Yukio Tono and Naho Kawamoto

Developing and calibrating "Can do" descriptors for dictionary use by EFL learners using the Rasch model

Abstract: This study developed a "can do" list of dictionary skills for English language learners and tested item difficulty using the Rasch model. The list, based on the Common European Framework of Reference for Languages (CEFR), aimed to prototype and validate skill levels similar to CEFR. It selected descriptors through a literature review spanning 30 years, focusing on encoding and decoding activities across four stages. A questionnaire based on these descriptors was completed by 223 Japanese university students learning English. After classifying items into "can do" and "can't do" categories using a dichotomous model, the Rasch analysis was used to assess item difficulties and perform fit analysis. The partial credit model was then applied for a more precise analysis. Results indicated instability in interpreting item difficulty but proved promising in estimating dictionary skill difficulty. Future work involves revising misfit items to enhance descriptor coherence for learners.

Keywords: dictionary reference skills, "can do" list, CEFR, descriptors, Rasch model, English as a foreign language (EFL)

1 Introduction

The field of dictionary user needs and skills analysis has gradually developed since the 1980s, with various empirical studies (e.g., Tono 2001) and a seminal summary paper in the field (Nesi 2015) being published. This development was largely spurred by the emergence of the so-called Big 5,[1] English monolingual learners' dictionaries, accompanied by a flourishing body of empirical research employing English learners as subjects (see Nesi (2015) for a review). Conversely, European dictionary use research has not solely focused on foreign language learners but has encompassed a broader audience,

1 *Oxford Advanced Learner's Dictionary* (OALD), *Longman Dictionary of Contemporary English* (LDOCE), *Collins COBUILD English Dictionary* (COBUILD), *Cambridge International Dictionary of English* (CIDE), and *Macmillan English Dictionary for Advanced Learners* (MEDAL).

Note: This paper is a revised version of the presentation given at the AILA 2023.

Yukio Tono and Naho Kawamoto, Tokyo University of Foreign Studies

https://doi.org/10.1515/9783111373294-007

including general users, interpreters, and translators (Atkins/Varantola 1997; Kosem et al. 2019).

In tandem with the evolution of these research domains, the learning environments in which users are situated have undergone significant transformations. The advent of the Big 4 in 1995 catalysed a resurgence in the market for printed dictionaries. Notably, within the realm of English learner's dictionaries, Harold E. Palmer and A.S. Hornby, British educators invited to enhance English language education in Japan, made considerable contributions. Their early research on vocabulary selection, collocation, etc., culminated in the development of the *Idiomatic Syntactic English Dictionary* in 1943. Subsequently, this work was re-published by Oxford University Press, evolving into the globally acclaimed *Oxford Advanced Learner's Dictionary* (OALD). Consequently, the global market for dictionaries catering to English language learners experienced a significant boom (Cowie 2000).

The 2000s witnessed an explosive expansion in internet usage, precipitating a shift from paper to electronic media for dictionaries. In Japan, the market for pocket electronic dictionaries underwent rapid expansion from the 1990s onwards, with the market size peaking at nearly one million units per year from 2000 to the late 2010s. While the merits and drawbacks of transitioning from paper to electronic dictionaries were debated (Koyama 2013; Lew 2013), the omnipresence of the internet led to the commonplace practice of accessing online dictionaries via smartphones or tablet PCs. This shift in interface fundamentally altered how users interact with dictionary information, necessitating a re-evaluation of pertinent variables within the field and a reexamination of critical issues.

In the 2020s, the widespread adoption of smartphones, machine translation tools such as DeepL, and the emergence of generative AI, represented by ChatGPT, have precipitated changes that have reverberated throughout the foundations of language education. While the relationship between these new information technologies and dictionary information warrants future discussion, the most immediate concerns in language education revolve around delineating the requisite skills, objectively defining the language competencies to be acquired, determining which should be internalized as inherent abilities, and discerning which aspects should be supplemented by AI and other tools to achieve educational objectives.

In parallel with the growing need to clearly define language competence, the Common European Framework of Reference for Languages (CEFR, hereafter) was published by the Council of Europe in 2001. The CEFR has since played an important role as a generic framework for foreign language learning, teaching, and assessment. It was designed for the first time as a generic "descriptive tool," detailing the skills and levels to be achieved across languages. In the updated 2020 version called the Companion Volume (CEFR/CV, henceforth), the main framework consists of seven major common reference levels: Pre-A1, A1, A2, B1, B2, C1, and C2, encompassing the areas of communicative language activities: reception, production, interaction, and mediation. The descriptions of what language users can do in a given skill at each level are defined using so-called “can do” descriptors. These “can do” descriptors are statistically determined in terms of their difficulty order.

Another feature of the CEFR is that, while defining communicative language activities, it also provides a detailed definition of the strategies used to perform these activities. Using this framework, we argue that it is possible to incorporate dictionary reference skills as part of the learning strategies. As will be discussed in the literature review, the current CEFR does not address the use of supplementary materials and references in detail. To address this gap, this study aims to create a set of new “can do” descriptors for dictionary use using the CEFR. In this way, dictionary reference skills can be redefined within the context of the CEFR. Furthermore, these dictionary reference skills can be linked to the CEFR language activities and levels as learning strategies, and can be used to investigate new dictionary skills in the evolving online learning environment.

To achieve these objectives, Kawamoto/Tono (2023) conducted a comprehensive literature review of previous studies on dictionary users. They extracted the main dictionary search skills for learning foreign languages, especially English, and created a tentative list of “can do” descriptors for dictionary use (see Section 2.3 for a summary). Next, these descriptors went through questionnaires administered to 223 participants to determine their confidence in performing each “can do” skill. This study presents the results of calibrating these descriptors based on the “can do” questionnaire responses, using Rasch analysis similar to that conducted by the CEFR.

2 Review

Here we will review the basic characteristics of “can do” descriptors (2.1) and how dictionary skills are treated in the CEFR descriptors (2.2). Section 2.3 summarises the making of the “can do” descriptor list for dictionary use reported in Kawamoto/Tono (2023).

2.1 Illustrative “can do” descriptors and dictionary reference skills

According to the CEFR-CV (2020), the idea of scientifically calibrating “can do” descriptors to a scale of levels comes originally from the field of professional training for nurses. It says, “what was needed was a systematic, informed observation by an expert nurse, guided by short descriptions of typical nursing competence at different levels of achievement.” (ibid, p. 35) This “can do” approach was transferred to language teaching and learning in the work of the Council of Europe and nowadays “can do” descriptors are applied to more and more disciplines in many countries in what is often referred to as a competence-based approach.

While the CEFR provides comprehensive specifications for different aspects of communicative language activities, communicative language competence, and communication and learning strategies, there is a notable gap when it comes to describing the skills required for utilizing reference materials, particularly dictionaries, for receptive and productive language activities. Although the original CEFR (2001) briefly acknowledged the importance of dictionary use within the Ability to learn (*savoir apprendre*) category, it provided limited details on how dictionaries should be effectively employed in language learning processes. The descriptors pertaining to dictionary skills were scattered throughout the CEFR and lacked specificity in terms of the types of information that learners should activate when utilizing dictionaries.

Recognizing the need for a more comprehensive treatment of dictionary reference skills, the CEFR-CV (2020) has revisited and enhanced the original framework, incorporating new perspectives on modes of communication, mediation skills, and online communication. While the CEFR-CV introduced a range of descriptors that incorporate dictionary use across different language activities and proficiency levels, the level of detail regarding the specific dictionary reference skills required remained limited. The CEFR-CV acknowledged the importance of dictionaries as aids for comprehension, writing, interaction, and mediation but did not provide explicit guidelines on which dictionary information should be accessed to address communication problems effectively. The following section will look at more detail about the CEFR descriptors related to dictionary use.

2.2 Review of descriptors related to dictionary use

2.2.1 The CEFR (2001)

In the CEFR (2001), dictionary skills were briefly mentioned as part of the broader Ability to learn (*savoir apprendre*) category. The descriptors highlighted the importance of skills and know-how, such as using a dictionary or navigating a documentation centre. For example, in Chapter 4, the following two descriptors contain the use of a dictionary, as in (1):

(1) a. READING CORRESPONDENCE C1: “Can understand any correspondence given the occasional use of a dictionary”
b. READING FOR INFORMATION AND ARGUMETN B2+: “Can understand specialised articles outside his/her field, provided he/she can use a dictionary occasionally to confirm his/her interpretation of terminology.”

As shown in (1a), descriptors often mention dictionary use with the expressions such as “given the occasional use of a dictionary,” which indicates a condition under which a given communicative activity is performed. Since this is a C-level descriptor, it expresses the almost perfect command of reading correspondence on the condition that the occasional use of a dictionary is available.

In Chapter 4.5 of the original CEFR (2001), the use of a dictionary was briefly mentioned as a part of communicative language processes. For comprehension, especially of written texts, the proper use of aids was recommended, including reference materials such as (a) dictionaries (monolingual and bilingual); (b) thesauruses; (c) pronunciation dictionaries; (d) electronic dictionaries, grammars, spell-checkers and other aids; and (e) reference grammars. This is the only place where different types of dictionaries were presented as aids of communicative language activities.

2.2.2 The CEFR-CV (2020)

The revised descriptors in the CEFR-CV treat dictionary use more extensively, providing a clearer understanding of its role in communicative language activities. For instance, in the domain of reading, if we compare the descriptors such as READING AS A LEISURE ACTIVITY B1 and B2, it is clear that they emphasize the regular use of a dictionary in B1 level (cf. 2b), which is less obvious in B2 level due to more advanced proficiency, to aid comprehension of reading novels for pleasure:

(2) a. READING AS A LEISURE ACTIVITY B2: Can read novels with a strong, narrative plot and that use straightforward, unelaborated language, <u>provided they can take their time and use a dictionary</u>. (Underline ours)
b. READING AS A LEISURE ACTIVITY B1: Can follow the plot of stories, simple novels and comics with a clear linear storyline and high frequency everyday language, <u>given regular use of a dictionary</u>. (Underline ours)

The descriptors shown in (3) to (6) cover the domain of written production and interaction, where descriptor scales such as OVERALL WRITTEN PRODUCTION Pre-A1, CREATIVE WRITING A2, OVERALL WRITTEN INTERACTION Pre-A1, and WRITTEN INTERACTION: CORRESPONDENCE Pre-A1 and A1 highlight the ability to convey basic information and compose messages using a dictionary as a reference:

(3) OVERALL WRITTEN PRODUCTION Pre-A1: Can give basic personal information (e.g. name, address, nationality), perhaps with the use of a dictionary.

(4) CREATIVE WRITING A2: Can compose an introduction to a story or continue a story, provided they can consult a dictionary and references (e.g. tables of verb tenses in a course book).

