clac CÍRCULO
de
lingüística
aplicada a la
comunicación

67/2016

ITEM AND TASK DIFFICULTY IN A B2 READING EXAMINATION: PERCEPTIONS OF TEST-TAKERS AND CEFR ALIGNMENT EXPERTS COMPARED WITH PSYCHOMETRIC MEASUREMENTS

Andrej Stopar and Gašper Ilc

University of Ljubljana

andrej stopar *at* ff uni-lj si, gasper ilc *at* ff uni-lj si

Abstract

The article presents a study of a CEFR B2-level reading subtest that is part of the Slovenian national secondary school leaving examination in English as a foreign language, and compares the test-taker actual performance (objective difficulty) with the test-taker and expert perceptions of item difficulty (subjective difficulty). The study also analyses the test-takers' comments on item difficulty obtained from a while-reading questionnaire. The results are discussed in the framework of the existing research in the fields of (the assessment of) reading comprehension, and are addressed with regard to their implications for item-writing, FL teaching and curriculum development.

Stopar, Andrej and Ilc, Gašper. 2016.
Item and task difficulty in a B2 reading examination: perceptions of test-takers and CEFR alignment experts compared with psychometric measurements.
Círculo de Lingüística Aplicada a la Comunicación 67, 318-342.
http://revistas.ucm.es/index.php/CLAC/article/view/53487
http://www.ucm.es/info/circulo/no67/stopar.pdf
http://dx.doi.org/10.5209/CLAC.53487

Key words: reading comprehension, difficulty perception, CEFR alignment judges, psychometric measurement

Contents

1. Introduction

Following the well-established distinction between the objective and subjective difficulty (Fulmer and Tulis 2013), the present study aims at determining possible correlations and interdependencies between these two types of difficulty, with special attention being paid to their importance for item-writers, test-, policy- and curriculum-developers as well as CEFR[1]-alignment experts.

The study draws on the reading comprehension subtest of the Slovenian national end-of-secondary-school-leaving exam in English, called the General Matura in English (henceforth GM), which has only recently been fully validated and aligned with the B2 level of the European CEFR scale (Bitenc Peharc and Tratnik 2014). For the purposes of the investigation, the GM reading subtest has been administered to a group of test-takers together with a while-reading questionnaire, in which the test-takers have commented on their perception of the item/task difficulty. In order to determine to what extent the objective difficulty correlates with the subjective difficulty, the study compares (i) the psychometric measurements of the reading subtest (objective difficulty) with (ii) the answers from the while-reading questionnaire as well as with the judgments of the language experts that have aligned the GM examination with the CEFR (subjective difficulty). The reason for including the language expert into the study is twofold. First, in our context, most of the language experts participating in the CEFR alignment project are also item-writers for the national examinations, and second we want to address the question of experts and their reported weak ability to predict the item/task difficulty (Alderson and Lukmani 1989; Sydorenko 2011). In addition, the in-depth analysis of the test-takers' while-reading questionnaire is employed to identify the underlying factors that can contribute to the item/task difficulty and influence test-taker performance.

We strongly believe that apart from theoretical implications, the results of our investigation will also have practical value especially in educational environments where the test-provider does not follow all the standardised test-design procedures as

[1] Common European Framework of Reference for Languages (Council of Europe 2001)

described in Green (2014) among others. For example, in Slovenia, the national high-stakes examinations, including the GM, are neither piloted nor pre-tested (Ilc, Rot Gabrovec, and Stopar 2014; Šifrar Kalan and Trenc 2014). Consequently, the item-writers, test-developers and (CEFR) alignment experts must solely rely on their subjective judgment regarding the item/task difficulty. Their misjudgement about the item/task difficult may negatively affect the test validity and reliability, which is an undesired result, particularly so in the case of high-stake examinations. Therefore, a better understanding of item/task difficulty may have positive ramifications for the test validity/reliability.

2. Reading Comprehension and Item/Task difficulty: Basic Tenets

Reading has long been treated as a cornerstone of foreign language (FL) teaching. The mid-20th-century notion of reading as one of the four discrete FL skills remains relevant today (Hinkel 2010), even though it has been amended by the findings of many studies showing that reading should not be treated as a single, monolithic skill but rather as a complex and extensive set of activities that involves multifarious skills. As Alderson (2000) points out, the lists of reading (sub-)skills and the descriptions of how they interact are numerous and various, depending on the theorist researching them. Some that have frequently surfaced in the literature and have persisted for decades include decoding, linguistic knowledge, knowledge of discourse structure, knowledge of the world, synthesis and evaluation, metacognitive knowledge, and others (Bloom et al. 1956; Grabe 1991; Koda 2005; Munby 1978; Urquhart and Weir 1998). Khalifa and Weir (2009) propose a detailed, 7-point taxonomic scale of the reading ability which involves (from the lowest to highest): word recognition, lexical access, syntactic parsing, establishing propositional meaning, inferencing, building a mental model, and creating a text-level structure.[2] While the lower levels mostly deal with lexis and syntax that are explicitly recoverable from the text, the higher levels focus on the contextual dimensions of reading such as recognizing the implicit meaning, connecting the text

