CEFR JOURNAL—RESEARCH AND PRACTICE
VOLUME 3

Investigating the difficulties for university learners of English in Japan of CEFR B1-level phrases

Takeshi Matsuzaki, Meiji University
Kevin Mark, Meiji University

https://doi.org/10.37546/JALTSIG.CEFR3-4

This study examines the variations in difficulty encountered by university learners of English in Japan with regard to English phrases that are classified as CEFR B1 level by the English Vocabulary Profile (EVP). Of the 332 English phrases categorized as B1 level at the time of investigation, 60 were identified as worthy of close examination for this research. An English phrase test was created, comprising two sections: one testing recognition and the other productive ability. Each section consisted of 60 test items embedded within 11 short written passages that were devised to provide context for the items used in this study. The test was administered to 360 university students in Japan, with the recognition section given first and the production section following immediately after. The results obtained show that there was a wide variation in the difficulty measures of the 60 phrases, and that recognition and production showed a discrepancy in terms of level of difficulty. The latter finding suggests that determination of difficulty based on a single processing mode is unlikely to provide an adequate indication of the difficulty of phrases, and that use should therefore be made of measurement instruments that assess both recognition and production ability. Implications for learning and pedagogy and for future directions for this line of research are discussed.

Keywords: English Vocabulary Profile (EVP), B1 level, phrases, difficulty, recognition, production

1 Introduction

The English Vocabulary Profile or EVP (http://vocabulary.englishprofile.org), a service provided by Cambridge University Press, describes a number of English phrases, in addition to different meanings of individual words, that are estimated to be 'typically known and used' by learners at different proficiency levels designated by the Common European Framework of Reference for Languages (CEFR; Council of Europe 2001; Harrison and Barker 2015). Because everyday language use teems with multi-word expressions (e.g., Erman and Warren 2000) and deviant use of such expressions by L2 learners can result in an increased and sustained processing burden for proficient speakers of the target language (Millar 2010; Stengers et al. 2011), the EVP serves as a valuable online resource for both L2 learners and teachers of English, enabling them to make more informed decisions as to what phrases to focus on at a given stage of learning English as a second language.

The primary objective of this study was to identify the difficulty measures of the selected B1 phrases that the authors identified as worthy of scrutiny, in the expectation that this will eventually lead to well-informed guidance for L2 learners of English in higher education in Japan and also for professionals working with them. The choice to target B1 phrases was made in consideration of the proficiency levels of college learners of English in Japan. We estimated, in accordance with the results of two surveys reported on in Negishi (2012), that the CEFR levels of the great majority of college learners of English

in Japan are likely to fall within A1 and A2 levels, and therefore phrases categorized under the B1 level were selected as items that represent reasonable target items for such learners to work on in order to advance to a higher level.

As far as the authors are aware, the only study to date that has looked into the difficulties for Japanese learners of English phrases sorted by the EVP is the investigation by Negishi, Tono and Fujita (2012). In their study, 100 English 'phrasal verbs' categorized by the EVP under CEFR A1 to B2 levels were examined. The researchers developed a fill-in-blanks phrasal verb test and administered it to some 1,600 Japanese learners, approximately 95% of whom were senior high school students. Negishi et al.'s study demonstrated that the difficulty measures of the phrasal verbs in each of the four levels examined vary to a large extent. Likewise, we anticipated that markedly differing difficulty measures would be found for our target B1 phrases.

One unique characteristic of our study is that it represents an attempt to delve into learners' depth of knowledge of the target phrases, which is not a matter that can be ticked off as simply present or absent: a learner at a certain learning stage may be able to recognize a phrase when it comes to comprehending input, while being unable to produce the same phrase as output. Unlike Negishi et al.'s study, our study sought a more textured understanding of the knowledge of B1 phrases possessed by college learners of English in Japan, by means of conducting an English phrase test that contains both recognition and production components.

2 Method

2.1 Participants

The participants in the study were undergraduate students in Japan taking courses taught by one of the co-authors. A total of 360 students took the two-section phrase test developed for the study, details of which are given below. The data for 59 students, however, were determined to be 'misfits' by the Rasch model, which we decided to adopt for our measurement due to the small sample at hand. The data for these misfit participants were then excluded from further difficulty analyses of the target phrases, as the primary objective of this study was to identify the difficulty measures of the target phrases, not to assess the learners' ability *per se.* This is a relatively large number of discounted misfits, but we were still able to ensure a sample size that is more than sufficient for Rasch measurement. The dispersion of the ability measures obtained of the 301 'fit' participants can be found in Table 2 and Figure 1 below.

2.2 Target B1 phrases

At the outset of the study all phrases assigned B1 level were extracted from the overall list of phrases on the EVP website, which gave the authors 332 candidates from which to choose for the investigation. Each one of these phrases was then carefully reviewed, albeit intuitively, in terms of whether the college students we teach use them in speaking or writing, and whether we think their meaning can be easily expressed by the same students in a circumlocutory way. Although this classification was solely based on intuition, we felt confident, as experienced teachers, in our ability to identify those phrases that we rarely find in our students' speech or written production, and which would be relatively difficult for these learners to bypass if they were to express the core meanings and functions of these phrases with their current English ability. On the basis of these criteria the 332 candidate phrases were narrowed down to 66.

In order to further reduce the total number, three major corpora (BNC, TenTen, and SkELL 3.8) were consulted to see the frequencies with which these phrases occur in them (see Table 1). However desirable it might have been to examine all 66 phrases, we wanted to also suppress the risk of negatively impacting performance on the test, as the test-takers would likely become increasingly fatigued by the sheer number of items. Since our ultimate goal with this study, as researchers who are also educators, was to acquire data that will be helpful to learners and practitioners, the most frequent phrases in the corpora were not

considered crucial candidates for this study, on the grounds that learners have a better chance of learning such phrases independently of direct instruction, by virtue of the fact that they will naturally encounter them more often than others. The top six high-frequency phrases on the list were thus excluded from our investigation, and the target phrases for this study were finalized with the remaining 60 phrases.

Table 1. ***The sixty-six candidate B1 English phrases for the present study***

Corpus frequency order	Phrase	Average frequency per million words	Corpus frequency order	Phrase	Average frequency per million words
1	at least	211.7	34	go wrong	7.9
2	you know	205.6	35	would rather	7.8
3	rather than	151.9	36	feel like/as if	7.8
4	I know	140.5	37	I bet (you)	7.0
5	at all	124.4	38	change your mind	6.6
6	used to do/be sth	72.9	39	fall asleep	5.8
7	up to 10, 20, etc.	61.4	40	either way	5.7
8	at the same time	54.6	41	again and again	5.7
9	keep doing sth	53.8	42	be up to sb	5.6
10	so far	43.8	43	it/that depends	5.5
11	after all	41.7	44	get worse	5.0
12	get down/into/off, etc.	41.3	45	things like that	5.0
13	make sb do sth	41.0	46	get sth wrong	5.0
14	not until	38.6	47	(just) in case	4.4
15	as long as	37.2	48	tell sb how/what/when to do sth	3.9
16	at first	35.6	49	as you know	3.9
17	not really	33.5	50	that sort of thing	3.9
18	in time	32.2	51	have sth in common	3.5
19	be supposed to do sth	29.0	52	for fun or for the fun of it	3.3
20	be worth sth/doing sth	27.4	53	feel sorry for	2.8
21	be willing (to do sth)	27.1	54	can't/couldn't help doing sth	2.8
22	take advantage of sth	23.1	55	tired of doing sth	2.8
23	no way	20.9	56	take a break/rest, etc.	2.3
24	get rid of sth	19.4	57	wait a minute	2.0
25	in advance	19.4	58	keep sb waiting	1.6
26	can afford	18.8	59	on purpose	1.6
27	do badly/well	14.2	60	be just about to do sth	1.1
28	ever since	13.5	61	if I were you	.8
29	get to know sb/sth	9.6	62	get cold/ill/late, etc.	.6
30	up to date	8.9	63	miss a chance/opportunity	.6
31	What if ...?	8.5	64	feel bad about sth/doing sth	.5
32	go badly/well, etc.	8.1	65	have been meaning to do sth	.4
33	get/become used to sb/sth/ doing sth	7.9	66	be up to sth	.3

2.3 Measurement instrument

Using the English phrase test designed in the study by Schmitt, Dörnyei, Adolphs and Durow (2004) as a model, a test consisting of two distinct sections, the combined results of which would be analyzed by means of the Rasch model, was developed for this study: a multiple-choice recognition section (henceforth, the R-section) and a fill-in-blanks production section (henceforth, the P-section).[1] As with Schmitt et al's study, the decision was made that all test items be embedded in some kind of meaningful context that would allow test-takers to process the target phrases in as naturalistic a way as possible. To this end, we wrote 11 short dialog or monolog contexts (seven and four each), which respectively incorporated five or six of the target phrases. The identical contexts were used for both sections, and 60 test items, each addressing one of the 60 target phrases, were prepared for each section.

As this study was going to be conducted with university students in Japan, the great majority of whom are estimated to be working towards the B1 level, we consulted the EVP website to ensure that none of the words used in the test would exceed B1 – including the directions for taking the test, the multiple-choice distractors in the R-section, and the synonymous expressions for the phrases in the P-section (see below for details). Limiting the test's vocabulary level to B1 level or below served to maintain the unidimensionality of the test items and the attribute of equal discriminative power – two of the prerequisites for using the Rasch model. This enabled us to assume that the determining factor for whether the participants could figure out the right answer to each item was restricted to their knowledge of the B1 phrase in question, and that the whole test would not be too difficult for them (cf. Shimada 2006).

