and the oral scores, mapped onto a common percentile scale to facilitate comparison. It shows that the mean ITNA scores are lower than the mean STRT scores. The t-tests showed that these differences were statistically significant (total: t [236] = −9.20, $p < .001$; oral: t [162] = −9.036, $p < .001$), with large (total: $d = -1.19$; oral $d = -1.41$) effect sizes. The correlation ($N = 118$) between total STRT and ITNA test scores was substantially higher ($r = .767$**) than the one between oral test scores ($N = 82$) ($\tau = .387$**).

It is important to reiterate that these mean results are calculated from test scores that included all criteria. As such, these scores are based on additional criteria that are not always operationalized in both tests. For that reason all further analyses only consider the criteria that both tests share to facilitate a direct comparison between corresponding CEFR-based criteria. Hence, the second research question aims to determine whether the levels in the corresponding criteria are used similarly.

2. Do the two tests apply the same CEFR-based level descriptors in the same way for similar task types?

In answering this second research question it is important to first establish to what extent the STRT and ITNA rating scales deviated from the original descriptors. The Jaccard index (Table 3) shows that on the whole the wording of the ITNA criteria stays closer to the exact wording of the CEFR than the wording of the STRT criteria. For example, the lowest Jaccard index for *ITNA ~ CEFR* is .27, but three of five *ITNA ~ CEFR* indices are substantially lower than that ($J \leq .10$). Because the wording of the descriptors in both tests sometimes deviates substantially from the CEFR original, it is logical that the overlap between the STRT and ITNA descriptors is typically not too big. The rating scale descriptors of Pronunciation in both tests stay closest to the CEFR wording, and as such the overlap between the STRT and ITNA descriptors is the largest for this criterion ($J = .44$).

Table 2. Descriptive Statistics: Overall and Oral Component.

	Total		Oral	
	ITNA	STRT	ITNA	Oral
N	118	118	82	82
$\bar{X}$	48.06	67.68	50.97	73.16
SD	20.09	11.72	19.69	10.33
Med	51.63	69.19	48.75	72.54
SE	1.84	1.07	2.17	1.14

Table 3. Jaccard Index for Rating Descriptor Pairs.

	ITNA ~ CEFR	STRT ~ CEFR	ITNA ~ STRT
Vocabulary	.53	.15	.10
Grammar	.30	.10	.06
Coherence	.27	.09	.08
Pronunciation	.80	.40	.44
Fluency	.88	.26	.29

For reasons of confidentiality and space, the rating scale descriptors can, unfortunately, not be repeated in their entirety, but a few examples, taken from confidential STRT and ITNA rating scale documents, may serve to illustrate how CEFR descriptors were paraphrased.

The B2 criterion for vocabulary in the ITNA adds (italics) to the CEFR descriptor: “has a good range of vocabulary for matters connected to his/her field and most general topics *and does not only use high-frequency words*.” The STRT vocabulary criterion, on the other hand, reads: “the lexical variation in the performance is sufficient to prevent frequent repetition of words.”

In both tests the descriptors for coherence include additions to the CEFR wording. The ITNA focuses on the sentence level: “sentences are linked logically and appropriate connectors are used when required.” The STRT raters are required to consider the text level as well: “The performance is one coherent whole [...] connectors are mostly used correctly and support the overall coherence.” Pronunciation$_{\text{STRT}}$ repeats the original B2 descriptor, but it is supplemented with a B1 characteristic (italics): “The pronunciation is clear and natural, *but with a foreign accent*.” This addition does not occur in the ITNA rating scale, where the cutoff point for pronunciation is B1, not B2.

The fluency descriptor in the ITNA rating scale has literally been copied from the CEFR: “can produce stretches of language with a fairly even tempo; although he/she can be hesitant as he/she searches for patterns and expressions, there are few noticeably long pauses.” A few additions (italics) were made in the STRT descriptor: “can produce stretches of language with an even tempo; although he/she can be hesitant as he/she searches *for the right expression*, there are *few noticeable or distracting* pauses.”