[2] The empirical studies have shown that the hierarchical ordering of the proposed levels should be slightly modified (Wu, 2011; Ilc and Stopar, 2014)

with the knowledge of the world as well as establishing intra-/inter-textual links. A similar proposal is put forward by Grabe (2009), and it distinguishes between word recognition, syntactic parsing, and proposition encoding (lower levels) from text processing strategies, background knowledge/inferencing, and understanding the discourse structure as well as the context of the reading act (higher levels). All of these levels are fully interconnected and '[c]omprehension cannot occur without the smooth operation of these processes' (Grabe 2009: 22). Given these assumptions, one would expect that there is a direct correlation between a taxonomic level and the reading comprehension difficulty: higher taxonomic skills should intrinsically be more difficult than the lower ones. However, the empirical study of Brunfaut and McCray (2015) has shown that such an overgeneralisation is problematic. According to their study, some readers have been able to aim their attention at higher order skills exclusively, making little use of lower level skills. This conclusion also supports previous claims that the difficulty level of a particular reading subskill cannot be directly linked to the taxonomic levels. For instance, Alderson and Lukmani (1989: 268) observe that some linguistically weaker test-takers perform overall 'somewhat better on the higher order questions than on lower order questions'. They attribute this fact to their non-linguistic cognitive skills abilities. Harding, Alderson and Brunfaut (2015: 7) again point out that reading skills also need to be closely linked with different cognitive processes, including working memory capacity, attention and the automaticity of word recognition. Due to these factors, the question of 'how to diagnose problems at the higher level, or problems related to the interactions between lower-and higher-level processes, is less clear' (ibid.).

Despite these observed and reported discrepancies between the taxonomic and difficulty levels, the contemporary FL teaching practices and policies, by and large, follow the assumption that the relative taxonomical ranking of a particular comprehension skill directly reflects the skill complexity and difficulty. This strategy is evident in the CEFR (Council of Europe 2001). The document, for instance, describes the reading ability of a B2 student as one that includes reading different types of discourse; dealing with 'contemporary problems', which can be interpreted part of the reader's knowledge of the world; and recognising 'particular attitudes or viewpoints' (Council of Europe 2001: 27). These descriptors can be directly linked with building a mental model and

inferencing taxonomic levels of Khalifa and Weir's (2009) classification. In contrast, the reading ability of an A2 student is defined by descriptors that are associated with lower taxonomic levels (word recognition and lexical access), for instance, 'can understand short, simple texts containing highest frequency vocabulary' (Council of Europe 2001: 69).

After the publication of Bachman's (1990) and Bachman and Palmer's (1996) seminal works on language testing, a lot of research has been dedicated to testing reading comprehension, and also to the relationship between factors that give rise to item/task difficulty. As Fulmer and Tulis (2013) observe, two different types of item/task difficulty have been discussed: the objective and the subjective difficulty. While the former mostly pertains to readability that can be objectively measured by using different tools and item/task analysis, the latter involves a subjective judgment of difficulty based on cognitive, motivational and emotional factors (Efklides 2002; Fulmer and Tulis 2013).

Discussing the objective difficulty, Freedle and Kostin (1993, 1999) analyse in detail factors such as vocabulary selection, sentence/passage length, topic (abstract vs. concrete), syntactic features (rhetorical organiser, referentials, fronting, negation), text organisation (topicalisation), and item type (explicit/implicit detail, explicit/implicit gist, textual organisation/structure). When addressing the relationship between the item type and difficulty, which is also discussed in this paper, Freedle and Kostin (1999: 18) observe that at least as far as the listening comprehension testing is concerned, the items that involve identifying the main idea and inference-application are easier than inference items. Lund (1991) reports that given the same language proficiency, test-takers find main-idea items and inference items easier than supporting idea items in the case of listening comprehension, whereas the situation is exactly the reverse with reading comprehension.

The perceived (i.e., subjective) difficulty involves both ability and affective variables. While the ability variables (intelligence, aptitude, cognitive style) are more permanent and can be diagnosed ahead of time, the affective variables (confidence, motivation, anxiety) are more temporary and less predictable (Robinson 2001: 32). Consequently, the reported discrepancies between the objective and subjective difficulty can be attributed to affective variables (Fulmer and Tulis, 2013).

The theoretical considerations involving the complexity of the reading process as well as FL testing (see above) have led authors such as Alderson (2000) and Spaan (2007) to suggest that a valid and reliable reading comprehension test should always contain an appropriate selection of tasks and texts that not only test the appropriate micro skills but also include tasks (and items) targeting the intended level of difficulty. Such a requirement, coupled with the requirements of the curricula increasingly aligned with the CEFR, presents a significant challenge for testing and assessment (Figueras 2012; Fulcher 2004). This is especially the case with examinations for which the curriculum also serves as the test construct. In our context, the GM test developers and item-writers are faced with the responsibility of creating valid tests that adhere to the requirements of their exam constructs, which, in turn, are rigidly aligned to the CEFR. The item-writers are also not supported by (external) validation and evaluation of the items (e.g., piloting, pre-testing). Thus, the test validity and reliability exclusively depends on the item-writers' and test-developers' judgments about item/task difficulty. The importance of pinpointing the desired difficulty level is also demonstrated by projects and studies focusing on relating examinations to the CEFR and identifying alternatives to (often impractical) piloting and pre-testing procedures (cf. Alderson et al. 2004; Bitenc Peharc, and Tratnik 2014; Cizek 2001; Council of Europe 2009; Hambleton and Jirka 2006; Kaftandijeva 2010; Little 2007; Martyniuk 2010; Sydorenko 2011). Along these attempts, the research presented herein explores to what extent the test-takers' as well as language experts' subjective perception of the item/task difficulty can be used as an alternative to piloting and pre-testing.

3. The Study

3.1. Context

The study presents three different reading comprehension tasks from the GM, in relation to the item difficulty as shown by psychometric measurements and the perception of test-takers and the CEFR-relating experts. For the purposes of the research, we have collected the test-takers' psychometric measurements, the test takers' answers to the while-reading questionnaire on the item difficulty, and the experts' judgments of item

difficulty. The reading tasks and the while-reading questionnaires were administered for the purposes of the present investigation only (i.e., they were not part of the GM administration); however, the GM administration guidelines were strictly followed.

