Yumsek *Language Testing in Asia* (2023) 13:3
https://doi.org/10.1186/s40468-022-00214-0

Language Testing in Asia

RESEARCH Open Access

Educational L2 constructs and diagnostic measurement

Meltem Yumsek*

*Correspondence:
myumsek@gmail.com

Ministry of Education,
06500 Ankara, Turkey

Abstract

Various stakeholders, such as policymakers and educators, require diagnostic feedback and actionable test results. One particular context in which fine-grained test results are of utmost importance is English language proficiency (ELP) assessments as they are used for critical decisions for students. Diagnostic classification models (DCMs) afford finer levels of feedback to improve learning outcomes. This study is part of a research project which implemented DCMs to the reading domain of a K–12 ELP assessment for grades 6–8 to evaluate its feasibility and utility. The research project is U.S.-based. However, the investigation is of relevance to the language testing community broadly. The paper discusses how to address second language reading construct within DCMs. Specifically, it details the identification of attributes underlying the construct and the development and selection of alternative *Q*-matrices.

Keywords: Diagnostic classification models, English learners, English proficiency assessments, K–12, Second language reading

Introduction

Increasingly, large-scale proficiency testing is used for accountability purposes (Haertel, 1999; Haertel & Herman, 2005) and as an instrument of educational reform across the world (Chalhoub-Deville, 2016). Diagnostic information is key in reform-oriented accountability tests to support the desired learning and teaching goals. The call for diagnostic results is manifested in educational laws as well. One example is the Every Student Succeeds Act (ESSA, 2015) in the U.S. where the law requires statewide assessments to "produce individual student interpretive, descriptive, and diagnostic reports…" (Section 111 [b][2][B][x]). In addition to policy demands, stakeholders such as educators favor access to diagnostic information that enhance teaching and learning effectiveness (Huff & Goodman, 2007; Kim et al., 2016; Lopez, 2019). Given the legal enforcement and stakeholders' interest, more efforts to yield diagnostic and instructionally-relevant results from large-scale tests are warranted.

The predominant measurement models that are used to estimate students' proficiency do not inherently provide the level of detail for teaching and learning (de la Torre, 2009). In some testing programs, subscores are reported to give more diagnostic feedback (de la Torre, 2009; Haberman, 2008; Kunnan & Jang, 2009). The subscores are generally

associated with language domains in language assessments. Yet, subscores themselves might not provide reliable information beyond the overall score (Haberman, 2008).

Diagnostic Classification Models (DCMs, Rupp et al., 2010) have emerged to purvey finer levels of information (de la Torre et al., 2010; de la Torre & Minchen, 2014). DCMs are multidimensional models which allow examining knowledge, abilities, processes, and strategies underlying the test performance (de la Torre, 2009; Yang & Embretson, 2007). These components are collectively referred to as attributes (de la Torre et al., 2010; de la Torre & Chiu, 2016; Henson, 2009; Rupp et al., 2010). Each test item is assumed to measure one or more attributes. DCMs estimate whether students possess the attributes or not (i.e., for binary attributes). Therefore, in addition to student responses, item-attribute information is necessary for the DCM estimation. This information is recorded in a *Q*-matrix (Tatsuoka, 1983) by experts. The *Q*-matrix is analogous to a factor structure. It shows how each item relates to an attribute. More specifically if an item is associated with an attribute, 1 is recorded in the *Q*-matrix. If an item is not associated with an attribute, 0 is recorded in the matrix.

DCMs have more utility because they allow a general ability to be defined at finer levels and yield information that hints areas where students need help (Templin & Hoffman, 2013). This, in turn allows class time and resources to be utilized more efficiently to address group-level problem areas (de la Torre, 2009; de la Torre & Minchen, 2014; Liu et al., 2018; Sessoms & Henson, 2018). In this way, learning opportunities are enhanced. DCMs also yield other useful information about the construct such as attribute hierarchies and relations which might help develop learning trajectories, and curricula (Templin & Bradshaw, 2014) or construct theories (Rupp et al., 2010).