After examining the correspondence in the wording of the rating criteria, it was determined whether the same levels of corresponding criteria in the STRT and ITNA rating scales can be considered equivalent. Table 4 shows how often CEFR levels were assigned to the same criteria in the ITNA and in both STRT tasks (coherence is not a rating criterion in the STRT argumentation task). In most cases the mode corresponds with the B2 level. In other words, on most criteria, most candidates scored B2 (e.g., 53 ITNA test takers scored B2 on vocabulary, and 8 scored C1). For grammar$_{\text{STRTpres}}$ and pronunciation$_{\text{ITNA}}$, B1 was the level most often assigned. No candidate scored A2 for grammar$_{\text{ITNA}}$ or coherence$_{\text{ITNA}}$ (ITNA assigned 0 A2, and 16 B1 ratings on coherence, while STRT scored 9 performances A2 and 31 B1).

For most criteria the probability of any given candidate to attain a score of B2 or more was found to be higher on the ITNA than on either of the STRT tasks (Table 5). The *p* values in Table 5 refer to the difference in probability between the two STRT tasks and the ITNA results. A given candidate has a 38% probability of attaining a B2 pronunciation score on the ITNA. The same candidate may, however, have a 68% probability of being rated B2 on the same criterion if he takes the STRT argumentation task. The difference between these probabilities is significant. In fact, vocabulary excepted, there is a consistent significant difference between the probability of attaining a score of at

Table 4. Frequencies of Assigned CEFR Levels.

		A2 (F)	B1 (F)	B2 (F)	C1 (F)
Vocabulary	ITNA	5	16	53	8
	STRT$_{\text{arg}}$	3	17	40	22
	STRT$_{\text{pres}}$	7	20	40	15
Grammar	ITNA	0	4	67	11
	STRT$_{\text{arg}}$	4	24	44	10
	STRT$_{\text{pres}}$	5	34	33	10
Coherence	ITNA	0	16	49	17
	STRT$_{\text{pres}}$	9	31	31	11
Pronunciation	ITNA	11	40	23	8
	STRT$_{\text{arg}}$	2	24	45	11
	STRT$_{\text{pres}}$	3	32	37	10
Fluency	ITNA	1	17	50	14
	STRT$_{\text{arg}}$	5	27	36	14
	STRT$_{\text{pres}}$	12	30	37	3

Table 5. Probability of Attaining B2 or Higher on STRT and ITNA.

	$P_{B2}^{STRTarg}$	$p^{\#}$	P_{B2}^{ITNA}	$p^{\#\#}$	$P_{B2}^{STRTpres}$
Vocabulary	.76	1.0	.74	.362	.67
Grammar	.66	.000	.95	.000	.52
Coherence			.80	.000	.51
Pronunciation	.68	.000	.38	.009	.57
Fluency	.61	.013	.78	.000	.49

Note. $p^{\#}$: *p* value for the difference in probability between $P_{B2}^{STRTarg}$ and P_{B2}^{ITNA}.
$p^{\#\#}$: *p* value for the difference in probability between $P_{B2}^{STRTpres}$ and P_{B2}^{ITNA}.

least B2 on the ITNA or one of the STRT tasks ($p < .05$). This indicates that the B2 threshold is interpreted or operationalized differently on the STRT and on the ITNA test.

The frequencies in Table 4 show regular discrepancies in STRT and ITNA judgments, and the probabilities in Table 5 indicate that STRT and ITNA judgments differ in severity from one criterion to the next (e.g., a B2 score on coherence is significantly harder to reach on STRT than it is on ITNA). As such, there likely is a different distribution in the ITNA and STRT scores for corresponding criteria.

Table 6 shows the ITNA score on both tasks combined in relation to the STRT argumentation task and the STRT presentation task. The relationship between the corresponding STRT and ITNA criteria is generally low ($\tau \leq .39^{**}$) and the agreement is generally weak ($k_w \leq .22$). The results presented in Table 6 imply that the STRT and ITNA results on corresponding criteria, used to assess the same people on highly comparable tasks, do not map onto each other well. The relationship is weakest for vocabulary and pronunciation. The correlation between the sums of these five corresponding criteria is weak as well ($\tau = .37^{**}$).

Having determined that the same levels in corresponding criteria are unlikely to be equivalent, multivariate linear regression (Table 7) was used to determine to what extent scores on the criteria that the STRT shares with the ITNA, predicted the ITNA scores. Three models were run. The first included the five STRT criteria from the two STRT tasks (Table 6), which significantly correlated at $\tau > .3$. This model explained 26% of the ITNA score variance ($R^2_{adj} = .2585$, $p < .000$). The second model included the seven STRT criteria that correlated significantly with the corresponding ITNA criterion, regardless of the strength of the relationship. The second model accounted for 27% of the

Table 6. Relationship Between Corresponding STRT and ITNA Criteria.