(5) OVERALL WRITTEN INTERACTION Pre-A1: Can convey basic information (e.g. name, address, family) in short phrases on a form or in a note, with the use of a dictionary.

(6) a. WRITTEN INTERACTION: CORRESPONDENCE Pre-A1: Can convey basic personal information in short phrases and sentences, with reference to a dictionary.
b. WRITTEN INTERACTION: CORRESPONDENCE A1: Can compose messages and online postings as a series of very short sentences about hobbies and likes/dislikes, using simple words and formulaic expressions, with reference to a dictionary.

The CEFR-CV also recognizes the role of dictionaries in mediation tasks, shown in (7) to (9). Mediation in its main sense involves "passing on to another person the content of a text to which they do not have access, often because of linguistic, cultural, semantic or technical barriers" (CEFR-CV, p.91). During this mediation process, dictionary consultation is sometimes needed. Descriptor scales such as MEDIATION: EXPLAINING DATA B1, MEDIATION: PROCESSING TEXT A1, and MEDIATION: TRANSLATING A WRITTEN TEXT A1 illustrate how dictionaries can support the interpretation, explanation, and translation of texts for mediation purposes:

(7) MEDIATION: EXPLAINING DATA B1: Can interpret and present in writing (in Language B) the overall trends shown in simple diagrams (e.g. graphs, bar charts) (with text in Language A), explaining the important points in more detail, given the help of a dictionary or other reference materials.

(8) MEDIATION: PROCESSING TEXT A1: Can, with the help of a dictionary, convey (in Language B) the meaning of simple phrases (in Language A) on familiar and everyday themes.

(9) MEDIATION: TRANSLATING A WITTEN TEXT A1: Can, with the help of a dictionary, translate simple words/signs and phrases (from Language A into Language B), but may not always select the appropriate meaning.

While the updated descriptors in the CEFR-CV offer different language use contexts involving dictionary use, the description of dictionary use itself is rather simple and there remains an opportunity for more detailed exploration. In the next section, we will summarise our previous paper (Kawamoto/Tono 2023) in which a tentative list of "can do" descriptors were prepared.

2.3 Dictionary skills descriptor development

The development of descriptors was initiated with a thorough literature review to identify key sources on dictionary reference skills. This encompassed dictionary use questionnaires (Atkins/Varantola 1997), dictionary look-up processes (Atkins 1996; Béjoint 1981; Béjoint 1994; Hartmann 1999; 2001; Hartmann/James 2002; Higuchi 2012; Lew/Galas 2008; Nesi 1999; Scholfield 1982;), dictionary skills tests (Tono 2001), empirical studies on dictionary users' reference skills (Béjoint/Moulin 1987; Harvey/Yuill 1997; Koyama 2013; Lew 2013; Lew/Mitton 2011; Nesi/Meara 1994; Neubach/Cohen 1988; Nuccorini 1992; Wingate 2004; Winkler 2001), and resources detailing dictionary workbooks and inventories of skills and strategies for dictionary use (Gavriilidou 2013, 2014; Gavriilidou/Mavrommatidou 2016; Mavrommatidou/Gavriilidou/Markos 2019).

As we produced descriptors, we emphasized the essential characteristics of these descriptors. For example, strategy inventories, like those by Mavrommatidou/Gavriilidou/Markos (2019), often illustrated user preferences, such as the choice of dictionaries or mediums such as paper or online. While informative, these aspects do not encapsulate the action-oriented nature that CEFR "can do" descriptors demand – specifically, the user's active performance in language-related tasks requiring dictionary consultation. Hence, our focus shifted to the actions, such as "finding the meaning of a word" or "identifying the correct collocation for a target word," rather than mere conditions or preferences for dictionary use.

The literature reviewed provided a wealth of information on dictionary skills, which we meticulously compiled into a database. Within this, each skill was classified with respect to the corresponding reference, types of information, and the involved look-up processes, such as planning before consultation, searching for the headword (macrostructure process), searching for the necessary information within the entry (microstructure process) and retrieving information and applying it to the context at hand (application/evaluation). Adopting North and Piccardo's (2019: 153) methodological approach, we collated significant dictionary-related actions and information from the literature to commence the descriptor development. This review included an analysis of descriptors from the CEFR and DIALANG projects, which often vaguely referenced "the use of a dictionary." Our task was to dissect these references and articulate the specific dictionary information needed for a range of communicative tasks.

Our target demographic was designated as users of English as a foreign language, spanning proficiency levels from A1 to B2. This also implies that the findings in this study need to be validated against users of different languages, which is beyond the scope of this paper. The exclusion of C-level users was intentional, acknowledging that the dictionary needs of those at higher proficiency levels – often involving specialized and technical vocabulary – differ markedly from the general language competence sought by A1-B2 learners through dictionary use. This decision enabled a concentrated effort to develop descriptors that closely resonate with the needs and abilities of learners up to the B2 level, enhancing the descriptors' relevance and utility for this audience.

Upon refining our focus, we curated a selection of 32 illustrative descriptors to include in the "can do" questionnaire (detailed in Appendix A).

3 Method

3.1 Purposes of the study

The present study aims to produce a list of scaled descriptors for L2 dictionary use. What is meant by "scaled" is to calibrate descriptors in the same method as the original CEFR.

The research questions in this study were formulated as follows:
RQ1. What characteristics can be found in the list of scaled "can do" descriptors for L2 dictionary use regarding the types of reference skills and their difficulty order?
RQ2. Is there any room for improvement in the formulation of "can do" descriptors?
RQ3. What relationship can be identified regarding the relationship between dictionary reference skills and the CEFR levels?

3.2 Participants

A total of 223 university students participated in this study, representing two different universities: a national university, where the students' academic level is high and most students major in foreign languages and cultures or area studies, and a private university, whose students' academic levels are average. The participant pool consisted mainly of non-English majors at the national university who were learning English as their third language, while the students at the private university belonged to either the Education or Computer Science department. It should be noted that the difference between the academic levels of the two universities should be taken into account. Overall, the students at the national university were more proficient in English and they reached B1+ to B2 level overall, whereas the students at the private university mostly belonged to A1 or A2.

3.3 Instruments: a questionnaire survey and Rasch analysis

3.3.1 The "can do" questionnaire

The questionnaire used in this study consisted of a total of 32 "can do" descriptors provided to the participants through a Google Forms survey. Participants were asked to indicate how well they could perform the task described in each descriptor of dictionary use using a 5-point Likert scale. A sample question item is shown in (10):

(10) "I can predict the base form of the problem word."
4: "Strongly agree" 3: "Agree" 2: "Neutral" 1: "Disagree" 0: "Strongly disagree"

Participants were instructed to rate their ability for each descriptor on the scale, with 4 indicating "Strongly agree," 3 indicating "Agree," 2 indicating "Neutral," 1 indicating "Disagree," and 0 indicating "Strongly disagree." They were asked to evaluate their abilities based on scenarios involving printed, electronic, and online dictionaries, excluding translation tools such as DeepL or Google Translate. While the "can do" descriptors were prepared and available in both English and Japanese, only the Japanese descriptors were utilized for this particular study.

3.3.2 Rasch analysis

Rasch analysis is a statistical method commonly used in educational and psychological research to evaluate the measurement properties of assessments, surveys, and other instruments. At its core, Rasch analysis aims to provide researchers with a common, interval-level metric by which both persons (i.e., individuals taking the assessment) and items (i.e., the questions or tasks within the assessment) are measured. This means that Rasch analysis facilitates the comparison of individuals' abilities or traits and the difficulty of assessment items on the same scale, allowing for precise measurement and meaningful interpretation. In this study, "persons" correspond to the participants of the questionnaire and "items" are descriptors.

One of the key features of Rasch analysis is its provision of fit statistics, which help researchers assess the extent to which individual persons or items conform to the Rasch model. Fit statistics indicate whether there are discrepancies between observed responses and expected responses according to the model. By identifying items or persons that do not fit the model well, researchers can refine their assessments and improve their measurement precision.

Furthermore, Rasch analysis allows for the evaluation of the coverage of the latent construct being measured. This involves assessing whether the range of items included in the assessment adequately represents the full spectrum of the construct being measured. Ensuring sufficient coverage is essential for accurately capturing individuals' abilities or traits across the entire range of the construct.

We used the extended Rasch modelling (*eRm*) package (Mair/Hatzinger 2007), which offers several functions to facilitate Rasch analysis, including "Estimates" and "Model diagnostics." The "Estimates" function in the *eRm* package provides estimates of individuals' abilities and item difficulties based on the Rasch model. These estimates are derived using conditional maximum likelihood estimation methods (Anderson

1972) and are represented on a common interval scale, allowing for direct comparisons between individuals and items.

On the other hand, the "Model diagnostics" function in the *eRm* package offers tools for assessing the adequacy of the Rasch model fit. This includes calculating INFIT and OUTFIT statistics, which measure the goodness of fit of individual items or persons to the Rasch model. Additionally, the function provides INFIT t-test statistics, which can be plotted for items or persons to visually assess model fit. Finally, the function includes testing procedures, such as Martin-Löf test (MLoef), to assess the unidimensionality of the measurement instrument, ensuring that it is measuring a single underlying trait or construct.

4 Results and Discussion

This section will report the results of two types of Rasch analysis based on the dichotomous Rasch Model (4.1) and the Partial Credit Model (4.2). After summarizing the fitted models and their diagnosis, we will move on to answering the three research questions (4.3).

4.1 Dichotomous Rasch model

The "can do" questionnaire asked participants to assess how well they think they can perform the tasks described in the "can do" descriptors. The present study compiled a list of 32 "can do" descriptors (see Appendix A), primarily covering four types of dictionary reference skills. The first type (Items No. 1–4, hereafter referred to as "before dictionary consultation") deals with the knowledge or skills needed to identify problematic words or phrases within a context and decide on look-up strategies, including the selection of appropriate dictionaries and the identification of problematic areas. The second type (Items No. 5–9, hereafter referred to as "macrostructure") deals with the selection of headwords and related skills such as predicting the base form or the meaning of the problem word. The third type (Items No. 10–28, hereafter referred to as "microstructure") encompasses a wide range of linguistic information presented under a headword as part of the microstructure. This includes details such as spelling, pronunciation (notated in katakana or IPA), stress, part of speech, countability, morphological information, distinctions between intransitive and transitive verbs, singular versus plural forms, usage labels (e.g., US/UK; formal/informal; spoken/written), verb codes or patterns, synonyms and antonyms, menus or signposts, etymology, and cross-referencing. The final type (Items No. 29–32, hereafter referred to as "application") consists of skills for applying dictionary information to construct new sentences, which includes analysing illustrative examples, verb patterns, and collocations.