A common area of DCM applications is second language (L2) ability. Although there has been considerable work on DCMs and L2 admission tests (e.g., Kim, 2015; Ravand, 2016; Sawaki et al., 2009; von Davier, 2008), DCMs for K–12 language assessments are scarce despite the great need. For instance, in the U.S. English learners (ELs) are students who immigrated to the U.S. or those born in the U.S. but speak a language other than English, and have difficulties in succeeding at schools due to their language proficiency. These students are required to take an English Language Proficiency (ELP) assessment until they attain required proficiency levels (ESSA, 2015). When students fail to pass the exam, they are placed in language support programs. ELP assessments are used for critical decisions such as classification and placement in instructional programs. Therefore, it is of utmost importance to provide detailed feedback about language development so that the achievement gap between L2 and L1 learners emphasized by various researchers (e.g., Deville & Chalhoub-Deville, 2011) can be closed.

The study is part of a U.S.-based research project which applied the DCM methodology to the reading domain of a large-scale assessment. This paper showcases how the reading construct in a K–12 ELP assessment was tackled within the DCM framework. The paper explicates a methodology for *Q*-matrix development and selection because a *Q*-matrix is a critical element of DCMs. The paper aims to discuss the identification of attributes underlying the L2 reading construct, the development of alternative *Q*-matrices, and the selection of the most suitable *Q*-matrix. The study employs a statistical validation approach for *Q*-matrix development and a holistic approach for *Q*-matrix selection. Given the purposes, the paper addresses the following research questions: (1)

What key attributes are represented in the reading section of a K-12 assessment? (2) Does a standards-based or expert-defined *Q*-matrix show a better fit?

Literature review

Dimensionality of the L2 reading construct

Because the multidimensionality of the construct is an important assumption under DCMs, the review of the literature will address the dimensionality of L2 reading construct. The debate regarding the divisibility and components of L2 reading is ongoing (Alderson, 2000; Weir, 2005). Although some research studies show that L2 reading construct cannot be separated properly (e.g., Alderson, 1990; Alderson & Lukmani, 1989; Rost, 1993), the unitary perspective is also found inadequate mainly because it can lead to issues of construct underrepresentation whereby critical reading components are ignored (Alderson, 2000; Urquhart & Weir, 1998). The field at large seems to favor a multi-component representation of the L2 reading construct. It is claimed that L2 reading includes several subskills (Koda, 2007, 2012), strategies (Weir, 2005), or a combination of processes, knowledge, and abilities (Hudson, 1996). Different taxonomies of reading skills were developed for test development (e.g., Davis, 1968). Empirical research has been undertaken to uncover the underlying L2 reading dimensions. Methodologies such as factor analysis, rule space model, verbal reports, and eye-tracking studies have been employed (e.g., Kim, 2009; Song, 2008; Buck et al., 1997; Brunfaut & McCray, 2015). Weir (2005) also argues that reading subcomponents are acknowledged in teaching and should be incorporated into testing. Furthermore, he posits that reporting reading as a separate domain requires the decomposition of this construct to distinguish it from other domains (Urquhart & Weir, 1998; Weir, 2005). Rather than overlooking the attributes, profiling them to uncover weaknesses seem to be a reasonable resolution (Urquhart & Weir, 1998). In conclusion, the consensus in the literature seems to favor a multidimensional representation of L2 reading.

Application of DCMs to L2 reading assessments

This multidimensional perspective of L2 reading has motivated DCM applications for L2 reading assessments. However, when these assessments are not developed for diagnostic purposes from the onset, DCMs are retrofitted to test data. In other words, dimensions or attributes are identified **ad hoc** as well as the *Q*-matrix which shows the relationship between each item and attribute. Although retrofitting presents some limitations such as low dimensionality, broadly defined attributes, or lack of a theory to determine attributes (Haberman & von Davier, 2007; Gierl & Cui, 2008; Deonovic et al., 2019), it can still provide useful detailed feedback. Attributes in a retrofitting context can be specified through alternative methods or a combination of them including expert input, verbal protocols with students, eye-tracking studies, and process data (Rupp et al., 2010). The reported literature shows that different methodologies have been utilized in different studies and no standard procedures have emerged (Kim, 2015).