	ITNA ~ $STRT_{arg}$		ITNA ~ $STRT_{pres}$	
	τ	k_w	τ	k_w
Vocabulary	.153	.031	.212*	.091
Grammar	.336*	.208***	.351**	.184***
Coherence			.386**	.216***
Pronunciation	.117	.122*	.212*	.207**
Fluency	.336**	.215**	.315**	.134*

Note. *$p < .05$, **$p < .01$, ***$p < .001$.
Overall correlation for summed criteria $\tau = .37^{**}$.

Table 7. Multivariate Linear Regression: $ITNA_{total}$ ~ $STRT_{arg+pres}$.

	B	*SE B*	*Β*	*P*
(Constant)	.457	3.811		.905
Grammar $STRT_{arg}$	1.547	1.851	.132	.406
Grammar $STRT_{pres}$	2.633	2.056	.241	.204
Fluency $STRT_{arg}$	3.709	1.436	.375	.012*
Fluency $STRT_{pres}$	−2.251	1.604	−.222	.165
Coherence $STRT_{pres}$	1.204	1.365	.128	.381

Note. Total R^2 *adjusted* is .2585 (p = < .000).

total ITNA score variance ($R^2_{adj} = .2706$, $p < .000$), but it did not significantly improve the model fit of the data, in comparison with the first model ($F(2, 74) = 1.6334$, $p < .000$). The third multivariate linear regression model included all nine STRT criteria that had corresponding ITNA criteria and explained 26% of the ITNA score variance ($R^2_{adj} = .2603$, $p < .001$). Because this model was not a significantly better predictor than the first ($F(4, 72) = 1.0476$, $p < .39$) or the second ($F(2, 72) = 0.4828$, $p < .62$), Table 7 shows the regression results of the first model. It indicates that only one predictor (fluency, in the argumentation task) significantly contributes to the model. Given the sample size, the estimates of the individual predictors should be treated with some caution; however, generalizing from the overall model fit can be done with a degree of confidence (Field *et al.*, 2012). All in all, the multiple regression analyses show that no more than 27% of the ITNA score variance can be explained by scores on corresponding STRT criteria.

When corresponding ITNA and STRT criteria were used in a pairwise linear regression (Table 8), the same trend emerges. Regression models with statistical significance ($p < .05$) based on the STRT criteria never explain more than 17.2% of the score variance in corresponding ITNA criteria (Table 8).

The results of the regression analyses above indicated that corresponding criteria are unlikely to measure the same construct at the same level. To ascertain whether the same criteria are equally difficult in the two tests, a multifaceted Rasch analysis was conducted on the basis of all corresponding rating criteria. This model (Table 9) reliably showed that the STRT test was the more difficult test. This does not automatically imply that the STRT is also the most difficult test overall, because both tests also include a written component, as well as criteria that occur in one test alone. But in a side-by-side comparison of both tests, based on criteria that can be compared, the STRT is the hardest.

Table 10 shows the results from the criteria measurement in the Rasch analysis. It is important that these results showed that corresponding criteria were never included in the same difficulty bands. This implies that the difficulty level of every ITNA criterion is significantly different from its

Table 8. Linear Regression on Criterion Level: $\text{ITNA}_{\text{total}} \sim \text{STRT}_{\text{total}}$.

	r^2_{adj}		*B*	*SE B*	*β*	*p*
Vocabulary	.022					.156
		(Constant)	2.279	.336		.000
		STRT_{arg}	−.076	.211	−.081	.719
		$\text{STRT}_{\text{pres}}$	.252	.200	.281	.211
Grammar	.172					.000
		(Constant)	2.306	.189		.000
		STRT_{arg}	.122	.103	.196	.239
		$\text{STRT}_{\text{pres}}$	.156	.096	.267	.109
Coherence	.170					.000
		(Constant)	2.148	.215		.000
		$\text{STRT}_{\text{pres}}$	.325	.077	.425	.000
Pronunciation	.067					.017
		(Constant)	1.664	.400		.000
		STRT_{arg}	−.339	.250	−.276	.179
		$\text{STRT}_{\text{pres}}$	.605	.240	.513	.014
Fluency	.133					.001
		(Constant)	1.983	.261		.000
		STRT_{arg}	.213	.127	.259	.098
		$\text{STRT}_{\text{pres}}$	.135	.130	.160	.305

Table 9. MFR: STRT and ITNA, Arranged by Measure.