The GM is a high-stakes exam, serving both as an achievement test (i.e. as a national secondary school-leaving exam), and as a proficiency test (i.e. as the tertiary education entrance exam). The GM is provided and administered by the Slovenian National Examinations Centre (RIC), and it comprises three obligatory and two elective subjects. One of the obligatory subjects is a FL. The GM in English consists of five subtests: the reading and listening comprehension, language in use, writing and speaking. The former four subtests are administered on the national level and marked externally; the last is administered by the Matura school committees using standardised prompts and criteria.

From 2008 to 2013, the RIC conducted a project that aligned all national examinations in English with the CEFR (Council of Europe 2001). The judging panel consisted of eleven to twelve experienced Slovenian education professionals. Most of the panellists are primary, secondary or tertiary teachers of English that cooperate with the RIC as item-writers and/or test-developers. The project's final report was published in 2014, claiming that the GM is aligned with the B2 level of the CEFR scale (Bitenc Peharc, and Tratnik 2014).

3.2. Participants

The data presented herein was collected from responses of a total of 83[3] test-takers, who are all non-native speakers of English. With regard to EFL and the CEFR-levels, they all share the same background: they have English as an FL1 subject in their curricula, and their expected proficiency level is, according to the curricula, B2. The test-takers were selected randomly from the GM population from different Slovenian secondary schools (last-year students, age range from 17 to 19). The participants were required to complete three different reading comprehension tasks that were originally administered

[3] The original number of participants was 100 but 17 test-takers did not complete the while-reading questionnaire.

by the RIC together with the accompanying while-reading questionnaire. Comparing the performance of the 83 test-takers included in the study with the performance of the test-takers that originally sat for the GM, we can observe a high level of consistency in correlations: 0.89, 0.77, and 0.87 for Tasks 1, 2, and 3 respectively, which suggests that our sample is representative of the GM test-taker population.

3.3. Instruments

3.3.1. Reading comprehension subtest

The three reading tasks included in the study were taken from the RIC test paper bank and were administered by the test provider in autumn 2009 to 1,022 test-takers (Tasks 1 and 2), and in spring 2013 to 4,375 test-takers (Task 3). The reason for selecting these three reading tasks for our research is twofold. First, Tasks 1 and 2 were also used by the panellists that aligned the GM reading subtest to the CEFR levels, so by using these two tasks, we have been able to compare the perception of item difficulty from the perspective of both the test-takers and the panellists. Second, Task 3 was selected intentionally to create a representative array of task-types that frequently occur in the GM reading subtests: Task 1 is an short-answer (SA) task type (Items 1–10), Task 2 (Items 11–20) is a gapped-text (GT) task type, and Task 3 (Items 21–28) is a multiple-choice (MC) task type. Following Freedle and Kostin's (1999) classification of items, we identified detail explicit (D-E) items (12 items), detail implicit (D-I) items (2 items), gist explicit (G-E) items (2 items), gist implicit (G-I) items (2 items), and items targeting at textual organisation/structure (O-S) (10 items). The items are presented in Table 1.

Table 1. The twenty-eight reading items used in the reading comprehension test

Item No.	Type	Target	Item No.	Type	Target	Item No.	Type	Target
1	SA	D-E	11	GT	O-S	21	MC	G-E
2	SA	D-E	12	GT	O-S	22	MC	D-E
3	SA	D-I	13	GT	O-S	23	MC	D-E

4	SA	D-E	14	GT	O-S	24	MC	D-E
5	SA	G-I	15	GT	O-S	25	MC	D-E
6	SA	D-E	16	GT	O-S	26	MC	G-E
7	SA	D-E	17	GT	O-S	27	MC	G-I
8	SA	D-I	18	GT	O-S	28	MC	D-E
9	SA	D-E	19	GT	O-S	-	-	-
10	SA	D-E	20	GT	O-S	-	-	-

3.3.2. While-Reading Questionnaire

Together with the reading tasks, the respondents were given a while-reading questionnaire, which had to be completed after answering each item. The respondents were asked to answer two questions for each item: (i) whether they found the item easy/moderate/difficult (a close-ended question), and (ii) what made the item easy/moderate/difficult (an open-ended question). We decided that the test-takers should evaluate the item difficulty level on a three-point scale (i.e. easy/moderate/difficult), so that the results can be directly compared with the facility values[4] from the official exam reports of the test provider. Taking into consideration the test provider's ranking of the items which coincides with those proposed in the relevant literature (see Bailey 1998) we assigned the numeric values to the participants' descriptive responses as follows. Test items marked as easy were assigned the value 0.95, items marked as moderate the value 0.50, and items marked as difficult the value 0.05.

To analyse the replies to the open-ended question, we have applied the method of clustering (Miles and Huberman 1994) which involves first identifying general topics and then breaking them down to more specific sub-topics.

[4] These are calculated with the classical test theory. Facility values together with other statistic data and their interpretation are included in the test-provider's final report published electronically for each administered exam (http://www.ric.si/splosna_matura/statisticni_podatki/?lng=eng).

3.3.3. CEFR Standard Setting

The experts' judgments presented herein are taken from the Slovenian alignment project. As stated in the project's final report (Bitenc Peharc and Tratnik 2014), the reading subtest of the GM was aligned to B2 level, its cut score set at 80%. During the standard setting procedure for the reading comprehension subtest, the panellists used the combination of the Angoff and the Basket Methods[5] (op. cit.: 10) in order to minimize the influence of a particular method on the final standard setting results, which is also in accordance with the recommendations for the CEFR-alignment projects (Council of Europe 2009: 61-65, 75-77; Kaftandijeva 2010: 131). Using the Basket Method, the experts ranked items as B1, B2 and C1, each abbreviation reflecting a CEFR level during the evaluation procedure. We converted their descriptive evaluations into numeric values in the same fashion as the participants' judgments about item difficulty. Since the GM targets the B2 level, we considered B1 items as easy, B2 items as moderate, and C1 items as difficult. Consequently, the numeric values of 0.05, 0.50 and 0.95 were assigned, respectively. As shown later (section 3.5), our proposed numeric conversion of the Basket judgments highly correlates with the experts' numeric item difficulty perception values based on the Angoff Method.