With regard to the R-section, three distractors were generated for each multiple-choice item, all devised in such a way that they would be as similar in meaning and length to the correct form as possible. It was thus expected that the ability to make the correct choice for an item would be based on knowledge of the phrase in question. This would be the case even if, importantly, participants chose distractors for other items in the same context (as the way in which the story was unfolding should have been clear to them). In other words, local independence – another condition to be met when applying the Rasch model – was secured in this section. Also, a fifth option ("I DON'T KNOW") was prepared for each item, the purpose of which was to minimize wild guesses – yet another condition to be satisfied for using the Rasch model – when participants had little or no clue about the target item in question. Below is an example item from the R-section.

Example:

Learning English is boring and it is also hard work. I **[1]** _____ lists of words.

1. (A) stay repeating
   (B) hold repeating
   (C) remain repeating
   (D) keep repeating
   (E) I DON'T KNOW

   [Answer: (D)]

Moving on to the description of the P-section, the primary motivation for including this component as well as the recognition section was to add a further dimension to our understanding of the participants' depth of knowledge of the target phrases. Inclusion of this section was also meant to further weaken the influence of any wild guesses in the R-section when evaluating the participants' ability measures (see Section 2.5 for scoring details).

1. The purpose of Schmitt et al.'s study was not to examine the difficulty measures of English phrases but to investigate the influence of learner characteristics, such as attitudes toward L2 learning, on the learning of multi-word expressions.

In devising the fill-in-blanks items in the P-section, we followed, although not entirely, the test design in the study by Schmitt et al. (2004), adopting blended elements of cloze and C-test techniques. In our study, each target phrase was first categorized in terms of how many 'keywords' (or lexically strong words) it contains, with lexically weak words deliberately excluded from that counting. For example, the phrase *be worth sth/doing sth* was classified as containing one keyword, and both *be* and *sth/doing sth* were discounted. There were, then, eight one-keyword phrases, 30 two-keyword phrases, 17 three-keyword phrases, four four-keyword phrases, and one five-keyword phrase.

In the case of one-keyword phrases, the keyword was substituted by a blank followed by the word's final letter.[2] It should be noted that we ensured that the length of each blank corresponded to that of the word in question, which constituted a hint for the respondents.[3] So the phrase *be just about to do sth*, for example, would appear as *was just* ____*t to say* (italics are used here but not in the test). Two-keyword phrases were displayed with the first word replaced by a blank with its final letter remaining, and the second word with a complete blank. So *get off* (registered as *get down/into/off, etc.* on the EVP website) appeared as ___*t* ____. Phrases of three keywords such as *get rid of sth* appeared as ___*t* ___*d of*, with the third keyword left untouched. Phrases containing four or five keywords were presented with two words being completely blanked (e.g., *at the* _____ _____ for *at the same time*).

Variations on this treatment were applied in a number of cases. When a word to be blanked with its final letter remaining ended with a plural or third-person singular *-s*, *-ing*, *-ed*, *-l*, *-ll*, *-ly*, *-e* or *-h*, one extra letter was left. Adjustments were also made, elaborated on below, when the blanks for an item were filled in an unanticipated way by one or more of the native speakers participating in the pilot. Also elaborated on below are some particular cases, such as where ___ ____*se* (for *on purpose*) was used instead of __*n* _____.

We wanted to see in this section whether the participants would be able to come up with the targeted phrase for each test item, given the context and the hints, rather than whether they could simply figure out the meaning of that phrase. A gloss of the meaning expressed by the item in question was therefore given to the participants in the right margin. With the provision of synonymous expressions in the P-section, local independence of the items – again, a prerequisite of the Rasch model – was expected to be protected.

An additional note regarding those synonymous expressions is that the parameter of each paraphrase was tailored so that the phrase in question would be neither too obvious nor too much of a riddle for the test-takers, and that creating the alternative wording, which had to adhere to such restrictions as using words up to B1 level, would be feasible on our part. Thus, for instance, the alternative expression prepared for the phrase *make sb do sth*, which was used in the sentence *Don't make me do that again!*, was *I don't want to have to do* (for the underlined words).

Each set of words corresponding to a paraphrase (i.e., each test item in the P-section) was shown in bold font with shaded background, as shown in the following example (with the paraphrase appearing in italics on the right).

Example:

WOMAN:	What do you do if you see a student sleeping in your class?	
MAN:	**It _____ds.** Sometimes I just make a joke.	(*I can't give the same answer in every situation*)

[Answer: It depends]

Draft items were piloted on three native-speaker university teachers of English in Japan.[4] Although the finalized test was administered with the R-section first and the P-section second, the native speakers

2. In Schmitt et al.'s study, the word's initial letter(s) was/were left and the rest of it were blanked.
3. No such adjustment in length was made in Schmitt et al.
4. One of them is no longer in Japan.

were, for this piloting, asked to complete the P-section before moving to the R-section. The main reason for doing so was that we wanted them to think about the blanks without the possibility of any inhibitory priming effects arising from their taking the R-section first: we assumed that they might come up with some words that we did not anticipate, and that they might even be unable to fill in some of the blanks. There were in fact some such cases, and there was also one item where one of the native speakers failed to fill in the blanks; the number of items in the P-section for which the native speakers performed unexpectedly was three, six, and seven respectively. There were also two cases in which the native speaker answers made us aware of multiple possible answers. In the case of the R-section, one of the native speakers made all the choices that we expected, while the two others made what we deemed to be mistakes with one and two items respectively. There were a further seven test items for which we discovered one of the distractors to also be an acceptable choice.

The pilot test thus led to revisions to each of the test items in question. The finalized recognition and production sections of the test are available in Appendices A and B, along with a list of the notes regarding the cases where spelling hints in the P-section were prepared in specific ways in Appendix C.

2.4 Procedure

The finalized test was administered in the following manner. First, all instructions were read to participants in Japanese in order to ensure that there would be no misunderstanding about the English instructions provided in the test booklets. A copy of the R-section booklet was then given to each participant. The participants were instructed to take as much time as they wished on this section. Upon completing the R-section, each student submitted their completed R-section booklet, received a copy of the P-section booklet, and then spent as much time as they wished on this second section, after which the test booklet was submitted.[5]

It should be kept in mind that the recognition section was conducted before the production section, which is the reverse of how the native speakers took the test in the pilot stage.[6] The results reported on below would thus have been different, quite possibly to a large extent, had the P-section been administered first. It can be assumed that items in the P-section would have been far more difficult to answer correctly without the learners' residual memory of the R-section. Indeed, it seems very likely that the main reason that the three native speakers were unable to fill in some blanks in the ways we anticipated is that they were asked to work on the P-section first. The R-section, on the other hand, would have been much less difficult had it followed the P-section. Either way, the memory trace from the section implemented initially would affect the test-takers' performance on the second section, and the judgment made for this study was that more informative data would be obtained by giving the R-section first.

It should also be noted about the procedure that we divided the participants into two groups of approximately the same number. Each group was given the contexts in reverse order to the other group. The first context for one group, entitled 'Learning English', thus appeared last for the other group, and so on. This reversal was adopted in order to average out the effect of cognitive fatigue on performance, as the total number of test items was quite large (60 x 2 = 120) although the same 11 contexts were repeated in the latter half of the test.

2.5 Scoring

For each P-section item for which the blanks were filled in correctly, 1 point was allotted, whereas 2 points were given for each correctly chosen item in the R-section.[7] This scoring method was adopted

5. It took about twenty minutes for the fastest participants to finish the entire test, and about fifty minutes for the slowest ones.
6. In Schmitt et al. (2004), the production test was administered before the recognition test, too.
7. In the scoring for the P-section, intelligible spelling errors were not penalized.

on the assumption, explained above, that the P-section must have been less difficult than it would have been had the R-section not been implemented first. That is, the participants' memory trace from taking the R-section can be assumed to have helped them fill in the blanks in the P-section, and thus there was a need to compensate for this priming effect on scoring.[8]

A further adjustment was that for each item that was answered correctly in both sections, 4 points were given instead of 3 points (i.e., the sum of 2 and 1) – the full score for this test was therefore 240 (4 x 60). The extra point was allotted largely as a way to diminish the effect of wild guesses on scoring. Wild guessing is a persistent issue in multiple choice tests that test givers have strived to eliminate, but one that is yet to be adequately addressed (Choi 1992); even though the participants in this study were encouraged to choose the option "I DON'T KNOW" when uncertain about an item in the R-section, there was no guarantee that they actually did this. Because the Rasch model presupposes a minimum level of wild guessing (which will also affect the degree to which equal discriminative power among the test items will be established), a participant in this study getting an item right in both sections was interpreted as evidence of not having guessed wildly for the phrase in question, and a bonus point was therefore justified. A further case for this bonus point can be made if one regards the ability to both correctly produce and recognize a phrase as evidence that the test-taker's knowledge of it goes beyond mere recognition level.

3 Results and discussion

3.1 Results of the phrase test

Table 2 summarizes the statistics for the measured persons and items, derived from Rasch measurement using the software *Winsteps*. Starting with the measured persons, the table only summarizes the data for the 301 fit participants, whose infit mean-square (MnSq)[9] figures range within .75 and 1.30 (cf. Bond and Fox 2007; McNamara 1996); as mentioned above, the data for 59 participants were counted out as misfits. The average θ (person ability measure) of the fit participants is .21, the standard deviation (SD) is .33, and the reliability of their estimated θ is sufficiently high (α = .91). Turning to the measured items, the average δ (item difficulty measure) of the items is .00, the SD is .35, and the reliability of the items' estimated δ is .98. With the average score of the participants being 133.6 out of 240 (about 56%) and this high level of reliability, this phrase test should be regarded as a highly reliable measurement instrument for the target B1 phrases in this study (although, as will be seen below, there were two misfit items). That is, similar item difficulty measures will be derived if the same test is administered to other learners, especially learners whose English proficiency level is about the same as or not too far off from that of the fit participants in this study.