	Measure	SE	InfitMnSq
STRT	.23	.09	1.02
ITNA	−.23	.10	.96

Note. Model, Sample: Separation 3.28, Strata 4.71, Reliability .92.

Table 10. MFR: STRT and ITNA Criteria, Arranged by Measure.

	Measure	SE	InfitMnSq
$\text{Pronunciation}_{\text{ITNA}}$	1.14	0.20	1.52
$\text{Fluency}_{\text{STRT}}$	0.47	0.20	1.14
$\text{Coherence}_{\text{STRT}}$	0.01	0.21	1.21
$\text{Grammar}_{\text{STRT}}$	−0.07	0.21	0.85
$\text{Pronunciation}_{\text{STRT}}$	−0.42	0.21	0.98
$\text{Vocabulary}_{\text{ITNA}}$	−0.6	0.22	0.94
$\text{Vocabulary}_{\text{STRT}}$	−1.02	0.22	0.88
$\text{Coherence}_{\text{ITNA}}$	−1.15	0.23	0.71
$\text{Fluency}_{\text{ITNA}}$	−1.21	0.23	0.99
$\text{Grammar}_{\text{ITNA}}$	−1.64	0.24	0.48

Note. Dotted line indicates difficulty band.

STRT counterpart. Moreover, the Rasch output generally aligns well with the probabilities displayed in Table 5. For example, pronunciation in ITNA is the most difficult criterion in the Rasch table and also had the lowest probability score. The probability scores for vocabulary were not significantly different, and in this table too, the measures of the vocabulary criteria of both tests are mapped closest to each other. Nevertheless, despite the results pertaining to the vocabulary scores, this study has yielded no data to indicate that corresponding CEFR-based criteria used to measure the same candidates in near-identical tasks can be considered equivalent.

Discussion and conclusion

The CEFR in its current form was not designed as a ready-to-use normative tool, and its descriptors are unsuitable for unaltered use in rating scales (Council of Europe, 2001; Figueras, 2012; Galaczi *et al.*, 2011; North, 2014a). Therefore, changing the CEFR descriptors to meet the needs of a test is common, recommended practice (North, 2007, 2014a; Weir, 2005b), and it is not unlikely for two tests that were aligned to the same CEFR level to differ substantially in terms of content or construct. In the minds of test score users, however, these tests might be considered equivalent exactly because they share the same CEFR level—a problem that this article has attempted to exemplify.

Every analysis in this study indicates the same trend: corresponding CEFR-based criteria in the ITNA and STRT rating scales are not equivalent. If they were, the correlations would be stronger, the kappa values would show more agreement, the linear regression model would explain more variance, and the same criteria would fall within the same Rasch difficulty bands. One explanation for the divergences can be found in the rating scale descriptors. Even though both tests started from the same CEFR descriptors, they diverged in interpretation and operationalization. The Jaccard index indicated that the descriptors are indeed quite dissimilar, as seen in some of the operationalizations discussed above. In short: the statistical analyses in this study fail to confirm the assumption that the STRT and ITNA descriptors interpret the same CEFR levels in an equivalent way and provide arguments to the contrary. Both tests have developed rating scales from the same source and adopted the same level system, but the relationship between equivalent criteria is weak. If the CEFR levels were true, unequivocal standards, this should not occur, but given the nature of the CEFR descriptors, the findings are not unexpected.

The root of the problem addressed in this article lies not so much in the CEFR itself as in the reification of its levels as standards (Fulcher, 2004). The CEFR is often referred to as a gold standard, because it is so eagerly used by all parties involved in European language testing, but there is one very important difference: exactness. The collective agreement that exactly one ounce of gold traded for exactly $20.67 made the gold standard the backbone of the global economic system for decades. The CEFR intentionally lacks such exactness, however, which makes it unusable as a standard. CEFR levels do not exist outside of the minds of the practitioners, and B2 is not an entity. As a standard, it shares less resemblance to screw threads or monetary systems than to primary colours. The colour blue has a marked beginning and an end but encompasses a range from light aquamarine to dark navy; it would be wrong to argue that only Pantone 2736C is the true blue. Likewise, it is problematic to consider the B2 level in one rating scale equivalent to the next, simply because both have been based on the same broad level.