A few studies merely detail attribute and *Q*-matrix development for L2 reading assessments. Both Jang (2009) and Sawaki et al. (2009) focus on the Test of English as a Foreign Language (TOEFL iBT) reading. Li and Suen (2013) explore the reading domain of the Michigan English Language Assessment Battery (MELAB). Sawaki et al. include

broader attributes (i.e., word meaning, specific information, connecting information, synthesizing and organizing) by reviewing test specifications and subskills reported in the literature. They also conduct substantive item analysis with content developers. The *Q*-matrix is coded by content experts and psychometricians in their research. In her study, Jang develops a more inclusive approach and draws from multiple sources including literature, test blueprints, assessments frameworks, statistical item, and dimensionality analysis, as well as textual analysis (e.g., word frequency, text length, rhetorical structure). Moreover, she incorporates think-aloud protocols to capture the actual processes endorsed by students and seeks to confirm the attributes identified by experts. Finer attributes (i.e., context-dependent vocabulary, context-independent vocabulary, syntactic and semantic links, explicit information, implicit information, inferencing, summarizing, mapping contrasting ideas into a framework) emerge in her study. Li and Suen also conduct verbal protocols with students to verify the initial set of skills identified by experts. They include vocabulary, syntax, extracting explicit information, and understanding implicit information as attributes for MELAB reading. All of these studies include *Q*-matrix validation based on just DCM parameters (e.g., combining skills, fixing item parameters).

Although all of the studies focus on the L2 reading construct and Sawaki et al. and Jang use the same test, the number and scope of attributes vary. Some variation might be expected, as reading comprehension is complex. The granularity of attributes should be aligned with the purposes of diagnostic information, theory, and estimation requirements (Rupp et al., 2010). Another aspect that is worth mentioning about the studies is the composition of the panel. Involving qualified experts in the identification of attributes and development of a Q-matrix is critical for successful implementation (Rupp et al., 2010). The expert panels in the studies reviewed tend to consist of graduate students, which can be attributed to the ease of access to this population. However, involving various experts, specifically the test developers, teachers, domain experts, and psychometricians is important due to their familiarity with the test, test-taker population, and DCM methodology (Kunina-Habenicht et al., 2012; Madison & Bradshaw, 2015). The purpose of this paper is to showcase a comprehensive methodology in specifying attributes and the *Q*-matrix. For this purpose, a panel of language testing experts including members from the test developer is convened and alternative *Q*-matrices and an empirical validation method are integrated to select an appropriate *Q*-matrix for DCM methodology.

The DCM used in the study

As mentioned earlier, a DCM application requires student responses to test items and a *Q*-matrix showing item-attribute relations. There are various models and they differ from each other with respect to attribute characteristics (e.g., binary vs. non-binary; compensatory vs. conjunctive). In general, a selected model posits whether each item measures the specified attributes in the *Q*-matrix. It estimates the probability of knowing attributes. Based on the status of each attribute (e.g., mastery vs. non-mastery), a student is assigned to a class. A class is like a profile showing the attributes a student attained or not. For instance, in the presence of two attributes, there will be four different profiles:

(1) mastery of the first attribute, (2) mastery of the second attribute (3) mastery of both attributes, or (4) non-mastery of the attributes.

In this study, the log-linear cognitive diagnostic model (LCDM; Henson et al., 2009) is used. In this model, the probability of a correct response is mathematically expressed as

$$\left(X_{ij}=1 \mid a_c\right)=\frac{\exp\left(\lambda_{0,j}+\lambda_j^T\ h\ \left(\alpha_c,q_j\right)\right)}{1+\exp\left(\lambda_{0,j}+\lambda_j^T\ h\ \left(\alpha_c,q_j\right)\right)} \tag{1}$$

where $\lambda_{0,\,j}$ represents the probability of a correct response for examinees who have not mastered any of the attributes. λ_j^T represents the main effects and their interactions. The combinations of main effects and their interactions are expressed by $h\ (\alpha_c,q_j)$. These terms can be rewritten as

$$\lambda_j^T\ h\ \left(\alpha_c,q_j\right)=\sum_{k=1}^{K}\lambda_{j,1,(k)}\alpha_{ck}q_{jk}+\sum_{k=1}\sum_{k'>1}^{K}\lambda_{j,2,(k,k')}\alpha_{ck}\ \alpha_{ck'}q_{jk}\ q_{jk'}+\ldots \tag{2}$$

The first term in Eq. 2 is the main effect and shows the increase in the correct response probability for mastering the given attribute (Madison & Bradshaw, 2015). The second term is the two-way interaction which indicates the increase in the correct response probability for possessing all required attributes (Madison & Bradshaw, 2015). The main advantage of the LCDM is that it does not require specifying attribute interactions (i.e., compensatory or conjunctive relationship) a priori (Henson et al., 2009; Rupp et al., 2010) and allows the relationship between attributes and the observed outcome to be items specific (Rupp et al., 2010). Different submodels may apply to different items (Rupp & Templin, 2011). Thus, this general model was preferred over other models due to, as explained earlier, a lack of comprehensive theory about L2 reading construct and attribute relations.