Figure 1, the distribution map of the diverse ability measures of the 301 fit participants and the varying difficulty measures among the 60 B1 phrases, illustrates one possibility to consider about the level of the phrases. While the difficulty measures for all items falling within a range of -1 to 1 should be interpreted as a corroboration of the EVP's CEFR level assignments of the phrases (cf. Negishi, Tono and Fujita, 2012), the gap between the most and least difficult phrases is arguably wide. The variances may thus suggest that certain phrases should perhaps be assigned two successive CEFR levels as opposed to a single distinct level. The variances could also suggest, if one wishes to adhere to a 'one level for one phrase' categorization, the validity of adopting the CEFR's newly-proposed 11 levels (Council of Europe 2018) in place of its previous six levels.[10] This is a tentative proposal, however, because the current study focused solely on the phrases on the B1 list, instead of including A2 and B2 phrases in the test.

8. Schmitt et al. (2004) argue that priming effects were minimized in their study, as they sandwiched three other language tests and a questionnaire between their productive and receptive phrase tests.
9. See: https://www.winsteps.com/winman/misfitdiagnosis.htm
10. So the least difficult phrases investigated in this study such as *if I were you* and *get to know sb/th* (see Table 3 below) can be labeled as A2-B1 (or more simply 'A2+'), and the most difficult ones such as *either way* and *for fun or for the fun of it* (also see Table 3) as B1-B2 (or 'B1+').

Table 2. ***Test statistics***

```
     SUMMARY OF 301 MEASURED Persons
-------------------------------------------------------------------------------
|          TOTAL                          MODEL          INFIT        OUTFIT    |
|          SCORE     COUNT     MEASURE    S.E.      MNSQ   ZSTD   MNSQ   ZSTD |
|-----------------------------------------------------------------------------|
| MEAN     133.6      60.0         .21     .09      1.00     .0   1.01     .1 |
| P.SD      37.5        .0         .33     .01       .11     .8    .14     .7 |
| S.SD      37.5        .0         .33     .01       .11     .8    .14     .7 |
| MAX.     231.0      60.0        1.47     .21      1.30    1.9   1.30    1.8 |
| MIN.      31.0      60.0        -.92     .09       .77   -1.8    .75   -1.7 |
|-----------------------------------------------------------------------------|
| REAL RMSE    .10 TRUE SD     .32  SEPARATION  3.26  Person RELIABILITY  .91 |
|MODEL RMSE    .10 TRUE SD     .32  SEPARATION  3.35  Person RELIABILITY  .92 |
| S.E. OF Persons MEAN = .02                                                  |
-------------------------------------------------------------------------------
Persons RAW SCORE-TO-MEASURE CORRELATION = .99
CRONBACH ALPHA (KR-20) Persons RAW SCORE "TEST" RELIABILITY = .92  SEM = 10.86

     SUMMARY OF 60 MEASURED Items
-------------------------------------------------------------------------------
|          TOTAL                          MODEL          INFIT        OUTFIT    |
|          SCORE     COUNT     MEASURE    S.E.      MNSQ   ZSTD   MNSQ   ZSTD |
|-----------------------------------------------------------------------------|
| MEAN     670.4     301.0         .00     .04      1.00     .0   1.01     .1 |
| P.SD     207.5        .0         .35     .00       .11    1.8    .15    1.8 |
| S.SD     209.2        .0         .35     .00       .12    1.9    .15    1.8 |
| MAX.    1097.0     301.0         .66     .06      1.46    7.3   1.65    8.0 |
| MIN.     294.0     301.0        -.89     .04       .71   -5.4    .70   -4.5 |
|-----------------------------------------------------------------------------|
| REAL RMSE    .04 TRUE SD     .35  SEPARATION  8.06  Items  RELIABILITY  .98 |
|MODEL RMSE    .04 TRUE SD     .35  SEPARATION  8.22  Items  RELIABILITY  .99 |
| S.E. OF Items MEAN = .05                                                    |
-------------------------------------------------------------------------------
Items RAW SCORE-TO-MEASURE CORRELATION = -1.00
Global statistics: please see Table 44.
UMEAN=.0000 USCALE=1.0000
```

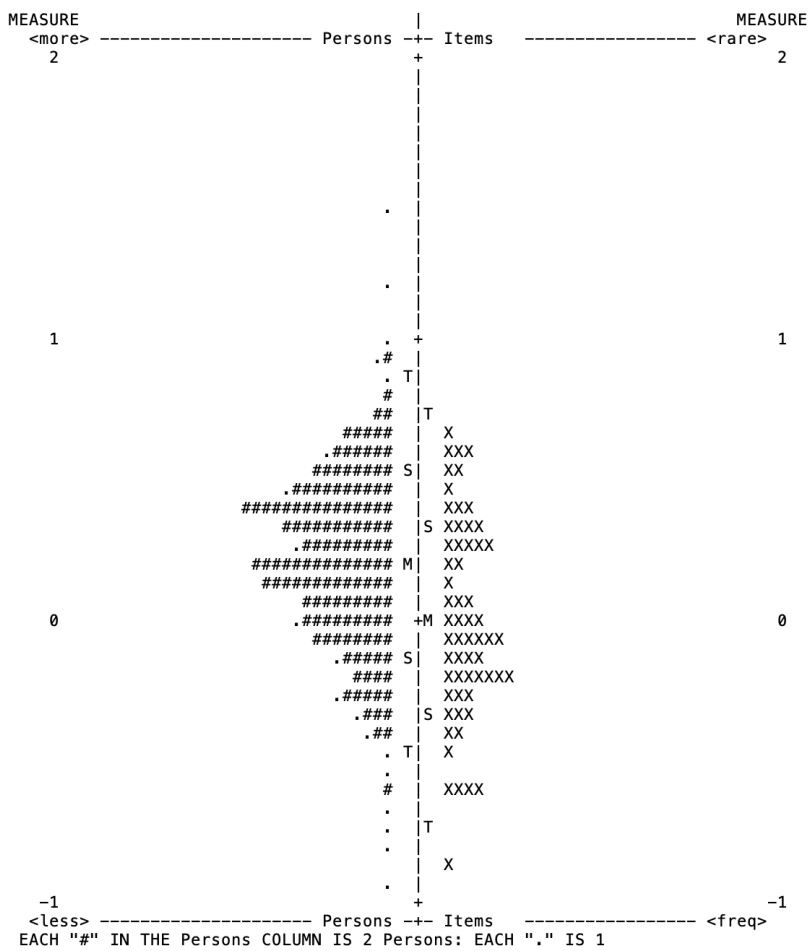

Figure 1. The person by item distribution map of the test.

Finally, the statistics for the individual items (Table 3) reveal the difficulty measures for the top 27 phrases (from *either way* to *feel bad about sth/doing sth*) as over 0.00. It is probably safe then to argue that learners whose proficiency level is either already at or approaching B1 level would benefit by focusing more on these phrases than the remaining 33 phrases. For learners stagnating at A2 (or even A1) level, the opposite may be the case, although of those 33 phrases, the test items created in this study for *keep sb waiting* (entry no. 41) and *in advance* (entry no. 8) were calculated by *Winsteps* as misfit items.