On the surface, using the same CEFR levels across tests may seem to increase score comparability and transparency, but this study offers no results to suggest that taking the CEFR as a starting point for rating scale development leads to equivalent ratings or increased transparency. Of course, the CEFR may be a useful inspiration for test developers to reflect on language proficiency levels, but it is not a standard that can simply be applied to reach equivalent scores. If the CEFR is to be a true catalyst for score comparability, it is not enough for two comparable tests to develop scales that draw on the same source, because this could create a false sense of uniformity. To attain true score equivalence in rating criteria, two comparable tests used in the same context could consider developing scales together, training raters together, or equating already developed rating scales (Kolen & Brennan, 2014 offer an overview of possible methods). Ensuring that CEFR levels are interpreted similarly in equivalent tests may rely on intertest collaboration and discussion, which has been a major goal of the CEFR all along. However, as the opening article of this special issue indicates, it is also a neglected one.

Limitations

Even though the demographics of the sample population are representative for the actual test population, the sampling methodology used in this study could qualify as convenience sampling. This may have limitations in terms of generalizability. This study used real-life performances and actual rating data. Performances were rated under the standard STRT and INTA rating conditions, and this may have impacted the scores. However, because it was the goal of this article to investigate differences between real-life scores to gauge the real-life impact of mismatching judgments, we decided to collect the data in actual rather that laboratory conditions. This had implications for the performances that could be compared. Because the ITNA regulations only allow candidates who passed the written component to take the oral exam, the number of respondents who could meaningfully be compared was reduced. Although all statistical assumptions were checked prior to the analyses (Purpura *et al.*, 2015), range restriction may have had an effect on the data, weakening the correlations.

Funding

This work was funded by the Research Foundation Flanders, FWO Vlaanderen, under Grant number G078113N.

References

ACTFL. (2012). *ACTFL proficiency guidelines 2012*. Alexandria, VA: American Council on The Teaching of Foreign Languages.

Alderson, J. (1991). Bands and scores. In J. C. Alderson, & B. North (Eds.). *Language testing in the 1990s* (pp. 71–86). London, England: Macmillan.

Alderson, J. (2007). The CEFR and the need for more research. *The Modern Language Journal*, *91*(4), 659–663. doi:10.1111/modl.2007.91.issue-4

Bachman, L. F., & Palmer, A. S. (1996). Language Testing in Practice: Designing and Developing Useful Language Tests. Oxford: Oxford University Press.

Bärenfänger, O., & Tschirner, E. (2012). Assessing evidence of validity of assigning CEFR ratings to the ACTFL Oral Proficiency Interview (OPI) and the Oral Proficiency Interview by computer (OPIc) (Technical Report 2012-US-PUB-1). Leipzig, Germany: Institute for Test Research and Test Development.

Barkaoui, K. (2011). Effects of marking method and rater experience on ESL essay scores and rater performance. *Assessment in Education: Principles, Policy & Practice*, *18*(3), 279–293. doi:10.1080/0969594X.2010.526585

Barkaoui, K. (2014). Multifaceted rasch analysis for test evaluation. In A. Kunnan (Ed.). *The companion to language assessment* (pp. 1301–1322). Hoboken, NJ: John Wiley & Sons, Inc.

Baztán, A. (2008). *La evaluación oral: Una equivalencia entre las guidelines de ACTFL y algunas escalas del MCER*. Granada, Spain: Universidad de Granada.

Council of Europe. (2001). *Common European framework of reference for languages: Learning, teaching, assessment*. Strasbourg, France: Author.