3.4. Research Questions

RQ1: How do test-taker perceptions/expert judgments of item difficulty correlate with test-taker performance?

RQ2: What are the characteristics of the reading comprehension items that exhibit the greatest differences between test-taker perceptions/expert judgments and psychometric statistics?

---

[5] The Basket Method builds directly on the connection between an item and the CEFR descriptors. To align an item to the CEFR level, a panellist has to establish at what CEFR level a test-taker can already answer the item correctly (Council of Europe 2009: 75). The Angoff Method, on the other hand, is based on the notion of 'minimally acceptable person' or a 'minimally competent candidate' at a targeted level. For each item, a panellist has to decide how likely it is that such a test-taker will answer correctly. (Council of Europe 2009: 61).

RQ3: What are test-takers' comments (in the while-reading questionnaire) on the factors that influence item difficulty?

3.5. Results

Table 2 below presents the findings on test-taker perceptions and expert perceptions (subjective difficulty) and test-taker actual performance (objective difficulty).

Table 2. Test-taker perceptions and expert judgments of Items 1–28 combined with test-taker actual performance

Item	Test-taker Perceptions (N=83)					Expert Judgments (N=11)						Test-taker Actual Performance (FV)
	Easy (%)	Moderate (%)	Difficult (%)	No answer (%)	Perceived FV (Converted)	B1 (Easy)	B2 (Moderate)	C1 (Difficult)	Perceived FV (Converted Basket)	Perceived FV (Angoff)	Average Perceived FV (Basket & Angoff)	
1	67 (81%)	10 (12%)	4 (5%)	2 (2%)	0.85	9 (82%)	2 (18%)	0 (0%)	0.87	0.75	0.81	0.98
2	65 (78%)	14 (17%)	2 (2%)	2 (2%)	0.85	9 (82%)	2 (18%)	0 (0%)	0.87	0.76	0.81	0.90
3	28 (34%)	13 (16%)	40 (48%)	2 (2%)	0.43	3 (27%)	8 (73%)	0 (0%)	0.62	0.67	0.65	0.84
4	32 (39 %)	23 (28%)	26 (31%)	2 (2%)	0.53	3 (27%)	8 (73%)	0 (0%)	0.62	0.66	0.64	0.87
5	26 (31%)	17 (20%)	38 (46%)	2 (2%)	0.43	1 (9%)	8 (73%)	2 (18%)	0.46	0.59	0.52	0.66
6	60 (72%)	16 (19%)	5 (6%)	2 (2%)	0.81	8 (73%)	3 (27%)	0 (0%)	0.83	0.79	0.81	0.96
7	67 (81%)	14 (17%)	0 (0%)	2 (2%)	0.87	7 (64%)	4 (36%)	0 (0%)	0.91	0.81	0.86	0.97
8	48 (58%)	24 (29%)	9 (11%)	2 (2%)	0.72	7 (64%)	4 (36%)	0 (0%)	0.79	0.77	0.78	0.97
9	51 (62%)	20 (24%)	10 (12%)	2 (2%)	0.73	7 (64%)	4 (36%)	0 (0%)	0.79	0.77	0.78	0.90
10	62 (75%)	15 (18%)	4 (5%)	2 (2%)	0.82	6 (55%)	5 (45%)	0 (0%)	0.75	0.75	0.75	0.82
11	57 (69%)	13 (16%)	3 (4%)	10 (12%)	0.83	10 (91%)	1 (9%)	0 (0%)	0.91	0.76	0.84	0.98
12	58 (70%)	12 (15%)	3 (4%)	10 (12%)	0.84	7 (64%)	4 (36%)	0 (0%)	0.79	0.71	0.75	0.95
13	47 (57%)	13 (16%)	13 (16%)	10 (12%)	0.71	6 (55%)	1 (9%)	4 (36%)	0.70	0.70	0.70	0.84
14	48	15	10	10	0.73	5	6	0	0.70	0.69	0.70	0.92

	(58%)	(18%)	(12%)	(12%)		(45%)	(55%)	(0%)				
15	47 (57%)	14 (17%)	12 (15%)	10 (12%)	0.72	4 (36%)	7 (64%)	0 (0%)	0.66	0.69	0.68	0.84
16	35 (42%)	11 (13%)	27 (33%)	10 (12%)	0.55	3 (27%)	8 (73%)	0 (0%)	0.62	0.67	0.65	0.74
17	50 (60%)	14 (17%)	9 (11%)	10 (12%)	0.75	7 (64%)	4 (36%)	0 (0%)	0.79	0.72	0.75	0.97
18	51 (62%)	15 (18%)	7 (8%)	10 (12%)	0.77	8 (73%)	3 (27%)	0 (0%)	0.83	0.76	0.79	0.91
19	55 (66%)	14 (17%)	4 (5%)	10 (12%)	0.81	9 (82%)	2 (18%)	0 (0%)	0.87	0.75	0.81	0.90
20	44 (53%)	16 (19%)	13 (16%)	10 (12%)	0.69	5 (45%)	6 (55%)	0 (0%)	0.70	0.69	0.70	0.84
21	43 (52%)	21 (25%)	6 (7%)	13 (16%)	0.74	n/a	n/a	n/a	n/a	n/a	n/a	0.97
22	23 (28%)	16 (19%)	31 (37%)	13 (16%)	0.45	n/a	n/a	n/a	n/a	n/a	n/a	0.51
23	32 (39%)	14 (17%)	24 (29%)	13 (16%)	0.55	n/a	n/a	n/a	n/a	n/a	n/a	0.66
24	27 (33%)	18 (22%)	25 (30%)	13 (16%)	0.51	n/a	n/a	n/a	n/a	n/a	n/a	0.70
25	25 (30%)	19 (23%)	26 (31%)	13 (16%)	0.49	n/a	n/a	n/a	n/a	n/a	n/a	0.55
26	24 (29%)	14 (17%)	31 (37%)	15 (18%)	0.46	n/a	n/a	n/a	n/a	n/a	n/a	0.81
27	14 (17%)	13 (16%)	43 (52%)	13 (16%)	0.31	n/a	n/a	n/a	n/a	n/a	n/a	0.56
28	42 (51%)	21 (25%)	7 (8%)	13 (16%)	0.73	n/a	n/a	n/a	n/a	n/a	n/a	0.55