Table 3. ***Test item statistics***

Items STATISTICS: MEASURE ORDER

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PTMEASUR-AL CORR.	PTMEASUR-AL EXP.	EXACT MATCH OBS%	EXACT MATCH EXP%	Items
44	294	301	.66	.05	.94	-.6	1.00	.0	.39	.35	30.9	34.0	either way
22	313	301	.62	.05	.94	-.6	1.06	.5	.23	.36	40.9	32.9	for fun or for the fun of it
54	313	301	.62	.05	1.11	1.2	1.22	1.9	.29	.36	18.3	32.9	be just about to do sth
10	318	301	.60	.05	.89	-1.3	1.06	.6	.28	.36	45.8	32.6	ever since
34	341	301	.56	.05	1.16	1.8	1.11	1.0	.35	.37	28.9	31.0	go badly/well, etc.
53	341	301	.56	.05	.91	-1.1	.98	-.2	.30	.37	39.2	31.0	have been meaning to do sth
49	392	301	.45	.04	1.15	2.0	1.21	2.1	.38	.38	16.3	27.2	do badly/well
56	404	301	.43	.04	1.19	2.6	1.39	3.8	.10	.39	27.2	26.7	be up to sth
57	422	301	.40	.04	.98	-.3	1.01	.2	.38	.39	19.3	25.8	that sort of thing
40	428	301	.39	.04	1.02	.3	1.03	.4	.36	.39	16.9	25.6	not until
24	442	301	.37	.04	.88	-1.8	.97	-.4	.38	.40	31.9	24.7	be worth sth/doing sth
37	456	301	.34	.04	1.03	.5	.99	.0	.45	.40	16.3	24.5	go wrong
20	469	301	.32	.04	.95	-.8	.91	-1.1	.52	.40	13.6	23.7	(just) in case
30	478	301	.30	.04	.85	-2.5	.84	-2.1	.54	.41	17.9	23.2	on purpose
60	487	301	.29	.04	1.22	3.3	1.21	2.6	.41	.41	11.0	22.9	would rather
33	488	301	.29	.04	1.03	.4	.98	-.2	.52	.41	11.0	22.9	be supposed to do sth
13	500	301	.27	.04	1.01	.2	1.02	.2	.43	.41	17.3	22.2	so far
35	506	301	.26	.04	.97	-.4	.96	-.5	.48	.41	13.6	22.0	What if ...?
59	520	301	.24	.04	.96	-.7	.94	-.8	.52	.41	15.0	21.0	can afford
5	527	301	.23	.04	1.01	.2	1.01	.2	.34	.42	17.9	20.8	after all
45	563	301	.17	.04	1.12	2.1	1.24	3.2	.22	.42	14.0	19.4	take a break/rest, etc.
47	597	301	.12	.04	.94	-1.0	.94	-.9	.41	.43	16.3	17.1	feel sorry for
29	612	301	.09	.04	1.03	.6	1.10	1.5	.23	.43	15.6	16.5	feel like/as if
32	618	301	.09	.04	.79	-4.1	.78	-3.5	.54	.43	18.3	16.4	get/become used to sb/sth/doing sth
28	624	301	.08	.04	1.02	.4	1.00	.1	.39	.43	11.6	16.4	not really
58	659	301	.02	.04	.98	-.4	.96	-.5	.40	.43	17.3	14.8	get sth wrong
46	663	301	.02	.04	1.01	.3	1.03	.4	.36	.43	12.3	14.8	feel bad about sth/doing sth
51	685	301	-.02	.04	1.15	2.6	1.17	2.4	.39	.43	10.3	14.0	tired of doing sth
7	692	301	-.03	.04	1.09	1.6	1.08	1.1	.42	.43	11.6	13.7	have sth in common
55	697	301	-.03	.04	1.04	.7	1.01	.1	.44	.43	9.6	13.6	can't/couldn't help doing sth
41	702	301	-.04	.04	1.46	7.3	1.65	8.0	.21	.43	9.0	13.4	keep sb waiting
15	722	301	-.07	.04	.98	-.3	1.00	.1	.46	.43	9.0	13.2	as long as
17	729	301	-.08	.04	.97	-.5	.94	-.9	.56	.43	6.0	12.4	it/that depends
50	732	301	-.09	.04	.86	-2.6	.86	-2.1	.52	.43	15.0	12.7	up to date
42	736	301	-.09	.04	.99	-.1	1.00	.0	.41	.43	9.0	12.4	up to 10, 20, etc.
25	766	301	-.14	.04	.83	-3.0	.81	-2.7	.59	.43	13.6	12.1	miss a chance/opportunity
8	774	301	-.15	.04	.71	-5.4	.70	-4.5	.58	.43	19.6	13.3	in advance
43	775	301	-.16	.04	1.08	1.3	1.05	.7	.39	.43	14.3	13.3	get down/into/off, etc.
19	776	301	-.16	.04	1.14	2.3	1.15	1.9	.37	.43	12.3	13.2	tell sb how/what/when to do sth
27	803	301	-.20	.04	.96	-.6	.97	-.3	.41	.43	13.3	13.3	at the same time
48	804	301	-.20	.04	.94	-1.0	.98	-.2	.47	.43	17.9	13.2	take advantage of sth
9	805	301	-.20	.04	1.00	.1	1.02	.3	.37	.43	12.3	14.0	at first
52	814	301	-.22	.04	.87	-2.3	.87	-1.6	.43	.43	19.6	13.8	get worse
4	816	301	-.22	.04	1.09	1.4	1.09	1.1	.38	.42	12.6	13.8	be willing (to do sth)
14	817	301	-.22	.04	1.01	.3	1.02	.3	.44	.42	14.3	14.0	no way
31	817	301	-.22	.04	1.09	1.4	1.17	2.0	.43	.42	10.6	14.0	get rid of sth
23	827	301	-.24	.04	.86	-2.3	.89	-1.3	.48	.42	16.6	14.2	things like that
38	843	301	-.27	.04	1.09	1.4	1.06	.7	.39	.42	16.9	15.5	make sb do sth
39	852	301	-.28	.04	.94	-.9	1.00	.1	.48	.42	17.6	16.9	in time
3	865	301	-.31	.04	.97	-.5	.90	-1.1	.50	.42	17.6	17.1	be up to sb
36	867	301	-.31	.04	.95	-.7	.96	-.4	.38	.42	21.3	18.2	wait a minute
16	873	301	-.32	.04	.92	-1.2	.90	-1.1	.47	.42	23.9	18.3	I bet (you)
26	905	301	-.38	.04	1.00	.1	.93	-.6	.44	.41	23.3	21.6	change your mind
1	920	301	-.41	.04	1.04	.5	1.08	.7	.33	.40	23.3	25.7	keep doing sth
11	947	301	-.46	.05	1.10	1.2	1.17	1.4	.34	.40	30.9	32.3	get cold/ill/late, etc.
2	994	301	-.57	.05	.92	-.9	.92	-.5	.42	.38	43.9	41.6	again and again
18	996	301	-.57	.05	1.08	.8	.93	-.5	.46	.38	44.5	41.6	fall asleep
12	997	301	-.58	.05	.95	-.5	.92	-.5	.44	.38	43.9	42.4	as you know
6	1006	301	-.60	.05	.95	-.4	.88	-.8	.37	.37	44.5	43.6	get to know sb/sth
21	1097	301	-.89	.06	.91	-.6	.71	-1.4	.41	.31	74.8	73.8	if I were you
MEAN	670.4	301.0	.00	.04	1.00	.0	1.01	.1			20.6	22.0	
P.SD	207.5	.0	.35	.00	.11	1.8	.15	1.8			12.4	10.8	

3.2 Implications for learning and teaching

Several further implications for learning and teaching emerge from the participants' performance on each section of the test. At this point, we only consider the relatively difficult phrases. Table 4 is a detailed description of the participants' performance on the 27 phrases rated as having a difficulty measure above 0.00. Phrases for which the accuracy rate is low in both sections (such as *either way*, *be just about to do sth*, *go badly/well, etc.*, *do badly/well*, *would rather*) are probably the biggest challenges to learners who are approximating to or currently at the B1 level, which suggests that they deserve more focused study and instruction than other phrases.

In the case of those phrases for which a relatively large percentage of the participants were able to choose the correct forms in the R-section but unable to successfully fill in the blanks in the P-section (such as *that sort of thing*, *(just) in case*, *on purpose*, *feel sorry for*, *get/become used to sb/sth/doing sth*), output training alone could be very effective for A2-B1 learners.

Where a large proportion of the participants performed correctly only in the P-section (such as *for fun or for the fun of it*, *ever since*, *have been meaning to do sth*, *be worth sth/doing sth*, *feel like/as if*), college learners are likely to be able to recognize the phrases (or their constituent words) without yet being sufficiently familiar with their collocational attributes. Such linguistic features might be best handled by explicit instruction that draws learners' attention to them.

The results as a whole, which reflect a wide range of performance, appear to us to indicate that the idea, mentioned above, that certain phrases be assigned two successive levels (or different levels using the CEFR's new 11 levels) would be more helpful if complemented by the potentially equally important proposal that distinction between production and recognition be made for level assignments.

Lastly, the participants' performance aside, focused study may be very effective if directed to low-frequency phrases (such as *be just about to do sth*, *have been meaning to do sth*, *be up to sth*, *on purpose*, *feel bad about sth/doing sth*), simply because learners, including those in the Japanese context, appear to have fewer opportunities to encounter and learn them in natural input.[11]

3.3 Caveats in interpreting the data

There are some caveats in interpreting the data gained in this study. To begin with, although measures were taken to minimize the influence of wild guesses, which would help to maintain parity of the discriminative power of the test items, it is possible that better test design and scoring could perhaps have further reduced that influence. In addition, priming effects were inevitable, given the content and procedural structure of the test. Thus, while the overall difficulty measures may have been roughly the same even if the administration of the R-section and the P-section had been reversed, the distribution of the accuracy rates for the two sections would probably have been somewhat different. Another issue is to do with how the blanks in the P-section were constructed: whether the rather complicated criteria governing their design may have affected the participants' performance. Issues with priming effects and the construction of the blanks weakened the integrative quality of the P-section and therefore the overall validity of the test as a tool to measure the actual difficulties of the target phrases. Last but not least, the range of the average difficulty measures found for the contexts, shown in Table 5, may not be negligible. It is important to note that this was the case even though the contexts were presented to the participants in two orders (see Section 2.4). There thus remains room for doubt as to whether local independence was sufficiently secured in this test.

11. The correlation between the phrases' difficulty measures and their average frequency figures derived from the three corpora consulted (BNC, TenTen, and SkELL 3.8) is virtually non-existent (r = -.072, p = .584), suggesting that a phrase's difficulty has little to do with its frequency.