Davies, A. (1984). Validating three tests of English language proficiency. *Language Testing*, *1*(1), 50–69. doi:10.1177/026553228400100105

De Jong, J. H. A. L., Becker, K., Bolt, D., & Goodman, J. (2014). Aligning PTE Academic Test Scores to the Common European Framework of Reference for Languages. Pearson. Retrieved from https://pearsonpte.com/wp-content/uploads/2014/07/Aligning_PTEA_Scores_CEF.pdf

Deygers, B., & Gorp, K. (2015). Determining the scoring validity of a co-constructed CEFR-based rating scale. *Language Testing*, *32*(4), 521–541. doi:10.1177/0265532215575626

Deygers, B., Zeidler, B., Vilcu, D., & Hamnes Carlsen, C. (2017). One Framework to Unite Them All? Use of the CEFR in European University Entrance Policies. *Language Assessment Quarterly*. doi:10.1080/15434303.2016.1261350

ETS. (2010). *Linking TOEFL iBT TM scores to IELTS scores—A research report*. Retrieved July 8 2015 from https://www.ets.org/s/toefl/pdf/linking_toefl_ibt_scores_to_ielts_scores.pdf

Field, A., Miles, J., & Field, Z. (2012). *Discovering statistics using R*. London, England: Sage Publications Ltd.

Figueras, N. (2012). The impact of the CEFR. *ELT Journal*, *66*(4), 477–485. doi:10.1093/elt/ccs037

Figueras, N., North, B., Takala, S., Van Avermaet, P., & Verhelst, N. (2009). *Relating language examinations to the common european framework of reference for languages: Learning, teaching, assessment (CEFR). A manual*. Stasbourg, France: Council of Europe.

Fulcher, G. (2004). Deluded by artifices? The Common European Framework and Harmonization. *Language Assessment Quarterly*, *1*(4), 253–266. doi:10.1207/s15434311laq0104_4

Fulcher, G. (2012). Scoring performance tests. In G. Fulcher, & F. Davidson (Eds.). *The Routledge handbook of language testing* (pp. 378–392). London, England and New York, NY: Routledge.

Fulcher, G., Davidson, F., & Kemp, J. (2011). Effective rating scale development for speaking tests: Performance decision trees. *Language Testing*, *28*(1), 5–29. doi:10.1177/0265532209359514

Galaczi, E., Ffrench, A., Hubbard, C., & Green, A. (2011). Developing assessment scales for large-scale speaking tests: A multiple-method approach. *Assessment in Education: Principles, Policy & Practice*, *18*(3), 217–237.

Gomez, P., Noah, A., Schedl, M., Wright, C., & Yolkut, A. (2007). Proficiency descriptors based on a scale-anchoring study of the new TOEFL iBT reading test. *Language Testing*, *24*(3), 417–444. doi:10.1177/0265532207077209

Harsch, C., & Martin, G. (2012). Adapting CEF-descriptors for rating purposes: Validation by a combined rater training and scale revision approach. *Assessing Writing*, *17*(4), 228–250. doi:10.1016/j.asw.2012.06.003

Harsch, C., & Rupp, A. (2011). Designing and scaling level-specific writing tasks in alignment with the CEFR: A test-centered approach. *Language Assessment Quarterly*, *8*(1), 1–33. doi:10.1080/15434303.2010.535575

Howell, D. (1997). *Statistical methods for psychology*. Belmont, CA: Duxbury.

Kane, M. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, *50*(1), 1–73. doi:10.1111/jedm.2013.50.issue-1

Khalifa, H., & Ffrench, A. (2009). Aligning Cambridge ESOL examinations to the CEFR: Issues and practice. *Cambridge Research Notes*, *37*, 10–15.

Kolen, M. J., & Brennan, R. L. (2014). Test Equating, Scaling, and Linking: Methods and Practices. Springer Science & Business Media.

Landis, J., & Koch, G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159–174. doi:10.2307/2529310

Linacre, J. (2012). *A user's guide to FACETS Rasch-model computer programs*. Retrieved November 20, 2013 from www.winsteps.com/a/facets-manual.pdf

Linacre, M. (2015). *Facets (Version 3.71.4)*. Beaverton, OR: Winsteps.com.

Little, D. (2007). The common European framework of reference for languages: Perspectives on the making of supranational language education policy. *The Modern Language Journal*, *91*(4), 645–655. doi:10.1111/modl.2007.91.issue-4

Lumley, T. (2002). Assessment criteria in a large-scale writing test: What do they really mean to the raters? *Language Testing*, *19*(3), 246–276. doi:10.1191/0265532202lt230oa

Manning, C. D., Raghavan, P., & Schütze, H. (2008). Introduction to Information Retrieval. New York: Cambridge University Press.