Firstly, we fitted the simplest model, called the Dichotomous Rasch Model. This model uses sum scores from these ordinal responses (i.e., 0 to 4) to calculate interval-level estimates that represent person locations (i.e., person ability or person achievement) and item locations (i.e., the difficulty to provide a correct or positive response) on a linear scale that represents the latent variable (the log-odds or "logit" scale). The difference between person and item locations can be used to calculate the probability for a positive response to a "can do" descriptor (x = 1), rather than a negative response (x = 0).

The results of the Rasch analysis for the dichotomous data provided valuable insights into the characteristics of the developed "can do" descriptors for dictionary skills. Firstly, the analysis confirmed that the result of the Martin-Löf test showed that unidimensionality is tenable (LR-value: 144.074; Chi-square df: 271; p=1.00), indicating that the descriptors represented a unidimensional construct, measuring a single underlying skill set related to dictionary use.

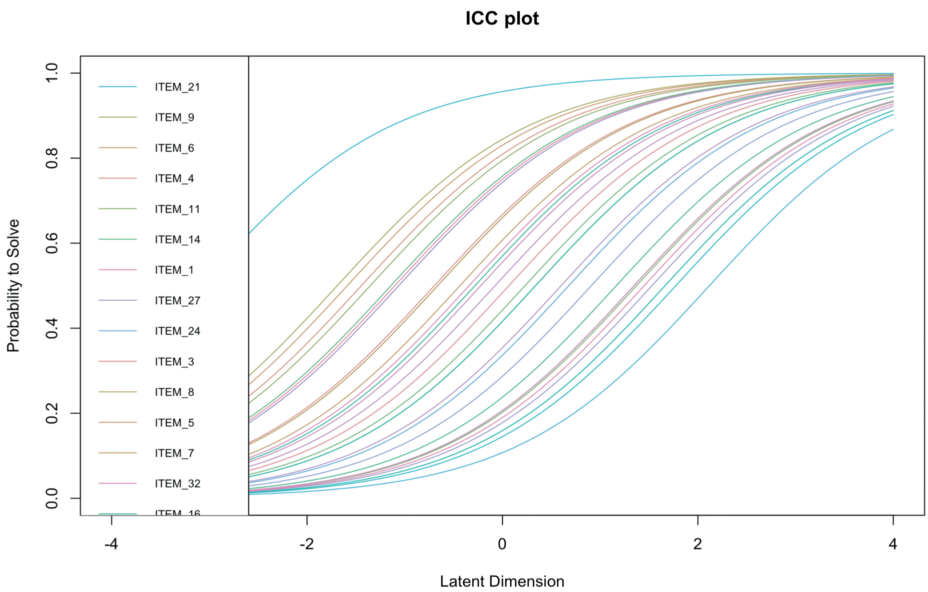

Figure 1: Joint ICC plot of dictionary skills descriptors.
Note: The ITEM number corresponds to the number of the descriptors in Appendix A. Due to limitations in controlling *eRm* graph functions, not all items are visible on the legend.

The Item Characteristic Curve (ICC) plot (see Figure 1) displayed a gradual progression of item difficulty across the latent dimension (see Appendix B). The plot demonstrated a smooth curve without any unusual intersections, indicating that the items fit well within the overall construct. Only one item showed a slight misfit, identified as Item 21, which was considered the easiest but exhibited some deviation from the expected pattern.

The Person-Item map, shown in Figure 2, visualized the relationship between participants' proficiency levels and the difficulty levels of the descriptors across the latent dimension (the logit scale). This diagram also shows that Item_21 is located toward the left-most end of the plots, showing that the item is considered extremely easy by most participants.

The next step involves investigating the fit of the items to the Rasch model. Item fit information was obtained using the item fit statistics presented (see Appendix C). The results provided further evidence that, overall, the data fit the Rasch model well, with only a few exceptions. Specifically, Items 14, 22, and 25 exhibited statistically significant values for OUTFIT *t*, indicating that the residuals for these items are more variable than what the Rasch model predicts.

4.2 Partial Credit Model

Given that the "can do" questionnaire features polytomous responses, such as the Likert scale format ranging from 0 to 4, dichotomizing the five response categories would be somewhat artificial. Therefore, we extended the Rasch model to polytomous models, employing the Partial Credit Model (PCM) using the *eRm* package. The full output of the fitted PCM is presented in Appendix D. With the five-category response scale for the "can do" questionnaire, we obtain four parameters per item corresponding to the four thresholds. By default, the first threshold of the first item is fixed to 0. The easiness parameters are also reported for this polytomous model.

The result of the Martin-Löf test showed that unidimensionality is again tenable (LR-value: 742.521; Chi-square df: 4095; p=1.00). Figure 3 illustrates the person-item map for the PCM. In this visualization, not only are the item locations represented by black dots, but also the threshold locations. This provides a better understanding of how participants responded to the descriptors compared to the dichotomous model. For instance, items marked with a (*) on the right vertical axis (Items 10 and 23) exhibited disordered thresholds (where 3 and 4 were not in the correct order). These items present a problem and may require either removal or improvement by collapsing the "Strongly Agree" and "Agree" categories.

Compared to the dichotomous model presented in Figure 2, the order of descriptors in the Partial Credit Model (PCM) may exhibit slight variations. However, the PCM offers more detailed insights into how participants responded to each item. When an item performs well, each response category (1 to 4) becomes the most likely response within a certain range of the latent dimension. Consequently, the Item Characteristic Curve (ICC) for each category typically exhibits a well-defined bell shape, with the curves shifting from easy to difficult along the latent dimension (as illustrated by Item No. 1 in Figure 4).

Despite minor changes, the PCM results also show that "can do" item No. 21 was found to be the easiest, while No. 22 proved to be the most challenging. Figure 5 provides a comparison of the Item Characteristic Curve (ICC) plots for Items No. 21 and 22. Due to the plotting of characteristic curves for each response category, joint plotting becomes too complicated and is not available. For Item No. 21, the curve line toward the right edge (light blue line) is the most prominent, indicating that a majority of respond-

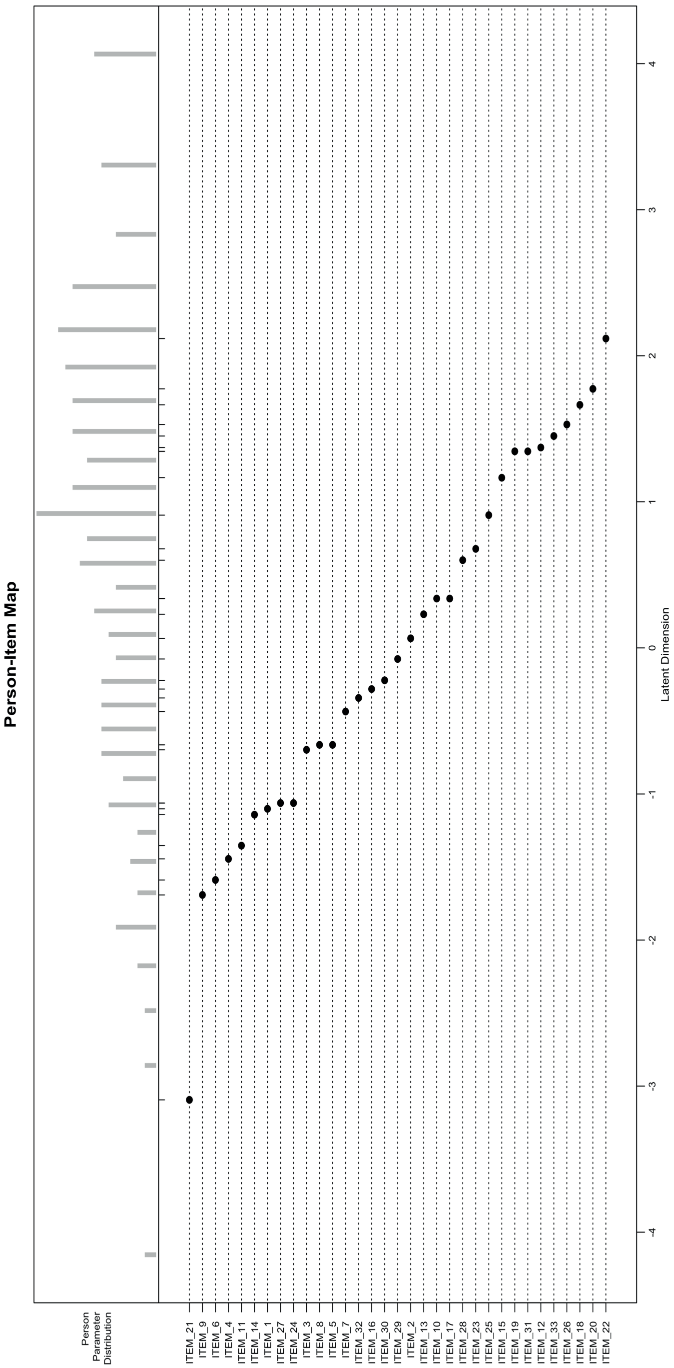

Figure 2: Person-Item map for the dichotomous model.