Methodology

Assessment and the participants

A large-scale K–12 ELP assessment measuring the social and academic language was used in this study. The assessment is used in the U.S. schools. ELP assessments are required by law to monitor language development of ELs for accountability purposes in the U.S. Scores are also anticipated to impact instructional decisions if students are placed in language support programs. The ELP assessment used in the study is also a standards-based assessment. The standards are akin to broad objectives. In this case, they describe the language necessary for academic success at schools. They guide language instruction as well as the assessments. The standards are accompanied by can-do-descriptors and language functions that outline students' use of language in different situations. The language functions are derived from the standards, teaching materials, and literature.

The ELP assessment measures ELs' language development in four domains (e.g., reading, listening, speaking, and writing). The reading domain for grade 6–8 ELs was selected for DCM analysis. The test form included 27 items organized around short paragraphs (50–130 words) or testlets (210–270 words) with a specific theme. The texts

Table 1 Attributes in the standards-based *Q*-matrix

Standards	Social Instructional Language
	Language of Language Arts
	Language of Math
	Language of Science
	Language of Social Studies
Key uses	Recount
	Explain
	Argue

accompanying the items also varied in types such as informative texts and stories. All items were multiple choice and scored dichotomously. A total of 23,942 ELs responded to the reading items. While about 40% of ELs were enrolled in grade 6, there were about 30% of ELs in each of grades 7 and 8. The student group was balanced concerning gender distribution with 46% of girls and 54% of boys. The majority of the ELs (about 80%) had a Hispanic background.

Q-matrix development

DCM was retrofitted to the test data and actual attributes underlying the test were not known. Thus, two alternative *Q*-matrices were developed in search of the most feasible solution to explain performance.

Standards-based Q-matrix

Test specifications are recommended as a practical and reasonable starting point for *Q*-matrix development (Li & Suen, 2013). The interpretations of the dimensions in the test specifications are also straightforward as they are familiar to the test users. In this respect, "standards" and "language functions" from the item specifications by the test developer, shown in Table 1, were used to build the first *Q*-matrix. Test items were designed to measure five standards that represent EL's ability to communicate in English in social and instructional settings and four academic content areas (math, science, language arts, and social studies). Moreover, three language functions that are associated with processes ELs engage with characterize the items. These functions are *Recount*, *Explain* and *Argue*. *Recount* relates to identifying main ideas, and details or summarizing information. *Explain* is about reasoning, sequencing, or comparing and contrasting information while *Argue* is concerned with identifying evidence or differentiating facts from opinions (see WIDA, 2012, 2016 for more information about the dimensions). Five standards and three language functions were combined to represent the attributes in this *Q*-matrix for theoretical and statistical reasons. Students are not likely to engage with an item by isolating a function from the targeted standard. Each item was also associated with one standard and language function. Using only standards or functions would result in a simple structure and cause information loss (Rupp & Templin, 2011). Despite the practicality of using standards and functions, it was anticipated that this *Q*-matrix alone would pose certain estimation and convergence challenges due to having a large number of attributes. Additionally, this matrix might not show a good fit. A quick

inspection of language functions also revealed that these functions were condensed and it is possible to define more fine-grained attributes that are also commonly encountered in L2 reading literature. Therefore, an alternative *Q*-matrix was included in the study.

Expert-defined[1] Q-matrix

A group of language testing experts (referred to as SMEs) was convened to develop the second *Q*-matrix. One SME was a professor of educational measurement. Two SMEs were graduate students specializing in language testing and they were quite familiar with the DCM methodology. They also taught ELs previously. The remaining SMEs were from the testing agency responsible for the development and validation of the assessment. They all had a graduate degree in language testing and one of them worked with DCMs before. Test developer involvement was particularly insightful due to their close acquaintance with the content of the test. Of the seven SMEs four were native speakers of English.

In developing this *Q*-matrix, task analysis (Sawaki et al., 2009) was used to draft the initial list of attributes with three SMEs. Task analysis involves responding to the items and recording the skills students are expected to engage with. First, the SME who was the researcher of this study compiled a list of common L2 reading attributes in the literature. Example attributes were meant to give an idea to SMEs in the task analysis and other attributes were expected to emerge. An introductory meeting was held to discuss the example attribute list. SMEs worked on example items together to start deriving the attributes. A common understanding of the task was ensured in the meeting. Three SMEs then completed the analysis of the actual test items individually. The group met again to review the attributes each SME elicited and shared their interpretations. A final list along with the operationalization of each attribute was created.