McNamara, T. (1996). *Measuring second language performance*. London, England: Longman.

Messick, S. (1989). Validity. In R. Linn (Eds). *Educational measurement* (pp. 13–103). New York, NY: Macmillan.

Moss, P. (2007). Reconstructing validity. *Educational Researcher*, *36*(8), 470–476. doi:10.3102/0013189X07311608

Norris, J. M. (2015). Statistical Significance Testing in Second Language Research: Basic Problems and Suggestions for Reform. Language Learning, 65(S1), 97–126.

NATO. (2014). *Standardization agreement STANAG 6001 language proficiency levels*. Brussels, Belgium: Bureau for International Language Coordination.

North, B. (2007). The CEFR illustrative descriptor scales. *The Modern Language Journal*, *91*(4), 656–659. doi:10.1111/modl.2007.91.issue-4

North, B. (2014a). *English profile studies. The CEFR in practice* (Vol. 4). Cambridge: Cambridge University Press.

North, B. (2014b). Putting the Common European Framework of Reference to good use. *Language Teaching*, *47*(02), 228–249. doi:10.1017/S0261444811000206

O'Sullivan, B., & Weir, C. (2011). Testing and validation. In B. O'Sullivan (Ed.). *Language testing: Theory and practice* (pp. 13–32). Oxford, England: Palgrave.

Papageorgiou, S. (2010). Investigating the decision-making process of standard setting participants. *Language Testing*, *27*(2), 261–282. doi:10.1177/0265532209349472

Papageorgiou, S., Xi, X., Morgan, R., & So, Y. (2015). Developing and validating band levels and descriptors for reporting overall examinee performance. *Language Assessment Quarterly*, *12*(2), 153–177. doi:10.1080/15434303.2015.1008480

Purpura, J. E., Brown, J. D., & Schoonen, R. (2015). Improving the validity of quantitative measures in applied language research. *Language Learning*, *65*(1), 36–73. doi:10.1111/lang.12112

Riazi, M. (2013). Concurrent and predictive validity of Pearson Test of English Academic (PTE Academic). *Papers in Language Testing and Assessment*, *2*(2), 1–27.

Roever, C., & McNamara, T. (2006). Language testing: The social dimension. *International Journal of Applied Linguistics*, *16*(2), 242–258. doi:10.1111/ijal.2006.16.issue-2

Sim, J., & Wright, C. C. (2005). The Kappa Statistic in Reliability Studies: Use, Interpretation, and Sample Size Requirements. Physical Therapy, 85(3),257–268.

Swender, E. (2010). A tale of two tests. STANAG and CEFR. Comparing the Results of side-by-side testing of reading proficiency, Paper presented at *BILC*, Istanbul.

Tannenbaum, R. J., & Wylie, C. E. (2008). *Linking English-language test scores onto the common European framework of reference: An application of standard-setting methodology*. Princeton, NJ: ETS.

Taylor, L. (2004). Issues of test comparability. *Cambridge Research Notes*, *15*, 2–5.

Tschirner, E., Bärenfänger, O., & Wisniewski, K. (2015). *Assessing evidence of validity of the ACTFL CEFR listening and reading proficiency tests (LPT and RPT) using a standard-setting approach* (Technical Report 2015-EU-PUB-2). Leipzig, Germany: Institute for Test Research and Test Development.

van Ek, J. A. (1975). *Systems development in adult language learning: The threshold level in a European-Unit/Credit system for modern language learning by adults*. Strasbourg, France: Council of Europe.

Vanbelle, S., & Albert, A. (2009). A note on the linearly weighted kappa coefficient for ordinal scales. *Statistical Methodology*, *6*(2), 157–163. doi:10.1016/j.stamet.2008.06.001

Weir, C. (2005a). *Language testing and validation*. New York, NY: Palgrave Macmillan.

Weir, C. (2005b). Limitations of the common european framework for developing comparable examinations and tests. *Language Testing*, *22*(3), 281–300. doi:10.1191/0265532205lt309oa

Zheng, Y., & De Jong, J. (2011). *Research note: Establishing construct and concurrent validity of pearson test of English academic*. Pearson Education Ltd. Retrieved March 2 2015 from http://pearsonpte.com/wp-content/uploads/2014/07/RN_EstablishingConstructAndConcurrentValidityOfPTEAcademic_2011.pdf