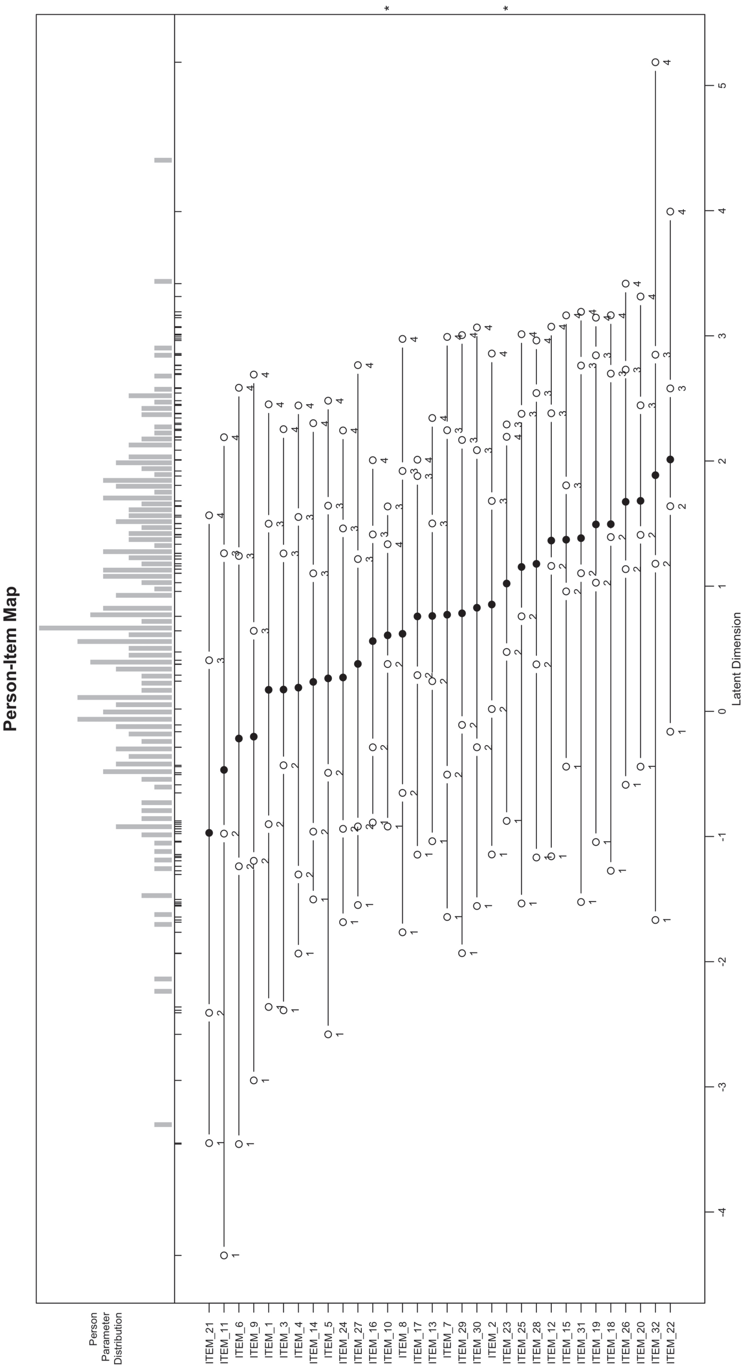

Figure 3: Person-Item map for the partial credit model.

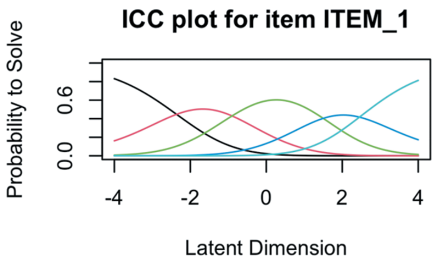

Figure 4: The ICC curve for Item No. 1.

ents are clustered around scale 4 on the latent dimension. It is worth noting that this package reports easiness parameters for the polytomous model (Padgett/Morgan 2020). Therefore, in this context, a higher score on the theta scale signifies easier items. Conversely, Item No. 22 demonstrates a majority of responses clustering around the "-4" area, indicating its extreme difficulty level.

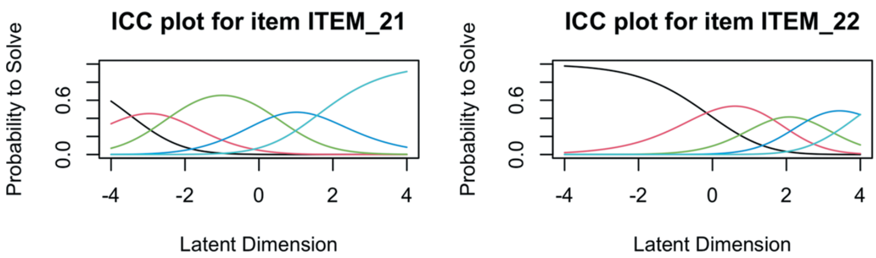

Figure 5: ICC plots for Items No.21 and No. 22.

Finally, we generated a Pathway Map, which plots the location of each item or person against its infit t-statistic (see Appendix E). Pathway maps serve as valuable tools for identifying misfitting items or persons within the Rasch model framework. Ideally, both items and individuals should exhibit infit t-statistics ranging between approximately -2 and +2, as indicated by the marked values. Figure 6 illustrates the Pathway Map based on the "can do" questionnaire for dictionary skills. The vertical green lines delineate the –2 and +2 thresholds.

Given that the Rasch model employs infit statistics rather than traditional residual analysis used in regression analysis, it is crucial to utilize these statistics to identify items that do not conform to the Rasch model and to explore the underlying reasons for their misfit. In Figure 6, several items (No. 2, 14, 22, 25) are identified as misfitting. Notably, these items were also flagged as problematic in the dichotomous model analysis (cf. Section 4.1). Their response patterns deviate from the expectations of the Rasch model, indicating a need for further investigation and potential refinement.

Figure 7 illustrates that there are numerous instances of misfitting persons. While this paper does not extensively explore the factors influencing individuals' misfit, it is

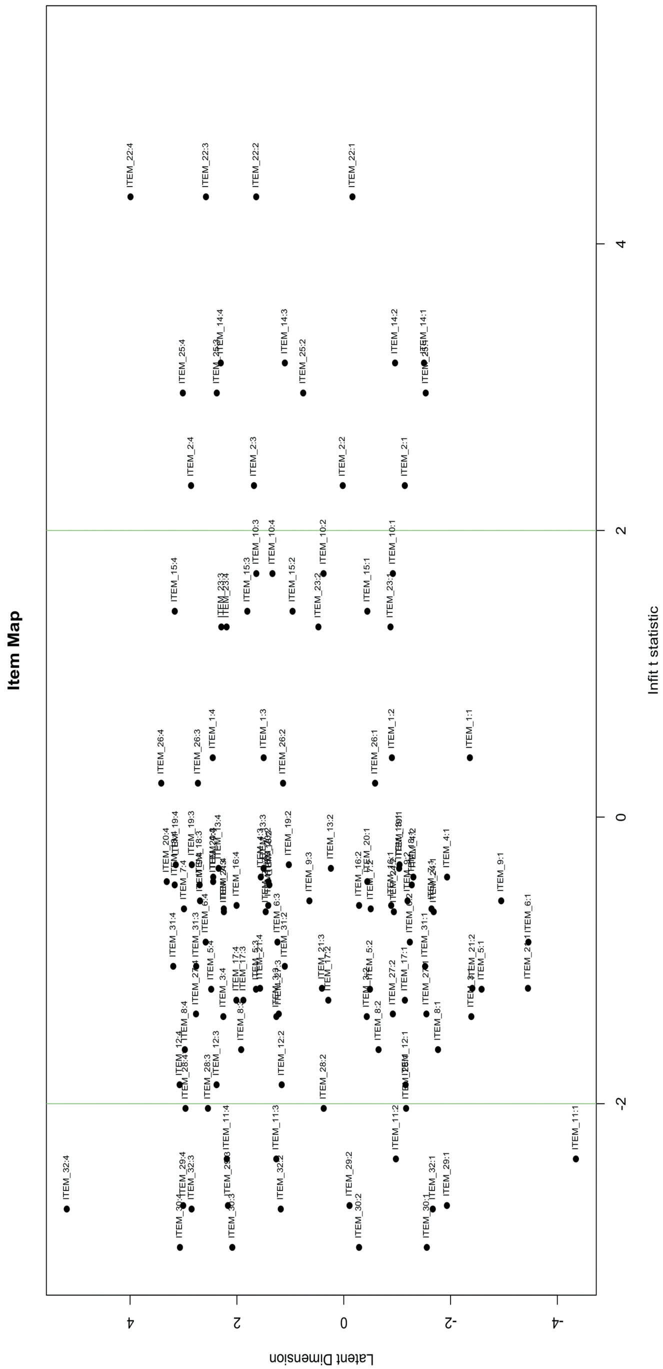

Figure 6: Path Way map (Item).

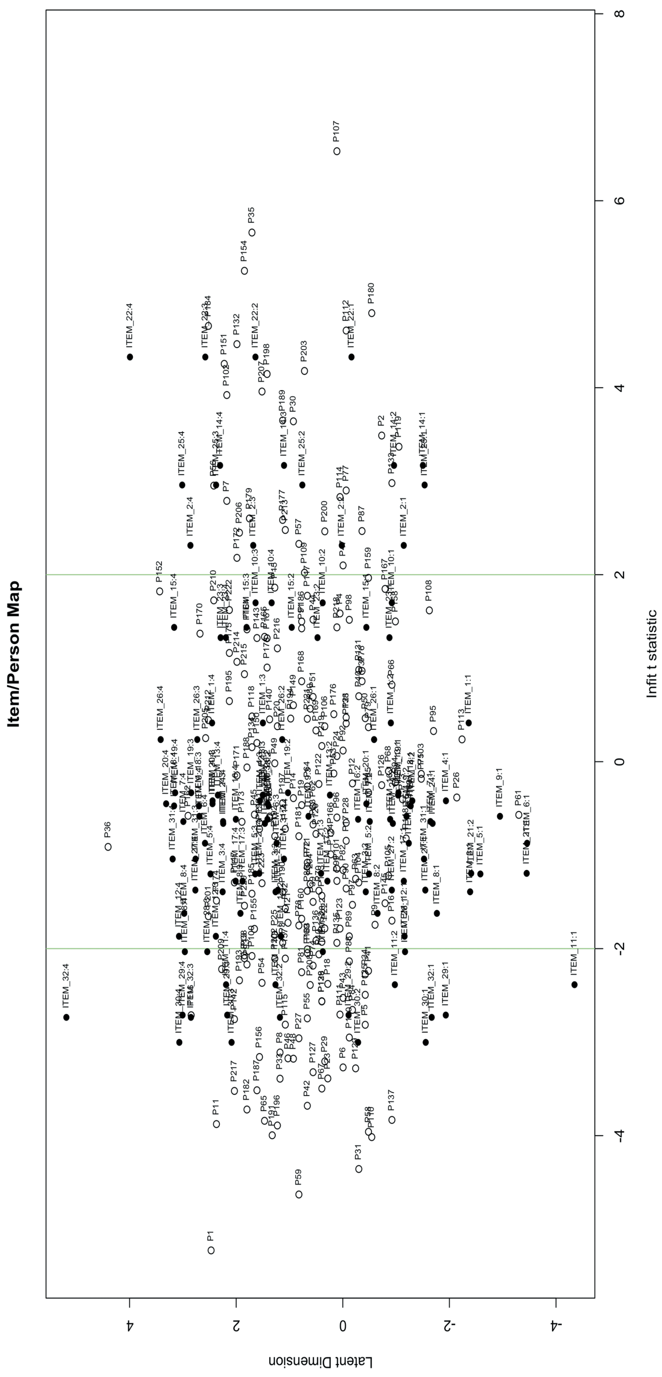

Figure 7: Pathway map (Item-Person).

crucial to highlight a significant aspect. Tono/Kawamoto (in press) conducted an experiment aimed at assessing whether users who responded positively to specific "can do" items could effectively execute the dictionary operations described in the descriptors. The findings of this experiment are mixed, revealing a noticeable disparity between users' perceptions of their abilities and their actual performance. This disparity underscores the importance of further exploration in understanding and addressing the gap between users' self-assessments and their real-world proficiency levels. Moving forward, investigating this issue will be essential for refining the "can do" list and ensuring its alignment with users' actual capabilities.