3.5.1. Correlations and Comparisons

With regard to RQ1, we find that the correlation between test-taker perceptions of item difficulty and test-taker performance (Items 1–28) is relatively high, at 0.73. The correlation between expert judgments and test-taker performance (Items 1–20) is very high, at 0.83.

The test-takers perceive the test as noticeably more difficult than it actually is (the respective average facility values for Tasks 1–3 are 0.67 and 0.82). Their predictions are most reliable for Task 1 (correlation: 0.70) and Task 2 (correlation: 0.88), whereas the correlation between the perceptions and the performance is the lowest for Task 3, at 0.44. In contrast, the gap between the average perceived facility value (0.53) and the average performance facility value (0.66) is the least noticeable for the same task.

Unfortunately, an identical comparison between the test-takers and the alignment experts is not possible since the data from the CEFR alignment project do not include expert difficulty judgments on the multiple-choice task (Task 3). Despite this limitation,

we can observe that even if the correlation analysis is confined to the first two tasks, the result remains unchanged; namely, the correlation between test-taker perceptions and performance is at 0.73.

Focusing on individual items from the tasks that the test-takers and the judges had in common, we can observe a high degree of agreement about the items perceived/judged as the most difficult or the easiest. For instance, the data show that four out of the five items that were perceived as the most difficult are the same in both groups (Items 5, 4, 16 and 3) – albeit not in the same order. Both groups also share the perception that four among the five easiest items in the first two tasks are Items 1, 2, 7 and 11.

For reasons of practicality, the detailed presentation of results in the following sections is limited to a maximum of five testing items that are (i) perceived as the most difficult; (ii) perceived as the easiest, and (iii) most noticeably misperceived.

3.5.2. Items Perceived as the Most Difficult

The five items that the test-takers perceived as particularly challenging in Tasks 1–3 are, in order of perceived difficulty, Items 27, 3, 5, 22 and 26. The first three require the test-takers to process implicit information; additional factors affecting their difficulty are comparison (Item 27), negation (Item 3) and the fact that all preclude syntactic/lexical lifting. Items 22 and 26 target explicit information but involve the processing of numerous details and key words (in the text and distractors).

The five items that the experts marked as the most difficult in Tasks 1–2 are, in order of perceived difficulty, Items 5, 4, 16, 3 and 15. Items 5 and 3 are presented above. Items 4 and 15 rely on the test-takers' being familiar with the C1[6] word 'flee' and the low-frequency, subject specific word 'doge' (chief magistrate of the Venetian Republic). Item 15 is cognitively demanding since it includes contrasting.

[6] The CEFR level as provided in the online dictionary Cambridge Dictionaries Online (based on the English Vocabulary Profile).

3.5.3. Items Perceived as the Easiest

The five items that the test-takers saw as the easiest in Tasks 1–3 are, starting with the easiest, Items 7, 1, 2, 12 and 11. The short-answer items test the ability to identify explicit details and allow the answers to be recovered verbatim. The gapped-text items are syntactically and lexically undemanding and contain explicit lexico-grammatical cohesion links to the rest of the text.

The same justification can be given for the five items that the experts judged as the easiest in Tasks 1–2: starting with the easiest, these are Items 7, 11, 2, 19 and 1.

3.5.4. Gaps between Perceptions and Performance

In line with RQ2, we also observed the characteristics of the reading comprehension items that exhibit the greatest differences between test-taker perceptions/expert judgments and psychometric statistics.

The five items most noticeably misperceived are Items 3, 26, 4, 27 and 8 (Tasks 1–3). All are judged as more difficult than they are according to statistics. The difficulty of these items is related to their implicitness (Items 3 and 8), demanding vocabulary (Items 4 and 26) and comparison (Item 27).

The items that the experts most noticeably misperceived (in Tasks 1–2) are items 4, 14, 17, 12, 3 and 8 (the sixth item is included in the analysis because the numerical gap between their perceptions and actual performance was identical for Items 3 and 8). Similarly to the test-takers, the experts perceive the items as being more difficult than they actually are. Items 3, 4 and 8 are discussed above, while for Items 12, 14 and 17, we can observe that they are structurally ambiguous (from a lexico-grammatical perspective more than one option fits the gap).

3.5.5. Test-takers' Qualitative Comments

In the while-reading questionnaire, the test-takers were asked to provide comments on the factors that influence item difficulty. Their responses are presented based on task-types. Given the unstable status of the affective variables (see above), we must mention that the perceived difficulty of the test-takers included in our study may be somewhat

different from the test-takers sitting for the GM examination due to different circumstances (different testing situation, motivation, etc.).