Table 4. ***The participants' performance on each section of the test***

δ		Item		Distribution of the participants			
Ranking	Measure	Phrase	Average frequency per million words	Correct in both R- & P-sections (M = 44%)	Correct only in R-section (M = 13%)	Correct only in P-section (M = 21%)	Incorrect in both sections (M = 22%)
1	0.66	either way	5.7	11%	15%	25%	50%
2	0.62	for fun or for the fun of it	3.3	10%	9%	45%	36%
	0.62	be just about to do sth	1.1	12%	23%	12%	53%
4	0.60	ever since	13.5	12%	3%	52%	33%
5	0.56	have been meaning to do sth	.4	13%	6%	48%	33%
	0.56	go badly/well, etc.	8.1	18%	8%	27%	48%
7	0.45	do badly/well	14.2	20%	22%	7%	51%
8	0.43	be up to sth	.3	19%	9%	39%	33%
9	0.40	that sort of thing	3.9	18%	26%	15%	41%
10	0.39	not until	38.6	20%	19%	23%	38%
11	0.37	be worth sth/ doing sth	27.4	22%	2%	53%	22%
12	0.34	go wrong	7.9	26%	8%	31%	35%
13	0.32	(just) in case	4.4	25%	23%	10%	42%
14	0.30	on purpose	1.6	23%	31%	4%	42%
15	0.29	be supposed to do sth	29.0	30%	11%	19%	40%
	0.29	would rather	7.8	31%	17%	6%	47%
17	0.27	so far	43.8	30%	5%	37%	29%
18	0.26	What if ...?	8.5	29%	13%	25%	33%
19	0.24	can afford	18.8	30%	21%	11%	38%
20	0.23	after all	41.7	29%	13%	34%	24%
21	0.17	take a break/ rest, etc.	2.3	33%	7%	44%	17%
22	0.12	feel sorry for	2.8	32%	27%	16%	25%
23	0.09	feel like/as if	7.8	36%	1%	58%	5%
	0.09	get/become used to sb/sth/ doing sth	7.9	34%	28%	13%	25%
25	0.08	not really	33.5	39%	7%	36%	18%
26	0.02	get sth wrong	5.0	39%	22%	18%	21%
	0.02	feel bad about sth/doing sth	.5	42%	9%	36%	14%

Note. Average frequencies per million words were derived from BNC, TenTen, and SkELL 3.8.

Table 5. ***Context-by-context average difficulty measures***

Context no.	Item no.	Theme of the context	Average δ
1	1-5	Learning English	-0.26
2	6-10	Romance	-0.08
3	11-16	Health	-0.23
4	17-21	Teaching	-0.28
5	22-26	Travel	0.05
6	27-32	Watching TV	0.02
7	33-38	Asking for help, but not in a direct way	0.15
8	39-44	Late for the test	0.08
9	45-50	Absent from school	0.08
10	51-55	A difficult relationship	0.18
11	56-60	Parents and marriage	0.28

4 Conclusion

This study has examined the variations in difficulty encountered by university learners of English in Japan with regard to English phrases that the EVP classifies as CEFR B1 level. It has demonstrated that while the B1 level indeed seems to be valid in a broad sense, Japanese university learners do also seem to encounter within this level considerable variations in difficulty. It can be presumed that this range of difficulty also applies to other phrases that were not investigated in this study. The most important assertion that can be confidently made on the basis of the data and analysis presented here is that the inherent difficulty of a phrase differs depending on whether the mode of language processing is production or recognition. The results thus suggest that while it may be seemingly helpful to assign a single level to a phrase, it may be more realistic and ultimately helpful to take account of the processing mode. More detailed accounts of the global difficulty of a phrase, as suggested in this paper, may well be helpful to learners and teachers, although defining global difficulty is a tremendously challenging task.

This study points to a number of future directions. The test developed for this study seems, despite its inherent limitations, worthy of replication to see if its findings are validated with learners of English in different contexts who have non-Japanese L1 backgrounds. Insights gained from the present study can also help in the design of new tests for measuring difficulties of other English phrases. Whatever measurement tool is developed, a phrase ought to be examined from at least the two aspects of recognition and production. The way this notion was handled in this study can easily be extended to listening. While applying it to speaking will be more challenging, it is certainly not out of the question to do so. Also, scientifically more rigorous data could be acquired if, say, a two-section test were to use two different sets of contexts so as to minimize priming effects. Testing involving a very large number of participants would help to resolve the issues associated with wild guessing and equal discriminative power, as such testing would allow for the adoption of a three-parameter logistic model.

In conclusion, it is hoped that, all caveats and limitations considered, this study's findings can lead in the direction of more information for L2 learners of English at the tertiary level within and outside of Japan, for classroom practitioners teaching them and for material developers. It is also hoped that the study has shed light on the complexities inherent in the comprehension and production of multi-word expressions in English, and that future research projects may benefit from these insights.

5 Acknowledgements

This work is supported by JSPS KAKENHI Grant-in-Aid (Number: 16K02855). The authors wish to thank the two anonymous reviewers for comments on the earlier version of this paper.

6 References

Bond, Trevor & Christine Fox. 2015. *Applying the Rasch model: Fundamental measurement in the human sciences* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Choi, Inn-Chull. 1992. *An application of item response theory to language testing.* New York: Peter Lang.

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

Council of Europe. 2018. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment. Companion Volume with New Descriptors.* Strasbourg: Council of Europe.

Erman, Britt & Beatrice Warren. 2000. The idiom principle and the open-choice principle. *Text, 20,* 29-62.

Harrison, Julia & Fiona Barker. 2015. *English Profile in Practice,* English Profile Studies volume 5. Cambridge: Cambridge University Press.

McNamara, Tim. 1996. *Measuring second language performance.* London: Longman.

Millar, Neil. 2010. The processing of malformed formulaic language. *Applied Linguistics, 32*(2), 1-21.

Negishi, Masashi. 2012. The development of the CEFR-J: Where we are, where we are going. In N. Tomimori, M. Furihata, K. Haida, N. Kurosawa, & M. Negishi (Eds.), *New Perspectives for Foreign Language Teaching in Higher Education: Exploring the Possibilities of Application of CEFR*, 105-116. Tokyo: WOLSEC, Tokyo University of Foreign Studies.

Negishi, Masashi, Yukio Tono & Yoshihito Fujita. 2012. A validation study of the CEFR levels of phrasal verbs in the English Vocabulary Profile. *English Profile Journal, 3*(1), 1-16.

Shimada, Takuji. 2006. On IRT-based language testing: Aren't there any blind spots? *Foreign language education: Theory and practice, 32,* 29-42.

Schmitt, Norbert, Zoltán Dörnyei, Svenja Adolphs & Valerie Durow. 2004. Knowledge and acquisition of formulaic sequences: A longitudinal study. In N. Schmitt (Ed.), *Formulaic Sequences: Acquisition, Processing, and Use*, 55-86. Amsterdam: John Benjamins Publishing Company.

Stengers, Helene, Frank Boers, Alex Housen & June Eyckmans. 2011. Formulaic sequences and L2 oral proficiency: Does the type of target language influence the association? *IRAL, 49*(4), 321-343.

7 Biographies

Takeshi Matsuzaki is an Associate Professor at Bunkyo Gakuin University, Japan. He has taught English at the university level in Japan for over 15 years. He holds an MS in TESOL from University of Pennsylvania and a PhD in Humanities from Tokyo University of Foreign Studies. His research interests include the learning and teaching of multi-word expressions in foreign language contexts and the complex and dynamic nature of second language learning.

Kevin Mark is a Professor at Meiji University, where he has been working since 1991. He argues that, for language teaching research and theory to be relevant, the specialized strands of applied linguistics need to be integrated by broad humanistic and educational principles. His innovative work in autonomy, global education, curriculum, CALL, materials writing and learner corpus development as a part of teaching addresses this apparently paradoxical and overarching question: How can mass education be made a more congenial and human process for teachers and learners alike?

Appendix A
B1 phrase recognition section of the test

Vocabulary Phrases (multiple choices)

NAME _____

Directions:

Each of the following pieces of language is spoken by one or two speakers. Each one contains five or six missing phrases. Choose from (A), (B), (C) or (D) and circle the letter for the phrase which fits best. If you are not sure, circle (E) for "I don't know."

Example

I'm **[1]** ___ a team of twenty people.

[1] (A) responsible of
(B) in responsibility for
(C) the charge of
(D) in charge of
(E) I DON'T KNOW

When you are told to, go on to the next page and start taking the test.

Learning English

In the following conversation between two friends, the man complains about how boring and hard learning English is, and the woman gives him some advice.

MAN:	Learning English is boring and it is also hard work. I **[1]** ___ lists of words. Every week I'm doing the same things, **[2]** ___.
WOMAN:	It doesn't have to be like that.
MAN:	What do you mean?
WOMAN:	Whether it's interesting to you or not **[3]** ___. There are actually many interesting ways to study. You only need to **[4]** ___ look for them. I'm sure you will discover that learning English does not have to be boring **[5]** ___.

[1] (A) stay repeating
(B) hold repeating
(C) remain repeating
(D) keep repeating
(E) I DON'T KNOW

[2] (A) in a repeating way
(B) in a frequent way
(C) again and again
(D) again and over
(E) I DON'T KNOW

[3] (A) is after you
(B) is based on you
(C) is on you
(D) is up to you
(E) I DON'T KNOW

[4] (A) be wanting to
(B) be meaning to
(C) be intending to
(D) be willing to
(E) I DON'T KNOW

[5] (A) at the end
(B) all over
(C) at last
(D) after all
(E) I DON'T KNOW

Romance

In the following conversation between two friends, the woman asks the man about how his relationship with his girlfriend started.

WOMAN:	How did you and your girlfriend **[6]** ___ each other?
MAN:	We met through a friend of ours. She told us we **[7]** ___.
WOMAN:	So you knew **[8]** ___ that you would like each other?
MAN:	No, but we hoped we would.
WOMAN:	How did you feel when you met her?
MAN:	**[9]** ___ I was a little embarrassed. But we very quickly started to feel comfortable with each other.
WOMAN:	And you've been together **[10]** ___?
MAN:	That's right!