4.3 Responses to the research questions

This section will examine the two versions of the Rasch model, the dichotomous model and the Partial Credit Model (PCM), in order to explore the answers to the three research questions below:
RQ1. What characteristics can be found in the list of scaled "can do" descriptors for L2 dictionary use regarding the types of reference skills and their difficulty order?
RQ2. Is there any room for improvement in the formulation of "can do" descriptors?
RQ3. What relationship can be identified regarding the relationship between dictionary reference skills and the CEFR levels?

4.3.1 RQ1: Characteristics of the scaled "can do" descriptors for L2 dictionary use

Appendix A presents the list of "can do" descriptors ordered by Easiness Parameters (beta) with 0.95 confidence intervals derived from the dichotomous Rasch model. It is important to recall that these descriptors were initially formulated with four search stages in mind:
Stage 1: Before dictionary consultation
Stage 2: Macrostructure (searching for the headword)
Stage 3: Microstructure (searching for necessary information under the entry)
Stage 4: Applying the retrieved information to the context for evaluation

Let us discuss the items within each category. Among the four items belonging to Stage 1, it became evident that the easiness parameters varied across descriptors. Item 4 ("I can identify the problem area: a word or a phrase."; beta = 1.44) and Item 1 ("I can decide whether to look up a word in a dictionary or to guess its meaning from the context."; beta = 1.10) were deemed relatively easy, followed by Item 3 ("I can identify a part of speech of the problem word."; beta = 0.69) and Item 2 ("I can choose the appropriate dictionary according to the problem."; beta = −0.07). Notably, Item 2 posed challenges, as it emerged as one of the misfitting items. Given the unpredictable responses from users, revising this descriptor for future surveys is imperative.

An observation worth noting is that Item 2 was perceived as difficult by advanced learners. While less proficient learners may interpret "the appropriate dictionary" simply as choosing between English-Japanese or Japanese-English dictionaries, more advanced learners have a broader array of options, including English monolingual dictionaries, thesauruses, and specialized dictionaries. This expanded choice pool can significantly complicate the decision-making process as proficiency levels increase.

Among the five items pertaining to macrostructure skills, Item 9 ("I can identify the meaning of the problem word when there is more than one meaning in the dictionary, by comparing each one with the context, assisted by example sentences, etc."; beta = 1.69) and Item 6 ("I can predict the base form of the problem word."; beta = 1.59) emerged as the easiest, while Items 5, 8, and 7, focusing primarily on headword selections, exhibited similar difficulty levels (beta = 0.43 to 0.66). Overall, macrostructure skills are generally perceived as easier compared to microstructure skills. However, it is worth noting that Item 9 may present greater difficulty than anticipated, particularly when users need to infer the meaning from the context before consulting a dictionary.

Items 10 to 28 are all associated with microstructure skills. The information sought under the headword, which users found relatively easy, included verb/adjective conjugations (Item 11; beta = 1.35), pronunciation using katakana notations (Item 14; beta = 1.14), synonyms and antonyms (Item 24; beta = 1.06), and translation equivalents (Item 27; beta = 1.06). Items with moderate difficulties encompassed stress symbols (Item 16; beta = 0.28), transitive/intransitive verbs (Item 13; beta = −0.23), countable/uncountable nouns (Item 10; beta = −0.34), stress in compound words (Item 17; beta = −0.34), sentence pattern codes (Item 28; beta = −0.60), verb patterns (Item 23; beta = −0.68), and word origins (Item 25; beta = −0.91). It is noteworthy that Item 14 was identified as a misfitting item. Users may have been confused about the nature of katakana notations. While katakana can be helpful for beginners, certain katakana notations may pose challenges even for advanced learners. Clarifying the descriptor to explicitly define "katakana notations" may alleviate user confusion and enhance comprehension.

The most challenging items included the International Phonetic Alphabet (Item 15; beta = −1.17), register information (spoken vs. written) (Item 19; beta = −1.35), attributive vs. predicative adjectives (Item 12; beta = −1.37), cross-referencing (Item 26; beta = −1.53), speech level (informal vs. formal) (Item 18; beta = −1.66), varieties (UK vs. US) (Item 20; beta = −1.77), and word meaning menu (Item 22; beta = −2.12). While many of these items are typically accompanied by usage labels or grammatical codes, such as [U/C] or [S/W], which may seem straightforward once users are acquainted with their meanings, the results indicate that these nuances are not explicitly taught in classrooms, and users may struggle to leverage this information effectively for understanding word usages.

The shallow nature of dictionary reference acts on smartphones or tablet PCs, coupled with limited exposure to opportunities for maximizing the utility of usage information, underscores the importance of dictionary skill training. Moving forward, it remains an empirical question whether these items are inherently challenging and warrant placement at their current difficulty levels on the latent dimension, or if they could be positioned at points of lesser difficulty. Addressing these questions is essential for enhancing dictionary literacy and optimizing users' utilization of dictionary resources.

Finally, four descriptors address the application of dictionary information to specific language contexts for evaluation. Among them, Items 32 ("I can look for appropriate collocates of the problem word from the collocation box and apply them to produce new sentences."; beta = 0.34), Item 30 ("I can identify verb patterns of the problem word in illustrative examples and apply them to produce new sentences."; beta = 0.22), and Item 29 ("I can identify the grammatical information of the problem word such as pre- and post-modification patterns in illustrative examples and apply them to produce new sentences."; beta = 0.08) were positioned in the middle on the latent dimension.

Conversely, Item 31 was deemed the most challenging (beta = −1.35), focusing on the skill of "identifying collocates of the problem word in illustrative examples and applying them to produce new sentences." It is intriguing to note that users perceived it as more difficult to discern collocation patterns from illustrative examples than to identify collocations in the collocation column. While this discrepancy may not significantly impact advanced users, it could pose substantial challenges for novice to intermediate level learners.

4.3.2 RQ2: Room for improvement in the formulation of "can do" descriptors

Upon examining the nature of descriptors and their relative difficulty levels as determined by both the user questionnaire and Rasch analysis, several suggestions for improvement emerge. Firstly, while the Rasch model effectively scaled the descriptors for L2 dictionary use, instances of user interpretation variance were prevalent, particularly evident in misfitting items such as "the appropriate dictionary," "katakana notations," and "word meaning menu," among others. To mitigate ambiguity, providing concrete examples alongside descriptors may enhance understanding. However, this approach could pose challenges, considering the lack of universally agreed-upon dictionary conventions. It would be beneficial to assess the impact of providing sample entries on the difficulty order, both with and without such examples.

Another pertinent issue is the alignment between users' claims and their actual lookup behaviour. Tono/Kawamoto (in press) investigated this discrepancy by administering actual dictionary lookup activities as tasks to a group of students. Their findings revealed occasional mismatches between user perception and behaviour, demonstrating instances where users were unable to perform tasks they believed they could, as well as instances where they easily executed certain tasks despite claiming inability in the questionnaire. This disparity underscores the influence of prior experience and dictionary skills training in the classroom. Therefore, in creating a list of scaled "can do" descriptors for L2 dictionary use, it is imperative not only to conduct a "can do" questionnaire survey but also to administer performance tests. This dual approach ensures the avoidance of misinterpretations of knowledge and skills illustrated in the descriptors, while verifying users' actual performance in relation to their questionnaire claims.

4.3.3 RQ3: How to link the dictionary skills descriptors to the CEFR levels

An important consideration lies in linking each descriptor to the CEFR levels. One approach involves integrating already calibrated descriptors for communicative language strategies from the CEFR-Common Reference Levels (CEFR-CV) into our descriptor survey as anchor items. By examining the theta scores for these existing strategy descriptors, we can potentially determine the cut-offs for various CEFR levels along the latent dimension.

Given the limited description of dictionary use in the CEFR documents, aligning every descriptor with a specific CEFR level may prove challenging. However, it seems logical to introduce certain lexicographical conventions, such as grammar codes or IPA symbols, after covering foundational English grammar structures and sound systems, typically at the end of the A1 level or the onset of the A2 level. Introducing these meta-cognitive learning devices at the A1 level may not be advisable. Similarly, the choice between monolingual versus bilingual dictionaries should likely be associated with the B levels, as bilingual dictionaries are often the initial option for beginners.

Therefore, in accordance with the original CEFR recommendations (Council of Europe 2020), a combination of intuitive, qualitative, and quantitative methods will be essential to estimate the appropriate CEFR level for each descriptor. Furthermore, linking each dictionary use descriptor with communicative language activities and strategies, particularly those involving reading, writing, and mediation, would be advantageous.

5 Conclusion

The present study embarked on a crucial mission: to develop and calibrate a comprehensive set of "can do" descriptors tailored specifically for users of English as a foreign language, with a paramount focus on optimizing the utilization of various types of dictionary information. While further refinements are undoubtedly necessary, our study unequivocally demonstrated the transformative potential of the Rasch model in enhancing questionnaire survey results, thereby enabling the determination of relative difficulty levels for each descriptor precisely. Examining the nuanced nature and distinctive characteristics of each descriptor, alongside rigorous empirical validation, we envision the evolution of a meticulously curated list of scaled "can do" descriptors for L2 dictionary use. Such a resource holds immense promise, serving as a cornerstone for future discussions and implementations of dictionary skills training grounded in the ongoing refinement of these indispensable descriptors.