The five items that the test-takers marked as the most difficult in Tasks 1–3 include two short-answer items (3 and 5) and three multiple-choice items (22, 26 and 27).

The 64 comments for the short-answer items (3 and 5) have been clustered as follows:

- The answer is not explicitly stated (62.50%): 'the answer has to be deduced from the text', 'the answer is not in the text';

- Issues with the question (25.00%): 'misleading', 'difficult to understand';

- Issues with the text (9.38%): 'the text is not clear', 'the article is ambiguous';

- General comments (3.12%): 'I don't know the answer', 'difficult'.

Items 22, 26 and 27 are perceived as very demanding owing to the following (82 comments):

- Issues with the text (31.71%): 'I had to reread the text', 'I don't understand the text', 'the text is not clear';

- Difficult vocabulary (19.51%): 'there are important words I don't understand', 'I don't understand the word "faultless"' (Item 26);

- The answer is not explicitly stated (18.29%): 'you have to read between the lines', 'not stated directly' (such descriptions refer exclusively to Item 27);

- Issues with the multiple-choice question (15.85%): 'the options are strange', 'two answers seem possible', 'all the answers refer to the whole paragraph';

- General comments (14.64 %): 'I don't know the answer', 'difficult'.

The comments about the easiest items refer to short-answer items and gapped-text items. The content and the distribution of the 161 explanations pertaining to the short-answer items (1, 2 and 7) are very homogenous:

-The answer is explicitly stated (86.96%): 'the answer can be quickly found in the text', 'this is mentioned in the text', 'you just copy it from the text';

- Comments about the question (9.33%): 'the question is clear';

- General comments (3.73%): 'logical', 'I know the answer'.

The 93 comments explaining the simplicity of the gapped-text items (11 and 19) include:

- The gap is (clearly) linked to the rest of the text (54.84%): 'connected to the rest', 'good text coherence', 'the only possible/logical/suitable answer';

- References to lexico-grammatical features (19.35%): 'known words', 'simple vocabulary and content', 'the structure of the sentence shows you where everything belongs';

- Comments about the text (19.35%): 'a clear text', 'a simple text';

- General comments (6.45%): 'easy'.

There may be some overlap between the first two categories of comments about the gapped-text items: the difference between 'a sensible continuation', 'the only logical answer' and 'the structure of the sentence shows you where everything belongs' is debatable. Conflating these categories would result in 74.19% of the comments referring to textual structure and organization.

The five items with the greatest gap between test-taker perceptions and performance are Items 3, 26, 4, 27 and 8. The 55 comments on the short-answer items (3, 4 and 8) include:

- Not explicitly stated in the text (54.55%): 'the answer is not evident' (mostly referring to detail implicit Items 3 and 8);

- Comments about the text (25.45%): 'the text is difficult to understand', 'the text is ambiguous';

- Comments about the question (16.36%): 'a tricky question', 'a misleading question';

- General comments (3.64%): 'illogical', 'I don't know'.

The 58 comments on the misperceived multiple-choice items (26 and 27) contain the following explanations:

- Comments about the text (29.31%): 'the text includes many pieces of information', 'the text is difficult to understand';

- Not explicitly stated in the text (25.86%): 'not stated in the text', 'requires reading between the lines' (these comments refer to Item 27 exclusively);

- Vocabulary issues (24.14%): 'many new words', 'faultless?' (see Item 26), 'I don't exactly understand the word "accurate"';

- Issues with the question (10.34%): 'all the options are similar' (Item 27);

- General comments (10.34%): 'I don't know the answer', 'difficult', 'rather demanding'.

4. Discussion and Conclusions

The study explores the relationship between objective and subjective difficulty of the GM reading comprehension text. Overall, the findings confirm the predictions of taxonomies proposed by Khalifa and Weir (2009) and Grabe (2009): the higher the taxonomic level, the more challenging the item is for the reader. The most frequent difficulties reported by our test-takers thus involve the higher-order skills of inferencing and text processing. Nevertheless, a common factor contributing to the difficulty of the test is also vocabulary, the recognition of which is ranked as a lower-order skill. We can observe that all these reading obstacles are reliably detected by both the test-takers and the expert judges.

Our empirical data show that test-takers are reliable judges of item difficulty. Their perceptions closely correlate (0.73) with their performance on the examination. This observation corroborates the previous findings of Apostolou (2010: 45-47) on test-taker perceptions of item difficulty in listening comprehension texts; thus we have proved that a similar conclusion can be extended to the reading comprehension as well. Another finding is that the CEFR alignment judges are even more accurate in their assessment of testing items (0.83), which is expected given their training and professional experience. This supports the findings of Fortus, Coriat and Fund (1998), who report a very similar correlation of 0.82 for trained judges assessing reading comprehension items. In contrast, our result partly refutes some previous studies (Alderson and Lukmani 1989; Sydorenko 2011) that claim that (experienced) item writer intuitions are weak predictor of item difficulty. We propose this difference is the result of the training that the CEFR-relating judges received. The observed correlations in herein attest to the reliability of

both test-taker difficulty perceptions and expert judgments, and thus prove their relevance for test-design, CEFR-alignment procedures, and assessment in teaching.

Despite the otherwise consistently high correlations between the perceptions (of both the test-takers and the experts) and psychometric data, we can observe that in some items the differences are quite pronounced. In the case of test-takers, this is typical for items that test implicit information, prevent recovering the answers verbatim, and contain low frequency vocabulary. Common issues are also with overall comprehension of the text – even with items that target explicit information. In the group of experts, the most problematic items are also related to less frequent language and, with regard to gapped-text items, to gaps that are structurally ambiguous and thus rely mostly on the comprehension of the context. Also noteworthy is that the most noticeably misperceived items often overlap with the items perceived/judged as the most difficult. It would appear that the perception of difficulty is intensified when test-takers or judges encounter the most challenging items.