[6]
(A) get to know
(B) become knowing
(C) grow into knowing
(D) turn into knowing
(E) I DON'T KNOW

[7]
(A) shared many things
(B) held many things in common
(C) had a lot in common
(D) got lots of common things
(E) I DON'T KNOW

[8]
(A) in advance
(B) before advance
(C) at the advance
(D) for advance
(E) I DON'T KNOW

[9]
(A) In the first time
(B) To begin with it
(C) First of all
(D) At first
(E) I DON'T KNOW

[10]
(A) ever after
(B) after that
(C) since that
(D) ever since
(E) I DON'T KNOW

Health

In the following conversation between a couple, the woman gives the man some advice about how he can lose weight.

MAN:	Do you think I'm **[11]** ___?
WOMAN:	Well, **[12]** ___, a man of your height should be under 70 kilos. **[13]** ___ you've been able to reach 80, right? Let me ask you something. Don't you think you are still eating too much pasta?
MAN:	There **[14]** ___ I am going to eat less pasta. I love it!
WOMAN:	Well, **[15]** ___ you eat so much, **[16]** ___ you won't be able to lose those 10 kilos.

[11]
(A) going thinner
(B) getting thinner
(C) turning thinner
(D) being thinner
(E) I DON'T KNOW

[12]
(A) since you know
(B) for you to know
(C) since it's known
(D) as you know
(E) I DON'T KNOW

[13]
(A) For now
(B) So far
(C) Before now
(D) For the past
(E) I DON'T KNOW

[14]
(A) is no way
(B) are no possibilities
(C) is not the possibility
(D) is not a way
(E) I DON'T KNOW

[15]
(A) as far as
(B) so far as
(C) as long as
(D) as if
(E) I DON'T KNOW

[16]
(A) you bet
(B) the bet is
(C) I bet
(D) it's a bet
(E) I DON'T KNOW

Teaching

In the following conversation between two teachers and colleagues, the woman asks the man how he deals with students sleeping in his class.

WOMAN:	What do you do if you see a student sleeping in your class?
MAN:	**[17]** ___. Sometimes I just make a joke. Once, when I was giving a private lesson, the girl I was teaching **[18]** ___ in front of me.
WOMAN:	What did you do?
MAN:	Well, I wondered for a few minutes, but then she woke up. So, I **[19]** ___ before coming to the next lesson.
WOMAN:	What did you tell her?
MAN:	I told her that she should drink at least three cups of coffee. She smiled, thanked me and said she would, **[20]** ___.
WOMAN:	**[21]** ___, I would have said at least four cups.

[17] (A) The case is different
(B) It depends
(C) Things change
(D) The choices are different
(E) I DON'T KNOW

[18] (A) went sleeping
(B) fell asleep
(C) just slept
(D) was sleepy
(E) I DON'T KNOW

[19] (A) told her what to do
(B) told her what should she do
(C) told her to do what
(D) told her what she does
(E) I DON'T KNOW

[20] (A) just in the case
(B) just in a case
(C) just in cases
(D) just in case
(E) I DON'T KNOW

[21] (A) If you were me
(B) If I am you
(C) If I could be you
(D) If I were you
(E) I DON'T KNOW

Travel

In the following conversation between a couple, the man wants the woman to go with him on a very cheap tour to Hawaii.

MAN:	I just heard about a really cheap three-night tour to Hawaii. The flight leaves on Friday afternoon and gets back on Monday at lunchtime.
WOMAN:	That's tomorrow afternoon. Are you crazy?
MAN:	It only costs 50,000 yen for everything. I think we should go, just **[22]** ___. We could go swimming and shopping and do **[23]** ___.
WOMAN:	You're right. It **[24]** ___ the price. It would be crazy to **[25]** ___ when it's so cheap.
MAN:	I'm glad you've **[26]** ___!

[22] (A) for the fun of it
(B) for the fun
(C) for its fun
(D) of the fun for it
(E) I DON'T KNOW

[23] (A) things like that
(B) a thing like that
(C) the stuff like that
(D) stuff such as that
(E) I DON'T KNOW

[24] (A) is worth
(B) is fair by
(C) is valuable for
(D) is right for
(E) I DON'T KNOW

[25] (A) fail an opportunity to go
(B) escape opportunities to go
(C) miss a chance to go
(D) give away chances to go
(E) I DON'T KNOW

[26] (A) decided to change minds
(B) changed your mind
(C) changed your decision
(D) decided to change
(E) I DON'T KNOW

Watching TV

In the following conversation, the mother wants her son to study harder.

MOTHER:	I told you not to watch TV and do your homework **[27]** ___. I know that you aren't doing a good job with your homework.
SON:	I am doing a good job!
MOTHER:	Most of your attention is going to the TV, isn't it?
SON:	**[28]** ___. Only a little.
MOTHER:	**[29]** ___ you're doing this just to annoy me. You're doing it **[30]** ___.
SON:	No, I'm not!
MOTHER:	I'm going to **[31]** ___ this TV tomorrow.
SON:	Don't do that!
MOTHER:	You'll **[32]** ___ it.

[27]
(A) in the same time
(B) at the same time
(C) in the same moment
(D) at the same moment
(E) I DON'T KNOW

[28]
(A) Not really
(B) Not much so
(C) Not really so
(D) Not very
(E) I DON'T KNOW

[29]
(A) I'm likely to think
(B) I feel I think
(C) I think it's likely
(D) I feel like
(E) I DON'T KNOW

[30]
(A) on purpose
(B) on your purpose
(C) for its purpose
(D) for the purpose of it
(E) I DON'T KNOW

[31]
(A) take off
(B) get rid of
(C) put off
(D) hand out
(E) I DON'T KNOW

[32]
(A) become used for
(B) become used with
(C) get used for
(D) get used to
(E) I DON'T KNOW

Asking for help, but not in a direct way

In the following conversation, the older sister wants her younger brother to help her with her smartphone.

SISTER:	Why are you watching TV? I thought **[33]** ___ doing your English homework.
BROTHER:	Don't bother me!
SISTER:	Did your English test **[34]** ___ today?
BROTHER:	No!
SISTER:	**[35]** ___ you fail again? Maybe you need some help. Shall I help you?
BROTHER:	**[36]** ___. What are you trying to do here? You're not really interested in my English, are you? Has something **[37]** ___ with your smartphone again? I spent an hour helping you with it yesterday. Don't **[38]** ___ that again!

[33]
(A) there was a need for you to be
(B) you had the need for
(C) you were supposed to be
(D) it was the importance of your
(E) I DON'T KNOW

[34]
(A) go well
(B) do well
(C) get well
(D) become well
(E) I DON'T KNOW

[35]
(A) What's possible if
(B) How possible is it that
(C) How can it be if
(D) What if
(E) I DON'T KNOW

[36] (A) Wait the second
(B) Wait a minute
(C) Hold the second
(D) Hold a minute
(E) I DON'T KNOW

[37] (A) gone wrong
(B) got wrong
(C) come bad
(D) become bad
(E) I DON'T KNOW

[38] (A) make me do
(B) get me
(C) want me for
(D) request me for
(E) I DON'T KNOW

Late for the test

In the following, the speaker is upset because she doesn't think she can get to school before her test starts.

With this train delay I'm not going to get to school **[39]** ____ for the test, even though it **[40]** ____ . I just sent an email to the professor, telling him what's happened. He answered that he can **[41]** ___ for a little while, but only **[42]** ___ ten minutes. It might be quicker if I **[43]** ___ the train and take a bus. But **[44]** ___, I'm going to be too late for the test.

[39] (A) in the time
(B) in time
(C) within the time
(D) within time
(E) I DON'T KNOW

[40] (A) is going to start not by ten
(B) is starting not until ten
(C) isn't starting by ten
(D) isn't going to start until ten
(E) I DON'T KNOW

[41] (A) keep everyone waiting
(B) keep everyone waited
(C) get everyone waiting
(D) get everyone waited
(E) I DON'T KNOW

[42] (A) before
(B) to
(C) up until
(D) up to
(E) I DON'T KNOW

[43] (A) take out
(B) take off
(C) get out
(D) get off
(E) I DON'T KNOW

[44] (A) in both ways
(B) either way
(C) each way
(D) in each way
(E) I DON'T KNOW

Absent from school

In the following, the speaker had an accident and has not gone to school for a while. His classmates have been really kind to him. He is feeling guilty, because he has actually been enjoying himself.

I've had to **[45]** ___ from school for the past three weeks, because of my accident. I **[46]** ___ about the fact that everyone at school has been **[47]** ___ me, because actually I've been enjoying myself. I've been able to **[48]** ___ the break to do a lot of studying. I'm sure that I am going to **[49]** ___ in my classes when I go back to school. I will be **[50]** ___, even though I have been absent for three weeks.

[45] (A) be at home
(B) be on vacation
(C) have time resting
(D) take a break
(E) I DON'T KNOW

[46] (A) have the wrong feeling
(B) have the bad feeling
(C) feel wrong
(D) feel bad
(E) I DON'T KNOW

[47] (A) having a sorry feeling for
(B) being sorry to
(C) feeling sorry for
(D) being sorry thinking about
(E) I DON'T KNOW

[48] (A) take advantage of
(B) make advantage of
(C) take advantage with
(D) make advantage with
(E) I DON'T KNOW

[49] (A) get well
(B) do well
(C) go well
(D) be well
(E) I DON'T KNOW

[50] (A) to date
(B) to the date
(C) up to date
(D) up to the date
(E) I DON'T KNOW

A difficult relationship

In the following, the speaker talks about a difficult classmate. Because of this classmate, she has been having a hard time.