Appendix A. A list of "can do" descriptors for dictionary use

Item No.	TYPE	Descriptors in English	Easiness estimate	Std. Error	Lower CI	Upper CI
21	Microstructure	I can identify the plural form of the problem word in the dictionary.	3.094	0.355	2.399	3.79
9	Macrostructure	I can identify the meaning of the problem word when there is more than one meaning in the dictionary, by comparing each one with the context, assisted by example sentences, etc.	1.691	0.226	1.248	2.134
6	Macrostructure	I can predict the base form of the problem word.	1.589	0.22	1.158	2.02
4	Before dictionary consultation	I can identify the problem area: a word or a phrase.	1.444	0.212	1.028	1.861
11	Microstructure	I can look at a dictionary and identify information on conjugational forms of verbs and adjectives.	1.354	0.208	0.946	1.761
14	Microstructure	I can identify the pronunciation of the problem word by looking at the katakana notation in the dictionary and know how to pronounce it.	1.142	0.198	0.753	1.531
1	Before dictionary consultation	I can decide whether to look up a word in a dictionary or to guess its meaning from the context.	1.102	0.197	0.716	1.487
24	Microstructure	I can look at the dictionary and identify synonyms and antonyms of the problem word.	1.062	0.195	0.68	1.445
27	Microstructure	If I am given more than one translation equivalent of the problem word, I can select the most appropriate one by reference to the context.	1.062	0.195	0.68	1.445
3	Before dictionary consultation	I can identify a part of speech of the problem word.	0.698	0.182	0.34	1.055
5	Macrostructure	I can determine which headwords in the dictionary provide information on the problem word or phrase.	0.664	0.181	0.308	1.019

8	Macrostructure	I can search the dictionary for the headword of the problem word and, when there are multiple parts of speech, homonyms, etc., I can use grammatical and semantic clues to identify the headword that provides the correct information.	0.664	0.181	0.308	1.019
7	Macrostructure	I can predict various spellings of words I hear on audio and can look them up in a dictionary to check the spelling.	0.436	0.175	0.092	0.779
32	Application	I can look for appropriate collocates of the problem word from the collocation box and apply them to produce new sentences.	0.343	0.173	0.004	0.682
16	Microstructure	I can identify the position of the stress in the problem word by looking at the stress marks in the dictionary.	0.282	0.172	−0.054	0.618
30	Application	I can identify verb patterns of the problem word in illustrative examples and apply them to produce new sentences.	0.222	0.17	−0.112	0.556
29	Application	I can identify the grammatical information of the problem word such as pre- and post-modification patterns in illustrative examples and apply them to produce new sentences.	0.076	0.168	−0.252	0.404
2	Before dictionary consultation	I can choose the appropriate dictionary according to the problem.	−0.065	0.165	−0.389	0.258
13	Microstructure	I can look at a dictionary and identify whether the problem word is an intransitive or transitive verb.	−0.23	0.163	−0.55	0.089
10	Microstructure	I can look at a dictionary and identify whether the problem word is a countable or uncountable noun.	−0.338	0.162	−0.655	−0.021
17	Microstructure	I can identify the position of the stress in the compound words in question by looking at the stress marks in the dictionary.	−0.338	0.162	−0.655	−0.021
28	Microstructure	I can understand the terms and abbreviations used for the sentence pattern of the problem word in question.	−0.601	0.16	−0.914	−0.287

(continued)

(continued)

Item No.	TYPE	Descriptors in English	Easiness estimate	Std. Error	Lower CI	Upper CI
23	Microstructure	I can identify the sentence pattern of the given verb in question by looking at the verb pattern codes in the dictionary.	−0.678	0.159	−0.991	−0.366
25	Microstructure	I can look at the dictionary and identify word origins of the problem word.	−0.909	0.159	−1.22	−0.597
15	Microstructure	I can identify the pronunciation of the problem word by looking at the International Phonetic Alphabet (IPA) in the dictionary and know how to pronounce it.	−1.165	0.16	−1.478	−0.852
19	Microstructure	I can identify the usage of the problem word (e.g. spoken vs. written) by looking at the usage labels in the dictionary.	−1.346	0.161	−1.661	−1.031
31	Application	I can identify collocates of the problem word in illustrative examples and apply them to produce new sentences.	−1.346	0.161	−1.661	−1.031
12	Microstructure	I can look at a dictionary and identify whether the problem word is an attributive or predicative adjective.	−1.372	0.161	−1.687	−1.056
26	Microstructure	I can recognise the symbols for cross-references relating to the problem word and refer to them as appropriate.	−1.53	0.163	−1.848	−1.211
18	Microstructure	I can identify the usage of the problem word (e.g. informal vs. formal, etc.) by looking at the usage labels in the dictionary.	−1.664	0.164	−1.986	−1.342
20	Microstructure	I can identify the usage of the problem word (e.g. UK vs. US) by looking at the usage labels in the dictionary.	−1.773	0.166	−2.099	−1.448
22	Microstructure	I can find out which meaning of the problem word I should look at from the dictionary's word meaning menu.	−2.118	0.173	−2.457	−1.778

Appendix B. Results of Rasch Model estimation (Dichotomous Model)

Results of RM estimation:

Call: RM(X = dicho)

Conditional log-likelihood: −2739.652
Number of iterations: 27
Number of parameters: 32

Item Easiness Parameters (beta) with 0.95 CI:

	Estimate	Std. Error	lower CI	upper CI
beta ITEM_1	1.102	0.197	0.716	1.487
beta ITEM_2	−0.065	0.165	−0.389	0.258
beta ITEM_3	0.698	0.182	0.340	1.055
beta ITEM_4	1.444	0.212	1.028	1.861
beta ITEM_5	0.664	0.181	0.308	1.019
beta ITEM_6	1.589	0.220	1.158	2.020
beta ITEM_7	0.436	0.175	0.092	0.779
beta ITEM_8	0.664	0.181	0.308	1.019
beta ITEM_9	1.691	0.226	1.248	2.134
beta ITEM_10	−0.338	0.162	−0.655	−0.021
beta ITEM_11	1.354	0.208	0.946	1.761
beta ITEM_12	−1.372	0.161	−1.687	−1.056
beta ITEM_13	−0.230	0.163	−0.550	0.089
beta ITEM_14	1.142	0.198	0.753	1.531
beta ITEM_15	−1.165	0.160	−1.478	−0.852
beta ITEM_16	0.282	0.172	−0.054	0.618
beta ITEM_17	−0.338	0.162	−0.655	−0.021
beta ITEM_18	−1.664	0.164	−1.986	−1.342
beta ITEM_19	−1.346	0.161	−1.661	−1.031
beta ITEM_20	−1.773	0.166	−2.099	−1.448
beta ITEM_21	3.094	0.355	2.399	3.790
beta ITEM_22	−2.118	0.173	−2.457	−1.778

(continued)

	Estimate	Std. Error	lower CI	upper CI
beta ITEM_23	−0.678	0.159	−0.991	−0.366
beta ITEM_24	1.062	0.195	0.680	1.445
beta ITEM_25	−0.909	0.159	−1.220	−0.597
beta ITEM_26	−1.530	0.163	−1.848	−1.211
beta ITEM_27	1.062	0.195	0.680	1.445
beta ITEM_28	−0.601	0.160	−0.914	−0.287
beta ITEM_29	0.076	0.168	−0.252	0.404
beta ITEM_30	0.222	0.170	−0.112	0.556
beta ITEM_31	−1.346	0.161	−1.661	−1.031
beta ITEM_32	0.343	0.173	0.004	0.682
beta ITEM_33	−1.450	0.162	−1.767	−1.134

Appendix C. Item fit statistics (Dichotomous model)

Itemfit Statistics:

	Chisq	Df	p-value	Outfit MSQ	Infit MSQ	Outfit t	Infit t	Discrim
ITEM_1	205.114	211	0.601	0.968	0.969	−0.016	−0.241	0.478
ITEM_2	224.286	211	0.253	1.058	1.138	0.405	1.686	0.417
ITEM_3	166.285	211	0.990	0.784	0.998	−0.914	0.008	0.504
ITEM_4	174.565	211	0.968	0.823	1.005	−0.423	0.079	0.441
ITEM_5	148.134	211	1.000	0.699	0.888	−1.386	−1.161	0.564
ITEM_6	241.305	211	0.075	1.138	1.128	0.477	0.945	0.314
ITEM_7	253.899	211	0.023	1.198	0.964	0.974	−0.374	0.500
ITEM_8	180.573	211	0.937	0.852	0.872	−0.599	−1.343	0.573
ITEM_9	179.523	211	0.943	0.847	0.892	−0.275	−0.737	0.461
ITEM_10	222.936	211	0.273	1.052	1.038	0.393	0.524	0.476
ITEM_11	130.433	211	1.000	0.615	0.935	−1.238	−0.489	0.498
ITEM_12	153.608	211	0.999	0.725	0.844	−1.892	−2.287	0.587
ITEM_13	152.963	211	0.999	0.722	0.836	−1.977	−2.242	0.629
ITEM_14	342.443	211	0.000	1.615	1.339	1.848	2.708	0.180
ITEM_15	232.688	211	0.146	1.098	1.064	0.693	0.918	0.445

(continued)

	Chisq	Df	p-value	Outfit MSQ	Infit MSQ	Outfit t	Infit t	Discrim
ITEM_16	187.207	211	0.879	0.883	0.999	−0.578	0.022	0.518
ITEM_17	188.239	211	0.868	0.888	1.002	−0.741	0.054	0.517
ITEM_18	200.807	211	0.681	0.947	0.860	−0.243	−1.928	0.530
ITEM_19	209.283	211	0.520	0.987	0.968	−0.031	−0.437	0.485
ITEM_20	184.140	211	0.909	0.869	0.945	−0.672	−0.695	0.484
ITEM_21	253.271	211	0.025	1.195	1.133	0.497	0.557	0.196
ITEM_22	307.753	211	0.000	1.452	1.238	1.852	2.623	0.260
ITEM_23	256.844	211	0.017	1.212	1.137	1.474	1.851	0.416
ITEM_24	213.766	211	0.434	1.008	1.046	0.124	0.446	0.403
ITEM_25	321.368	211	0.000	1.516	1.295	3.244	3.868	0.290
ITEM_26	219.633	211	0.327	1.036	1.133	0.268	1.765	0.396
ITEM_27	175.617	211	0.964	0.828	0.872	−0.545	−1.167	0.541
ITEM_28	163.833	211	0.993	0.773	0.873	−1.721	−1.806	0.606
ITEM_29	144.776	211	1.000	0.683	0.810	−2.019	−2.439	0.647
ITEM_30	146.158	211	1.000	0.689	0.797	−1.833	−2.527	0.658
ITEM_31	185.508	211	0.897	0.875	0.944	−0.787	−0.779	0.518
ITEM_32	194.186	211	0.791	0.916	0.903	−0.376	−1.107	0.565
ITEM_33	138.463	211	1.000	0.653	0.802	−2.397	−2.919	0.608

Appendix D. Results of the Partial Credit Model

Results of PCM estimation:

Call: PCM(X = data)

Conditional log-likelihood: −7254.763
Number of iterations: 101
Number of parameters: 127

Item Easiness Parameters (beta) with 0.95 CI:

	Estimate	Std. Error	lower CI	upper CI
beta ITEM_1.c1	2.362	0.465	1.451	3.272
beta ITEM_1.c2	3.263	0.459	2.364	4.163
beta ITEM_1.c3	1.764	0.478	0.828	2.701
beta ITEM_1.c4	−0.688	0.518	−1.703	0.327
beta ITEM_2.c1	1.144	0.267	0.620	1.667
beta ITEM_2.c2	1.125	0.277	0.583	1.668
beta ITEM_2.c3	−0.555	0.317	−1.176	0.065
beta ITEM_2.c4	−3.414	0.410	−4.217	−2.610
beta ITEM_3.c1	2.389	0.426	1.553	3.224
beta ITEM_3.c2	2.820	0.428	1.980	3.659
beta ITEM_3.c3	1.558	0.447	0.683	2.433
beta ITEM_3.c4	−0.696	0.483	−1.642	0.250
beta ITEM_4.c1	1.935	0.450	1.053	2.817
beta ITEM_4.c2	3.239	0.434	2.388	4.089
beta ITEM_4.c3	1.686	0.455	0.794	2.578
beta ITEM_4.c4	−0.758	0.498	−1.733	0.218
beta ITEM_5.c1	2.581	0.453	1.692	3.470
beta ITEM_5.c2	3.072	0.454	2.182	3.962
beta ITEM_5.c3	1.429	0.475	0.498	2.360
beta ITEM_5.c4	−1.053	0.519	−2.070	−0.037
beta ITEM_6.c1	3.457	0.758	1.971	4.943
beta ITEM_6.c2	4.695	0.749	3.227	6.163
beta ITEM_6.c3	3.454	0.752	1.980	4.928
beta ITEM_6.c4	0.869	0.773	−0.646	2.384
beta ITEM_7.c1	1.643	0.331	0.995	2.292
beta ITEM_7.c2	2.149	0.330	1.502	2.796
beta ITEM_7.c3	−0.097	0.374	−0.830	0.636
beta ITEM_7.c4	−3.088	0.477	−4.022	−2.154
beta ITEM_8.c1	1.765	0.357	1.065	2.465
beta ITEM_8.c2	2.417	0.354	1.723	3.111
beta ITEM_8.c3	0.497	0.387	−0.263	1.256
beta ITEM_8.c4	−2.479	0.475	−3.410	−1.547
beta ITEM_9.c1	2.949	0.640	1.693	4.204

(continued)

	Estimate	Std. Error	lower CI	upper CI
beta ITEM_9.c2	4.143	0.632	2.905	5.381
beta ITEM_9.c3	3.500	0.635	2.256	4.744
beta ITEM_9.c4	0.809	0.662	−0.489	2.107
beta ITEM_10.c1	0.920	0.243	0.444	1.397
beta ITEM_10.c2	0.543	0.264	0.026	1.059
beta ITEM_10.c3	−1.094	0.314	−1.710	−0.479
beta ITEM_10.c4	−2.429	0.337	−3.089	−1.769
beta ITEM_11.c1	4.347	1.030	2.327	6.366
beta ITEM_11.c2	5.324	1.021	3.323	7.324
beta ITEM_11.c3	4.061	1.017	2.068	6.055
beta ITEM_11.c4	1.871	1.021	−0.129	3.872
beta ITEM_12.c1	1.158	0.223	0.722	1.595
beta ITEM_12.c2	−0.003	0.257	−0.507	0.500
beta ITEM_12.c3	−2.384	0.329	−3.029	−1.740
beta ITEM_12.c4	−5.457	0.475	−6.388	−4.526
beta ITEM_13.c1	1.038	0.253	0.541	1.534
beta ITEM_13.c2	0.797	0.269	0.269	1.325
beta ITEM_13.c3	−0.704	0.308	−1.308	−0.100
beta ITEM_13.c4	−3.048	0.371	−3.775	−2.320
beta ITEM_14.c1	1.503	0.374	0.770	2.236
beta ITEM_14.c2	2.464	0.362	1.755	3.173
beta ITEM_14.c3	1.362	0.383	0.610	2.113
beta ITEM_14.c4	−0.940	0.427	−1.777	−0.103
beta ITEM_15.c1	0.442	0.202	0.045	0.839
beta ITEM_15.c2	−0.516	0.237	−0.980	−0.052
beta ITEM_15.c3	−2.321	0.289	−2.887	−1.756
beta ITEM_15.c4	−5.485	0.430	−6.327	−4.643
beta ITEM_16.c1	0.888	0.274	0.351	1.425
beta ITEM_16.c2	1.176	0.276	0.635	1.716
beta ITEM_16.c3	−0.238	0.313	−0.851	0.376
beta ITEM_16.c4	−2.244	0.360	−2.949	−1.539
beta ITEM_17.c1	1.145	0.254	0.647	1.643
beta ITEM_17.c2	0.856	0.270	0.326	1.386

(continued)

	Estimate	Std. Error	lower CI	upper CI
beta ITEM_17.c3	−1.024	0.321	−1.653	−0.394
beta ITEM_17.c4	−3.034	0.369	−3.757	−2.311
beta ITEM_18.c1	1.274	0.223	0.837	1.711
beta ITEM_18.c2	−0.118	0.261	−0.629	0.392
beta ITEM_18.c3	−2.816	0.352	−3.505	−2.127
beta ITEM_18.c4	−5.981	0.527	−7.013	−4.948
beta ITEM_19.c1	1.044	0.219	0.616	1.473
beta ITEM_19.c2	0.017	0.249	−0.471	0.505
beta ITEM_19.c3	−2.828	0.347	−3.507	−2.148
beta ITEM_19.c4	−5.973	0.523	−6.998	−4.948
beta ITEM_20.c1	0.442	0.190	0.071	0.814
beta ITEM_20.c2	−0.968	0.235	−1.428	−0.508
beta ITEM_20.c3	−3.412	0.319	−4.038	−2.787
beta ITEM_20.c4	−6.726	0.512	−7.731	−5.722
beta ITEM_21.c1	3.449	1.098	1.297	5.602
beta ITEM_21.c2	5.858	1.069	3.762	7.954
beta ITEM_21.c3	5.450	1.063	3.367	7.533
beta ITEM_21.c4	3.884	1.058	1.810	5.958
beta ITEM_22.c1	0.163	0.179	−0.188	0.513
beta ITEM_22.c2	−1.476	0.232	−1.931	−1.021
beta ITEM_22.c3	−4.056	0.328	−4.699	−3.412
beta ITEM_22.c4	−8.048	0.668	−9.358	−6.738
beta ITEM_23.c1	0.875	0.230	0.425	1.326
beta ITEM_23.c2	0.401	0.250	−0.089	0.891
beta ITEM_23.c3	−1.891	0.319	−2.516	−1.265
beta ITEM_23.c4	−4.084	0.381	−4.831	−3.337
beta ITEM_24.c1	1.684	0.382	0.935	2.434
beta ITEM_24.c2	2.623	0.373	1.893	3.354
beta ITEM_24.c3	1.162	0.398	0.381	1.943
beta ITEM_24.c4	−1.082	0.440	−1.945	−0.219
beta ITEM_25.c1	1.535	0.257	1.031	2.039
beta ITEM_25.c2	0.776	0.280	0.227	1.325
beta ITEM_25.c3	−1.601	0.343	−2.273	−0.929

(continued)

	Estimate	Std. Error	lower CI	upper CI
beta ITEM_25.c4	−4.614	0.471	−5.537	−3.691
beta ITEM_26.c1	0.587	0.198	0.199	0.976
beta ITEM_26.c2	−0.549	0.234	−1.007	−0.091
beta ITEM_26.c3	−3.279	0.331	−3.927	−2.631
beta ITEM_26.c4	−6.696	0.550	−7.774	−5.617
beta ITEM_27.c1	1.548	0.371	0.821	2.274
beta ITEM_27.c2	2.467	0.360	1.761	3.173
beta ITEM_27.c3	1.250	0.382	0.501	2.000
beta ITEM_27.c4	−1.515	0.445	−2.388	−0.643
beta ITEM_28.c1	1.168	0.246	0.685	1.651
beta ITEM_28.c2	0.793	0.263	0.278	1.308
beta ITEM_28.c3	−1.750	0.334	−2.404	−1.096
beta ITEM_28.c4	−4.713	0.463	−5.620	−3.805
beta ITEM_29.c1	1.932	0.333	1.280	2.584
beta ITEM_29.c2	2.041	0.339	1.375	2.706
beta ITEM_29.c3	−0.127	0.381	−0.874	0.619
beta ITEM_29.c4	−3.133	0.482	−4.079	−2.188
beta ITEM_30.c1	1.555	0.310	0.949	2.162
beta ITEM_30.c2	1.842	0.313	1.229	2.456
beta ITEM_30.c3	−0.244	0.356	−0.941	0.453
beta ITEM_30.c4	−3.311	0.465	−4.222	−2.399
beta ITEM_31.c1	1.523	0.245	1.044	2.003
beta ITEM_31.c2	0.420	0.274	−0.116	0.956
beta ITEM_31.c3	−2.343	0.360	−3.047	−1.638
beta ITEM_31.c4	−5.535	0.533	−6.580	−4.489
beta ITEM_32.c1	1.667	0.252	1.173	2.162
beta ITEM_32.c2	0.488	0.281	−0.063	1.039
beta ITEM_32.c3	−2.361	0.367	−3.080	−1.642
beta ITEM_32.c4	−7.547	1.073	−9.650	−5.444

Appendix E. The Item fit statistics for the Partial Credit Model

Itemfit Statistics:

	Chisq	df	p-value	Outfit MSQ	Infit MSQ	Outfit t	Infit t	Discrim
ITEM_1	231.067	222	0.324	1.036	1.036	0.410	0.414	0.594
ITEM_2	275.753	222	0.008	1.237	1.220	2.456	2.313	0.536
ITEM_3	197.522	222	0.880	0.886	0.878	−1.283	−1.392	0.682
ITEM_4	212.201	222	0.670	0.952	0.958	−0.462	−0.419	0.625
ITEM_5	196.273	222	0.892	0.880	0.892	−1.314	−1.201	0.666
ITEM_6	202.042	222	0.828	0.906	0.919	−1.009	−0.872	0.635
ITEM_7	204.846	222	0.789	0.919	0.936	−0.808	−0.640	0.620
ITEM_8	186.398	222	0.961	0.836	0.850	−1.754	−1.623	0.676
ITEM_9	213.859	222	0.640	0.959	0.944	−0.423	−0.584	0.617
ITEM_10	283.557	222	0.003	1.272	1.158	2.596	1.699	0.619
ITEM_11	180.763	222	0.980	0.811	0.800	−2.195	−2.385	0.700
ITEM_12	176.444	222	0.989	0.791	0.829	−2.306	−1.868	0.700
ITEM_13	214.745	222	0.624	0.963	0.966	−0.383	−0.358	0.668
ITEM_14	297.426	222	0.001	1.334	1.318	3.287	3.169	0.483
ITEM_15	257.095	222	0.053	1.153	1.133	1.584	1.436	0.591
ITEM_16	211.348	222	0.685	0.948	0.943	−0.551	−0.617	0.688
ITEM_17	196.650	222	0.889	0.882	0.887	−1.304	−1.278	0.706
ITEM_18	218.775	222	0.549	0.981	0.950	−0.157	−0.473	0.634
ITEM_19	217.598	222	0.571	0.976	0.965	−0.223	−0.332	0.634
ITEM_20	209.570	222	0.715	0.940	0.955	−0.607	−0.450	0.648
ITEM_21	196.796	222	0.887	0.882	0.896	−1.290	−1.195	0.651
ITEM_22	323.997	222	0.000	1.453	1.467	4.117	4.328	0.380
ITEM_23	242.551	222	0.164	1.088	1.123	0.945	1.327	0.588
ITEM_24	206.002	222	0.772	0.924	0.937	−0.797	−0.661	0.651
ITEM_25	285.973	222	0.002	1.282	1.305	2.755	2.959	0.456
ITEM_26	219.143	222	0.542	0.983	1.020	−0.153	0.237	0.598
ITEM_27	196.411	222	0.891	0.881	0.874	−1.291	−1.374	0.663
ITEM_28	179.227	222	0.984	0.804	0.818	−2.203	−2.033	0.695
ITEM_29	170.074	222	0.996	0.763	0.763	−2.693	−2.711	0.737