A closer analysis of the reading items also reveals that the items perceived as the most demanding involve processing implicit information and main ideas. Such a finding confirms Lund's (1991) study that established these factors as challenging in a reading test and supports the idea that there is a link between the difficulty of the skill and its taxonomical position (cf. Freedle and Kostin 1999; Khalifa and Weir 2009). However, we also observe that some of the items that were perceived as very difficult are detail explicit items. Our analysis and the test-takers' responses in the while-reading questionnaire indicate that such items include some other factor contributing to their difficulty, such as overall text comprehension issues and, quite frequently, the presence of challenging or over-demanding vocabulary. This also demonstrates the impact that linguistic factors have even on lower order questions: a test-taker may be cognitively capable of higher-level processing (in terms of Grabe 2009) but still fails to answer a question owing to a word recognition issue. In comparison, the items perceived as the least demanding involve the identification of explicit details that can be recovered verbatim from the text. With regard to the gapped-text items, we conclude that the main factors contributing to their simplicity are syntactic and lexical accessibility. This, too, is consistent with the relevant literature: for instance, the impact of sentence length and complex vocabulary has been shown in Freedle and Kostin's work (1993), and both

notions are included in the CEFR descriptors. The above findings are well supported by the responses collected in the while-reading questionnaire.

The study also highlights some valuable insights in task-types. Firstly, the results (quantitative and qualitative) further attest to the importance of using a variety of task-types in a test (see Alderson 2004 and Weir 2005, for instance). Secondly, while all are consistently perceived as more difficult than they really are, the average facility values perceived by the test-takers are the most accurate for the most difficult of the three tasks in this study, i.e. the multiple-choice task. This finding is relevant in light of Sydorenko's (2011: 43) claim that item writers seem to have difficulties distinguishing between intermediate and advanced level items. If item writers fail to distinguish some difficulty levels, then this gap can be filled by including the perceptions of test-takers who have proved to be very successful in predicting the average facility values of the most difficult task in our study. Admittedly, judging from the observed correlations, the experts are not very likely to fail in their predictions; however, in contexts where item-writers are unable to receive sufficient training, such an alternative to piloting and pre-testing procedures is desirable.

The findings presented herein will not only help test-developers and item-writers predict item/task difficulty and give them an insight into test-takers' perception of difficulty but also provide practical implications for FL teaching and curriculum development. For instance, the study shows that the micro skills in reading comprehension that B2-level students feel most insecure about include searching for main ideas and, perhaps most significantly, reading for implicit information. In addition, the data indicate that more emphasis should be placed on the strategies of tackling unknown vocabulary. Such skills, incidentally, are already part of the CEFR descriptors for the level B2, which serve as the curricular basis for the national reading test analysed in this study.

References

Alderson, J. C., and Lukmani, Y. (1989) Cognition and Reading: Cognitive Levels as Embodied in Test Questions. Reading in a Foreign Language 5 (2), 253-270.

Alderson, J. C. (2000) Assessing Reading. Cambridge: Cambridge University Press.

Alderson, J. C., Figueras, N., Kuijper, H., Nold, G., Takala, S., and 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.Working paper. Retrieved February 26, 2014, from Lancaster University Publications & Outputs: http://eprints.lancs.ac.uk/44/1/final_report.pdf

Apostolou, E. (2010) Comparing Perceived and Actual Task and Text Difficulty in the Assessment of Listening Comprehension. Papers from the LAEL PG 2010 (pp. 26-47). Lancaster: Department of Linguistics and English Language (LAEL), Lancaster University. Retrieved April 7, 2014, from http://www.ling.lancs.ac.uk/pgconference/v05/Apostolou.pdf

Bachman, L. F. (1990) Fundamental Considerations in Language Testing. Oxford: Oxford University Press.

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

Baily, K. M. (1998) Learning about language assessment: Dilemmas, decisions, and directions. Boston: Heinle Cengage Learning.

Bitenc Peharc, S., and Tratnik, A. (2014) Umestitev nacionalnih izpitov iz angleščine v skupni evropski okvir. Zaključno poročilo o izvedbi projekta. Ljubljana: Državni izpitni center.

Bloom, B. S., Engelhart, M. D., Furst, E. J., Hill, W. H., and Krathwohl, D. R. (1956) Taxonomy of educational objectives: The classification of educational goals. Handbook I: Cognitive domain. New York: David McKay Company.

Brunfaut, T. and McCray, G. (2015) Looking into test-takers' cognitive processes whilst completing reading tasks: A mixed-method eye-tracking and stimulated recall study. (ARAGs Research Reports – Online. Vol. 1, No. 1). London: British Council.

Cizek, G. J. (Ed.). (2001) Setting Perfomance Standards, Concepts, Methods, and Perspectives. Mahwah, New Jersey, London: Lawrence Erlbaum Associates, Publishers.

Council of Europe. (2001) Common European Framework of Reference for Languages: Learning, Teaching, Assessment. Cambridge: Cambridge University Press.

Council of Europe. (2009) Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEFR). A Manual. Strasbourg: Language Policy Division.

Efklides, A. (2002) Feelings and judgments as subjective evaluations of cognitive processing: how reliable are they? Psychology: The Journal of the Hellenic Psychological Society 9, 163e184.

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

Fortus, R., Coriat, R. and Fund, S. (1998) Prediction of item difficulty in the English subtest of Israel's inter-university psychometric entrance test. In A. Kunnan (Ed.), Validation in Language Assessment: Selected Papers from the 17th Language Research Colloquium (pp. 61-87). Long Beach, Mahwah, New Jersey: Lawrence Erlbaum Associates, Inc.