I have a classmate who is always saying negative things about other people. I'm really **[51]** ___ to her. I'm afraid she is **[52]** ___, and I **[53]** ___ say something to her about it. The other day **[54]** ___ something to her when she asked me for advice about a problem she was having. She said she **[55]** ___ saying negative things about other people.

[51] (A) tired to listen
(B) tired of listening
(C) tiring listening
(D) tiring to listen
(E) I DON'T KNOW

[52] (A) getting worse
(B) going worse
(C) getting bad
(D) going bad
(E) I DON'T KNOW

[53] (A) mean to have had to
(B) have had the meaning to
(C) have meant that I
(D) have been meaning to
(E) I DON'T KNOW

[54] (A) it was about the time to say
(B) it was the time for saying
(C) I was just close to saying
(D) I was just about to say
(E) I DON'T KNOW

[55] (A) couldn't stop from
(B) wasn't able to stop from
(C) couldn't help
(D) didn't help
(E) I DON'T KNOW

Parents and marriage

In the following, the speaker talks about his parents. He is worried about how they will accept the news about his marriage.

My Mom and Dad want to know what **[56]** ___ each week. Usually we talk about my part-time job, my friends, my studies and **[57]** ___. They usually think I make decisions without enough thought, which isn't true. So I want to be sure that they won't **[58]** ___ it wrong this time. I am going to tell them that I am getting married next week! I am sure that they will tell me to wait until after I finish my college education. But she comes from a very rich family, so they don't need to worry that **[59]** ___ this. And they **[60]** ___ have a rich son than a poor one!

[56] (A) stuff I have been doing
(B) thoughts have been in me
(C) I have been up to
(D) I have been about
(E) I DON'T KNOW

[57] (A) the thing of that sort
(B) that sort of thing
(C) things that are similar
(D) similar things of that sort
(E) I DON'T KNOW

[58] (A) get
(B) understand
(C) have
(D) hold
(E) I DON'T KNOW

[59] (A) I can't have the money for
(B) my money is not enough for
(C) I can't afford to do
(D) my money cannot do
(E) I DON'T KNOW

[60] (A) will want to
(B) will like to
(C) would like to
(D) would rather
(E) I DON'T KNOW

This is the end of the test. Close this booklet and follow the instructions.

Appendix B
B1 phrase production section of the test

Vocabulary Phrases (filling in blanks)

NAME _____

Directions:

Each of the following pieces of language is spoken by one or two speakers. Each one contains five or six phrases with missing letters or words. You can see the meaning of the phrase on the right side.

Here is an example:

Phrase	Meaning
I'm ___ _____ge of a team of twenty people.	*(I'm responsible for)*

Answer

I'm *in charge* of a team of twenty people.

Shorter or longer lines such as "___" or "_____" give you an idea about how many letters are missing. Try to fill in the missing letters or words. If it is too difficult, just go on quickly to the next phrase with missing letters or words.

When you are told to, go on to the next page and start taking the test.

Learning English

In the following conversation between two friends, the man complains about how boring and hard learning English is, and the woman gives him some advice.

MAN:	Learning English is boring and it is also hard work. I **___p repeating** lists of words. Every week I'm doing the same things **a___ and ___n**.	*(continue to repeat)* *(feeling annoyed about doing the same thing so many times)*
WOMAN:	It doesn't have to be like that.	
MAN:	What do you mean?	
WOMAN:	Whether it's interesting to you or not **is __p __ you**. There are actually many interesting ways to study. You only need to **be wi___g to look for** them. I'm sure you will discover that learning English does not have to be boring **a___r a__**.	*(is something you are responsible for)* *(be positive about looking for)* *(even if you have not thought so until now)*

Romance

In the following conversation between two friends, the woman asks the man about how his relationship with his girlfriend started.

WOMAN:	How did you and your girlfriend **__t to kn__** each other?	*(become familiar with)*
MAN:	We met through a friend of ours. She told us **we had a lot __ _____**.	*(many of our interests and attitudes were the same)*
WOMAN:	So you knew **__ _____ce** that you would like each other?_	*(before meeting each other)*
MAN:	No, but we hoped we would.	
WOMAN:	How did you feel when you met her?	
MAN:	**__t ____t** I was a little embarrassed. But we very quickly started to feel comfortable with each other.	*(In the beginning)*
WOMAN:	And you've been together **e____ s_____**?	*(since then)*
MAN:	That's right!	

Health

In the following conversation between a couple, the woman gives the man some advice about how he can lose weight.

MAN:	Do you think I'm **____ting thinner**?	*(becoming thinner)*
WOMAN:	Well, **___s you ____w**, a man of your height should be under 70 kilos. **S___ ___r** you've been able to reach 80, right? Let me ask you something. Don't you think you are still eating too much pasta?	*(although I know you already realize this)* *(Until now)*
MAN:	**There is** _ _**y** I am going to eat less pasta. I love it!	*(It is impossible)*
WOMAN:	Well, **___s ____g as you eat** so much, **I ___t** you won't be able to lose those 10 kilos.	*(if you continue to eat) (I am sure)*

Teaching

In the following conversation between two teachers and colleagues, the woman asks the man how he deals with students sleeping in his class.

WOMAN:	What do you do if you see a student sleeping in your class?	
MAN:	**It _____ds**. Sometimes I just make a joke. Once, when I was giving a private lesson, the girl I was teaching **__ell _____** in front of me.	*(I can't give the same answer in every situation)* *(started to sleep)*
WOMAN:	What did you do?	
MAN:	Well, I wondered for a few minutes, but then she woke up. So, I just **told her what ___ ___** before the next lesson.	*(told her what she should do)*
WOMAN:	What did you tell her?	
MAN:	I told her that she should drink at least three cups of coffee. She smiled, thanked me and said she would, **____t in ___se**.	*(to make sure that she would not start to sleep during a lesson)*
WOMAN:	**___ I _____ you**, I would have said at least four cups.	*(In your position)*

Travel

In the following conversation between a couple, the man wants the woman to go with him on a very cheap tour to Hawaii.

MAN:	I just heard about a really cheap three-night tour to Hawaii. The flight leaves on Friday afternoon and gets back on Monday at lunchtime.	
WOMAN:	That's tomorrow afternoon. Are you crazy?	
MAN:	It only costs 50,000 yen for everything. I think we should go, just **for the ___n ___ it**. We could go swimming and shopping and do **____gs ___ke that**.	*(because it will be enjoyable)* *(other similar activities)*
WOMAN:	You're right. **It is _____th the price**. It would be crazy to **m____ a ____ce to go** when it's so cheap.	*(For that price, we should join the tour)* *(not take an opportunity to go)*
MAN:	I'm glad **you've _____ged your ____d**!	*(you are not thinking the way you were)*

Watching TV

In the following conversation, the mother wants her son to study harder.

MOTHER:	I told you not to watch TV and do your homework **at the ____ ____**. I know that you aren't doing a good job with your homework.	*(together)*
SON:	I am doing a good job!	
MOTHER:	Most of your attention is going to the TV, isn't it?	
SON:	**___t ____lly**. Only a little.	*(That's not completely true)*
MOTHER:	**I ___el ___ke** you're doing this just to annoy me. You're doing it **___ ____se**.	*(My feeling is that)* *(because you want to annoy me)*
SON:	No, I'm not!	
MOTHER:	I'm going to **___t ___d of this TV** tomorrow.	*(take this TV out of the house)*
SON:	Don't do that!	
MOTHER:	You'll **___t __sed to it**.	*(soon start to think that you are fine without a TV)*

Asking for help, but not in a direct way

In the following conversation, the older sister wants her younger brother to help her with her smartphone.

SISTER:	Why are you watching TV? I thought you **were _____sed to be** doing your English homework.	*(should be)*
BROTHER:	Don't bother me!	
SISTER:	**Did your English test __o w____** today?	*(Were you successful in your English test)*
BROTHER:	No!	
SISTER:	**____t ___** you fail again? Maybe you need some help. Shall I help you?	*(What will happen if)*
BROTHER:	**____t a ____te**. What are you trying to do here? You're not really interested in my English, are you? Has something **____ne _____g** with your smartphone again? I spent an hour helping you with it yesterday. **Don't ____ke me do** that again!	*(Stop, because I have a question)* *(stopped working properly)* *(I don't want to have to do)*

Late for the test

In the following, the speaker is upset because she doesn't think she can get to school before her test starts.

With this train delay I'm not going to get to school **i___ ____me for the test**, even though **it's ____ going to start ___il** ten. I just sent an email to the professor, telling him what's happened. He answered that he can **___p everyone _____** for a little while, but only **__p ___** ten minutes. It might be quicker if I **___t ____** the train and take a bus. But **____r ___**, I'm going to be too late for the test.	*(early enough to take the test) (it is going to start at)* *(make everyone wait)* *(a maximum of)* *(move out of)* *(whether I stay on the train or take a bus)*

Absent from school

In the following, the speaker had an accident and has not gone to school for a while. His classmates have been really kind to him. He is feeling guilty, because he has actually been enjoying himself.