(continued)

	Chisq	df	p-value	Outfit MSQ	Infit MSQ	Outfit t	Infit t	Discrim
ITEM_30	164.338	222	0.999	0.737	0.741	−3.013	−3.003	0.751
ITEM_31	201.314	222	0.837	0.903	0.897	−0.983	−1.041	0.652
ITEM_32	167.007	222	0.998	0.749	0.761	−2.864	−2.735	0.712

6 Bibliography

6.1 Dictionaries and other sources

CIDE = *Cambridge international dictionary of English*. Ed. by Procter, Paul. Cambridge: Cambridge University Press. 1995.

CEFR = *The common European framework of reference for languages: learning, teaching, assessment*. Ed. by Council of Europe. 2001 [https://rm.coe.int/1680459f97;lastaccess:21.05.2023].

CEFR/CV = *The common European framework of reference for languages: learning, teaching, assessment: companion volume*. Ed. by Council of Europe. 2020 [https://rm.coe.int/common-european-frame-work-of-reference-for-languages-learning-teaching/16809ea0d4;lastaccess:21.05.2023].

COBUILD = *Collins COBUILD English dictionary*. Ed. by Sinclair, John M. London: Harper Collins Publisher. 1995 [1987 Ed. J.M.Sinclair].

DIALANG = A diagnostic language assessment system for adult learners. Författare: A. Huhta, S. Luoma, M. Oscarson, K. Sajavaara, S. Takala, A. Teasdale. I Council of Europe,Common European Framework of Reference for Languages: Learning,Teaching, Assessment. Case Studies. (2002) Strasbourg: Council of Europe Publishing.

ISED = *Idiomatic syntactic English dictionary*. Ed. by Hornby, Albert S., Gatenby, E.V., and Wakefield, H. Tokyo: Kaitakusha 1943.

LDOCE = *Longman dictionary of contemporary English*. Ed. by Hornby, Albert S. Oxford: Oxford University Press 1995 [1948 Comp. A.S. Hornby].

MEDAL = *Macmillan English dictionary for advanced learners*. Ed. by Rundell, Michael. Oxford: Macmillan Education 2002.

OALD = *Oxford advanced learner's dictionary*. Ed. by Crowther, Jonathan. Oxford: Oxford University Press 1995 [1948 Comp. A. S. Hornby].

6.2 Academic literature

Andersen, Erling B. (1972): The numerical solution of a set of conditional estimation equations. In: *Journal of the Royal Statistical Society* 34:1, 42–54 [https://doi.org/10.1111/j.2517-6161.1972.tb00887.x].

Atkins, Beryl T. Sue (1996): Bilingual dictionaries: past, present and future. In: Gellerstam Martin/Järborg, Jerker/Malmgren, Sven-Göran/Norén, Kerstin/Rogström, Lena/Papmehl, Catarina Röjder (eds.): *Euralex '96 proceedings*. Göteborg University, 515–546.

Atkins, Beryl T. Sue/Varantola, Krista (1997): Monitoring dictionary use. In: *International Journal of Lexicography* 10:1, 1–45.

Béjoint, Henri (1981): The foreign student's use of monolingual English dictionaries: A study on language needs and reference skills. In: *Applied Linguistics* 2:3, 207–222.
Béjoint, Henri (1994): *Tradition and innovation in modern English dictionaries*. Oxford: Oxford University Press.
Béjoint, Henri/Moulin, André (1987): The place of the dictionary in an EFL programme. In: Cowie, Anthony P. (ed.): *The dictionary and the language learner: Papers from the EURALEX seminar at the University of Leeds*. Tübingen: Niemeyer, 97–113.
Cowie, Anthony P. (2000): *English dictionaries for foreign learners: a history*. Oxford: Oxford University Press.
Gavriilidou, Zoe (2013): Development and validation of the strategy inventory for dictionary use (S.I.D.U.). In: *International Journal of Lexicography* 26:2, 135–153.
Gavriilidou, Zoe (2014): Translation, cultural adaptation and preliminary psychometric evaluation of the English version of 'Strategy Inventory for Dictionary Use' (S.I.D.U.). In: Abel, Andrea/Viettori, Chiara/Ralli, Natascia (eds.): *Proceedings of the XVI Euralex international congress*. Institute for Specialised Communication and Multilingualism, 225–235.
Gavriilidou, Zoe/Mavrommatidou, Stavroula (2016): Construction of a tool for the identification of electronic dictionary users' skills: test specification and content validity. In: *Proceedings of the XVII Euralex International Congress*. Tbilisi, 168–178.
Harvey, Keith/Yuill, Deborah (1997): A study of the use of a monolingual pedagogical dictionary by learners of English engaged in writing. In: *Applied Linguistics* 18:3, 253–278.
Hartmann, Reinhard R.K. (1999): Case study: The Exeter university survey of dictionary use. In: Hartman-n,Reinhard R.K. (ed.): *Dictionaries in language learning*. Berlin, Freie Universität Berlin, 36–52.
Hartmann, Reinhard R. K. (2001): *Teaching and researching lexicography*. Harlow et al.: Longman.
Hartmann, Reinhard R.K./James, Gregory (2002): *Dictionary of lexicography*. London: Routledge.
Higuchi, Masayuki (2012): *Eigo jiten katuyou gaido – jisho no joho wo yomitoru tame no hissu chisiki* (A guide to English dictionaries). Kaitakusha.
Kawamoto, Naho/Tono, Yukio (2023): Developing illustrative "can do" descriptors for L2 dictionary use: a preliminary version. In: *Asialex 2023 Proceedings*. Yonsei University, 103–112.
Kosem, Iztok/Lew, Robert/Müller-Spitzer, Carolin/Silveira, Maria Ribeiro/Wolfer, Sascha (2019): The image of the monolingual dictionary across Europe. The results of the European survey of dictionary use and culture. In: *International Journal of Lexicography* 32:1, 92–114.
Koyama, Toshiko (2013): Enhancing learners' e-dictionary skills through strategy training. In: Kwary,Deny A./Musyhda,Lilla/Wulan,Nur (eds.): *Lexicography and dictionaries in the information age: Selected papers from the 8th ASIALEX International Conference*, *Bali*, *Indonesia*, Airlangga University, 173–178.
Lew, Robert (2013): Online dictionary skills. In: Gantar, Polona/Kallas, Jelena/Kosem, Iztok/Krek, Simon/Langemets, Margit/Tuulik, Maria (eds.): *Electronic lexicography in the 21st Century: Thinking Outside the Paper*. Institute for Applied Slovene Studies/Eesti Keele Instituut, 16–31.
Lew, Robert/Galas, Katarzyna (2008): Can dictionary skills be taught? The effectiveness of lexicographic training for primary-school-level Polish learners of English. In: *Proceedings of the XIII EURALEX International Congress*. Universitat Pompeu Fabra, 1273–1285.
Lew, Robert/Mitton, Roger (2011): Not the word I wanted? How online English learners' dictionaries deal with misspelled words. In: *Proceedings of eLex 2011*. Institute for Applied Slovene Studies, 165–174.
Mair, Patrick/Hatzinger, Reinhold (2007): Extended Rasch modeling: the eRm package for the application of IRT models in R. In: *Journal of Statistical Software* 20:9: 1–20.
Mavrommatidou, Stavroula/Gavriilidou, Zoe/Markos, Angelos (2019): Development and validation of the strategy inventory for electronic dictionary use (SIEDU). In: *International Journal of Lexicography* 32:4, 393–410.
Nesi, Hilary (1999): The specification of dictionary reference skills in higher education. In: Hartmann, R.R.K. (ed.): *Dictionaries in Language Learning*. Berlin, Freie Universität Berlin, 53–67.
Nesi, Hilary (2015): Dictionary use by English language learners. In: *Language Teaching* 47:1, 38–55.

Nesi, Hilary/Meara, Paul (1994): Patterns of misinterpretation in the productive use of EFL dictionary definitions. In: *System* 22:1, 1–15.
Neubach, Abigail/Cohen, Andrew D. (1988): Processing strategies and problems encountered in the use of dictionaries. In: *Dictionaries: The Journal of the Dictionary Society of North America* 10, 1–19.
North, Brian/Piccardo, Enrica (2019): Developing new CEFR descriptor scales and expanding the existing ones: constructs, approaches and methodologies [https://www.dgff.de/assets/Uploads/ZFF-2-2019-01-North-Piccardo.pdf;lastaccess:21.05.2023].
Nuccorini, Stefania (1992): Monitoring dictionary use. In: Tommola, Hannu/Varantola, Krista/Salmi-Tolonen, Tarja/Schopp, Jürgen (eds.): *EURALEX '92 Proceedings*. University of Tampere, 89–102.
Padgett, R. Noah/Morgan, Grant (2020): Using the *eRm* package for Rasch modeling. In: *Measurement Interdisciplinary Research and Perspectives*, July 2020 [DOI: 10.1080/15366367.2020.1732155].
Scholfield, Phil (1982): Using the English dictionary for comprehension. In: *TESOL Quarterly* 16:2, 185–194.
Tono, Yukio (2001): *Research on dictionary use in the context of foreign language learning*. Tübingen: Niemeyer.
Tono, Yukio/Kawamoto, Naho (in press): Developing illustrative "Can Do" descriptors for dictionary use by foreign language learners. In: *LEXICOGRAPHY: Journal of ASIALEX*.
Winkler, Birgit (2001): Students working with an English learner's dictionary on CD-ROM. In: *Information technology and multimedia in English language teaching conference, Hong Kong, June 1–2, 2001*. Hong Kong, The Hong Kong Polytechnic University, 227–254.
Wingate, Ursula (2004): Dictionary use – the need to teach strategies. In: *Language Learning Journal* 29, 5–11.