Freedle, R. O., and Kostin, I. W. (1993) Item Difficulty for Expository Prose Passages for Three Item Types: Main Idea, Inference, and Supporting Idea Items. Princeton, New Jersey: Educational Testing Service.

Freedle, R., and Kostin, I. (1999) Does the text matter in a multiple-choice test of comprehension? the case for the construct validity of TOEFL's minitalks. Language Testing 16 (2), 2-32. Retrieved April 17, 2014, from http://ltj.sagepub.com/content/16/1/2. DOI: 10.1177/026553229901600102

Fulcher, G. (2004) Deluded by Artifices? The Common European Framework and Harmonization. Language Assessment Quarterly 1(4), 253-266. DOI: 10.1207/s15434311laq0104_4

Fulmer, S. M., and Tulis, M. (2013) Changes in interest and affect during a difficult reading task: Relationships with perceived difficulty and reading fluency. Learning and Instruction 27, 11-20. DOI:10.1016/j.learninstruc.2013.02.001

Grabe, W. (2009) Reading in a Second Language. Moving from Theory to Practice. Cambridge: Cambridge University Press.

Grabe, W. (1991) Current developments in second-language reading research. TESOL Quarterly 25 (3), 375–406.

Green, A. B. (2014) Exploring Language Assessment and Testing: Language in Action. London and New York: Routledge.

Hambleton, R., and Jirka, S. (2006) Anchor-based methods for judgmentally estimating item statistics. In S. Downing and T. Haladyna (Eds.), Handbook of test development (pp. 399-420). Mahwah, Nj: Erlbaum.

Harding, L., Alderson, C. and Brunfaut, T. (2015) Diagnostic assessment of reading and listening in a second or foreign language: elaborating on diagnostic principles. Language Testing. Prepublished January 27, 2015. DOI:10.1177/0265532214564505.

Hinkel, E. (2010) Integrating the four skills: Current and historical perspectives. In R. B. Kaplan (Ed.), Oxford Handbook in Applied Linguistics, Second Edition (pp. 110-126). New York: Oxford University Press.

Ilc, G. and Stopar, A. (2014) Validating the Slovenian national alignment to CEFR: The case of the B2 reading comprehension examination in English. Language Testing. Prepublished December 23, 2014. DOI:10.1177/0265532214562098.

Ilc, G., Rot Gabrovec, V., and Stopar, A. (2014) Relating the Slovenian secondary-school English language national examinations to the CEFR: Findings and implications. Linguistica 54 (1), 293-308. DOI:10.4312/linguistica.54.1.293-308.

Kaftandijeva, F. (2010) Methods for Setting Cut Scores in Criterion-referenced Achievement Tests. A comparative analysis of six recent methods with an application to tests of reading in EFL. Cito, Arnhem: EALTA. Retrieved February 5, 2014, from http://www.ealta.eu.org/documents/resources/FK_second_doctorate.pdf

Khalifa, H., and Weir, C. J. (2009) Examining Reading: Research and Practice in Assessing Second Language Reading. Cambridge: Cambridge University Press.

Koda, K. (2005) Insights into Second Laguage Reading. A Cross-Linguistic Approach. Cambridge: Cambridge University Press.

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/j.1540-4781.2007.00627_2.x

Lund, R. J. (1991) A Comparison of Second Language Listening and Reading Comprehension. The Modern Language Journal, 75(2), 196-204. Retrieved April 28, 2014, from http://www.jstor.org/stable/328827. DOI:10.1111/j.1540-4781.1991.tb05350.x

Martyniuk, W. (Ed.). (2010) Aligning Tests with the CEFR: Reflections on Using the Council of Europe's Draft Manual (Studies in Language Testing). Cambridge: Cambridge University Press.

Miles, M. B. and Huberman, A. M. (1994) An Expanded Sourcebook: Qualitative Data Analysis. Second Edition. London: London Sage Publications.

Munby, J. (1978) Communicative Syllabus Design. Cambridge: Cambridge University Press.

Robinson, P. (2001) Task Complexity, Task Difficulty, and Task Production: Exploring Interactions in a Componential Framework. Applied Linguistics 22 (1), 27-57. DOI:10.1093/applin/22.1.27

Spaan, M. (2007) Evolution of a test item. Language Assessment Quarterly 4, 279-293. DOI: 10.1080/15434300701462937

Sydorenko, T. (2011) Item Writer Judgments of Item Difficulty Versus Actual Item Difficulty: A Case Study. Language Assessment Quarterly 8, 34-52. DOI: 10.1080/15434303.2010.536924

Šifrar Kalan, M., and Trenc, A. (2014) Relating reading comprehension in the Spanish as a foreign language national exam to the CEFR: some aspects of evaluation. Linguistica 54 (1), 309-323. DOI:10.4312/linguistica.54.1.309-323.

Urquhart, A. H., and Weir, C. (1998) Reading in a Second Language: Process, Product and Practice. London, New York: Longman

Wu, R. Y. (2011) Establishing the Validity of the General English Proficiency Test Reading Component through a Critical Evaluation on Alignment with the Common European Framework of Reference (Unpublished doctoral dissertation). University of Bedfordshire, Bedford, Luton Bedfordshire, and Milton Keynes, Buckinghamshire, UK. Retrieved June 12, 2014, from http://uobrep.openrepository.com/uobrep/bitstream/10547/223000/1/wu%20ESTABLISHING%20THE%20VALIDITY%20OF%20THE%20GENERAL.pdf

Received: March 22, 2015

Accepted: July 2, 2016

Published: September 23, 2016

Updated: September 27, 2016