I've had to **___ke a ____k** from school for the past three weeks, because of my accident. I **___el ___d about** the fact that everyone at school has been **___ling ___y for** me, because actually I've been enjoying myself. I've been able to **___ke a_____ge of** the break to do a lot of studying. I'm sure that I am going to **___o w___** in my classes when I go back to school. I will **be ___p to ___te**, even though I have been absent for three weeks.	*(be absent for a short period)* *(have a guilty feeling about)* *(giving kind attention to)* *(use and not waste)* *(be successful)* *(have the latest information)*

A difficult relationship

In the following, the speaker talks about a difficult classmate. Because of this classmate, she has been having a hard time.

I have a classmate who is always saying negative things about other people. **I'm really ___red ___ listening to her**. I'm afraid she is **___ting ___se**, and I **have been m___ing to say** something to her about it. The other day I **was just ___t to say** something to her when she asked me for advice about a problem she was having. She said she **___dn't ___lp saying** negative things about other people.	*(I don't like listening to her anymore)* *(saying more and more negative things about people)* *(have been thinking that I should say)* *(was on the point of saying)* *(was unable to stop herself from saying)*

Parents and marriage

In the following, the speaker talks about his parents. He is worried about how they will accept the news about his marriage.

My Mom and Dad want to know what I have **been ___p ___** each week. Usually we talk about my part-time job, my friends, my studies and **___t ___t of thing**. They usually think I make decisions without enough thought, which isn't true. So I want to be sure that they won't **___t it _____g** this time. I am going to tell them that I am getting married next week! I am sure that they will tell me to wait until after I finish my college education. But she comes from a very rich family, so they don't need to worry that I **___n't _____d to do this**. And they **___d ____** have a rich son than a poor one!	*(been doing and thinking about)* *(things like these)* *(make a mistake about what I am saying)* *(am not able to do this because I don't have enough money) (prefer to)*

This is the end of the test. Close this booklet and follow the instructions.

Appendix C

Notes regarding the cases where spelling hints in the P-section were prepared in specific ways.

No.	Entry on the EVP website	No. of keywords	Keyword(s)	Test item	Notes on the creation and revision of items
1	keep doing sth	1	keep	____p repeating	
2	again and again	3	again and again	a_____ and _____n	In the pilot, one native speaker wrote *on and on* for the draft item ____*n and* ____*n*; the hints were adjusted accordingly.
3	be up to sb	2	up to	is __p ___ you	
4	be willing (to do sth)	1	willing	be wi_____g to look for	In the pilot, one native speaker wrote *trying* for the draft item _____*g*; the hints were adjusted accordingly.
5	after all	2	after all	a____r a___	In the pilot, two native speakers wrote *for ever* and *for you* respectively for the draft item ____*r* __; the hints were adjusted accordingly.
6	get to know sb/ sth	3	get to know	___t to kn___	In the pilot, one native speaker did not write anything for the draft item ___*t to* ___*w;* the hints were adjusted accordingly.
7	have sth in common	3	have, in common	we had a lot ___ _____	For a three-keyword phrase with a sandwiched filler (*sth*), both the last two keywords were made complete blanks.
8	in advance	2	in advance	___ _____ce	(1) As the first word (*in*) is a preposition, it was totally blanked instead of the second, content word (*advance*). (2) In the pilot, one native speaker wrote *at once* for the draft item ___ _____ *ce*; however, no change was made.

No.	Entry on the EVP website	No. of keywords	Keyword(s)	Test item	Notes on the creation and revision of items
9	at first	2	at first	__t ____t	(1) As the first word (*at*) is a preposition, it was totally blanked instead of the second, content word (*first*). (2) In the pilot, one native speaker wrote *I admit* for the draft item __ ____*t*; the hints were adjusted accordingly.
10	ever since	2	ever since	e____ s_____	In the pilot, one native speaker wrote *after that* for the draft item ____*r* _____; the hints were adjusted accordingly.
11	get cold/ill/late, etc.	1	get	___ting thinner	
12	as you know	3	as you know	__s you ____w	
13	so far	2	so far	S__ ___r	In the pilot, two native speakers wrote *To date* and *To now* respectively for the draft item __*o* ___; the hints were adjusted accordingly.
14	no way	2	no way	There is ___ ___y	If the second word (*way*) were totally blanked, there could be an alternative word to fill (*chance*); the final letter was therefore left.
15	as long as	3	as long as	__s ____g as you eat	
16	I bet (you)	2	I bet	I ___t	As the first word is a pronoun, it was kept as is, and the second word was replaced by a blank with the final letter remaining.
17	it/that depends	2	it depends	It _____ds	As the first word is a pronoun, it was kept as is, and the second word was replaced with a blank with its final letters remaining.
18	fall asleep	2	fall asleep	__ell _____	
19	tell sb how/ what/when to do sth	4	tell, what to do	told her what ___ ___	
20	(just) in case	3	just in case	____t in ___se	
21	if I were you	4	if I were you	___ I _____ you	
22	for fun or for the fun of it	5	for the fun of it	for the ___n ___ it	With the paraphrase created for this phrase (*because it will be enjoyable*), it was expected that the word *fun* would be too difficult, if there were no hint, for test-takers to come up with; the final letter was therefore left.
23	things like that	3	things like that	_____gs ___ke that	
24	be worth sth/ doing sth	1	worth	It is ____th the price	
25	miss a chance/ opportunity	3	miss a chance/ opportunity	m____ a _____ce to go	In the pilot, one native speaker wrote *pass (up) a chance to go* for the draft item __*ss a* ____*ce to go*; the hints were adjusted accordingly.
26	change your mind	3	change your mind	you've _____ged your ____d	
27	at the same time	4	at the same time	at the _____ _____	

No.	Entry on the EVP website	No. of keywords	Keyword(s)	Test item	Notes on the creation and revision of items
28	not really	2	not really	___t ____lly	If the second word (*really*) were totally blanked, there could be an alternative word to fill (e.g., *most*); the final letters were therefore left.
29	feel like/as if	2	feel like	I ___el ___ke	If the second word (*like*) were totally blanked, there could be an alternative word to fill (*that*); the final letters were therefore left.
30	on purpose	2	on purpose	___ _____se	As the first word (*on*) is a preposition, it was totally blanked instead of the second, content word (*purpose*).
31	get rid of sth	3	get rid of	___t ___d of this TV	
32	get/become used to sb/sth/ doing sth	3	get used to	___t __sed to it	
33	be supposed to do sth	1	supposed	were _____sed to be	
34	go badly/well, etc.	2	go well	Did your English test __o w____	In the pilot, one native speaker wrote *go well/ okay/fine* for the draft item __*o* ____; the hints were adjusted accordingly.
35	What if ...?	2	what if	____t ___	
36	wait a minute	3	wait a minute	____t a _____te	
37	go wrong	2	go wrong	___ne _____g	If the second word (*wrong*) were totally blanked, there could be an alternative word to fill (*bad*); the final letter was therefore left.
38	make sb do sth	1	make	Don't ___ke me do	
39	in time	2	in time	i__ ___me for the test	In the pilot, all three native speakers wrote *on time for the test* for the draft item __ ___*me for the test*; the hints were adjusted accordingly.
40	not until	2	not until	it's ___ going to start ___il	If the second word (*until*) were totally blanked, there could be an alternative word to fill (*till*); the final letters were therefore left.
41	keep sb waiting	2	keep, waiting	____p everyone _____	
42	up to 10, 20, etc.	2	up to	___p ___	
43	get down/into/ off, etc.	2	get off	___t ____	
44	either way	2	either way	_____r ____	
45	take a break/ rest, etc.	3	take a break	___ke a _____k	
46	feel bad about sth/doing sth	3	feel bad about	__el __d about	
47	feel sorry for	3	feel sorry for	___ling _____y for	
48	take advantage of sth	3	take advantage of	____e a_____ge of	In the pilot, one native speaker wrote *take charge of* for the draft item ___*ke* _____*ge of*; the hints were adjusted accordingly.

No.	Entry on the EVP website	No. of keywords	Keyword(s)	Test item	Notes on the creation and revision of items
49	do badly/well	2	do well	__o w____	In the pilot, one native speaker wrote *do well/great* for the draft item __*o* ____; the hints were adjusted accordingly.
50	up to date	3	up to date	be __p to ___te	
51	tired of doing sth	2	tired of	I'm really __red __ listening to her	
52	get worse	2	get worse	____ting _____se	With the paraphrase created for this phrase (*saying more and more negative things about people*), it was expected that the word *worse* would be too difficult, if there were no hint, for test-takers to come up with; the final letters were therefore left.
53	have been meaning to do sth	1	meaning	have been m____ing to say	In the pilot, one native speaker wrote *meaning/planning* for the draft item ____*ning*; the hints were adjusted accordingly.
54	be just about to do sth	1	about	was just ____t to say	
55	can't/couldn't help doing sth	2	couldn't help	____dn't ___lp saying	(1) For the first word (*couldn't*), the third letter from the last (*d*) was left because tense was not tested. (2) If the second word (*help*) were totally blanked, there could be an alternative word to fill (*stop*); the final letters were therefore left.
56	be up to sth	2	up to	been __p ___	
57	that sort of thing	4	that sort of thing	____t ____t of thing	If both *that* and *sort* were totally blanked, an alternative phrase (*this kind*) would become possible; both words' final letter was therefore left.
58	get sth wrong	2	get, wrong	___t it _____g	With the paraphrase created for this phrase (*make a mistake about what I am saying*), it was expected that the word *wrong* would be too difficult, if there were no hint, for test-takers to come up with; the final letter was therefore left.
59	can afford	2	can afford	___n't _____d to do this	In the pilot, one native speaker wrote *don't want* for the draft item ___*n't* _____; the hints were adjusted accordingly.
60	would rather	2	would rather	_____d _____	