After this stage, SMEs from the test developer were involved in the process. All seven SMEs were invited to map the final attribute list with the items individually. They were provided with task descriptions, coding examples, test items, the attribute list, and item and distractor analyses. They were also requested to provide a rationale for their attribute choices. This information was sought to understand their mapping and resolve the potential disagreement. Specifically, SMEs' ratings were updated, when necessary, based on these descriptions. Mainly, if the rationale for selecting attributes contradicted the selection or the attribute was missing despite being mentioned, the rating was corrected. SMEs also rated their confidence for the coding of each item on a scale from 1 (low) to 5 (high). Due to the participation of multiple raters, the variability and similarity of SME mappings were also examined. Fleiss Kappa, which is an inter-rater agreement index for categorical ratings among multiple raters, was reported for this purpose.

All *Q*-matrices received from SMEs were reviewed by the researcher to build the final *Q*-matrix. An attribute was coded for an item if it was selected by four or more of the seven SMEs. Because the *Q*-matrix is coded by experts, and experts indicate their subjective views about item-attribute relationships, it is not always correctly specified (de la Torre & Minchen, 2019; Kang et al., 2019). To minimize the impact of misspecification,

[1] The term is first used by Reid et al. (2018).

some empirical methods are established to validate a *Q*-matrix and correct misspecifications. In this study, the *Q*-matrix was also validated using de la Torre and Chiu's (de la Torre & Chiu, 2016) general discrimination index (GDI, de la Torre & Minchen, 2019) to assess its statistical viability. In this method, the attribute combination yielding the highest difference in the correct response probabilities of masters and non-masters of attributes is defined to be the correct combination for an item. In doing this, the proportion of variance accounted for (PVAF) is estimated for each attribute combination and compared against a criterion value. $PVAF > .95$ rule was adopted for the study. When multiple attribute combinations yield high variance, the simplest combination with fewer attributes is chosen. Because GDI is a statistical method, it is possible to create a plausible *Q*-matrix based on statistical evidence, which might not be necessarily tenable from the theoretical perspective. Therefore, item content, item blueprints, experts' initial ratings, and rationale were reviewed again to determine the plausibility of the suggestions. Furthermore, to explore whether the statistical modifications were due to chance, the sample data was randomly divided into two as training and validation samples. Figure 1 shows the distribution of scores across samples. GDI was run with two samples to see whether the same modifications hold.

Evaluation and *Q*-matrix selection

After the two *Q*-matrices were developed, the LCDM was fit student response data using both Q-matrices separately. Relative and absolute fit indices were combined to evaluate model fit and select the best-fitting *Q*-matrix. Concerning relative fit, Akaike Information Criteria (AIC; Akaike, 1987), and Bayesian Information Criteria (BIC, Schwarz, 1978) were reported. A smaller value indicates the best-fitting model for both indices. Absolute fit indices are also recommended for comparing alternative matrices (Kunina-Habenicht et al., 2012; Li et al., 2016). The following six absolute fit indices were used for comparisons: average root mean squared error of approximation (RMSEA; von Davier, 2006), mean absolute deviation of item correlations (MADcor; DiBello et al., 2007), mean absolute deviation of item residual covariances (MADres; McDonald & Mok, 1995), standardized root mean square residuals (SRMSR; Maydeu-Olivares, 2013), mean absolute deviation of Q3 (MADQ3 (Yen, 1984), and $M\chi^2$ (Chen & Thissen, 1997). In addition to being common indices, previous studies report an acceptable to good performance for the selected indices (Lei & Li, 2016; Li et al., 2016). Generally, as these indices approach zero, the fit improves (Kunina-Habenicht et al., 2012).

Results

Research question 1: Key attributes underlying the test items

The standards-based Q-matrix

By using standards and language functions in test specifications the standards-based Q-matrix in Table 2 was constructed. As shown in Table 3 all attributes related to the standards were associated with six items except for the *Social Instructional Language* attribute which was related with three items. For attributes related to functions, *Recount* was measured 10 times, *Explain* 13 times, and *Argue* only 4 times. Each attribute related to the standards was measured together with an attribute related to language functions except for *Recount-Language of Science*, *Explain-Social and Instructional Language*, and

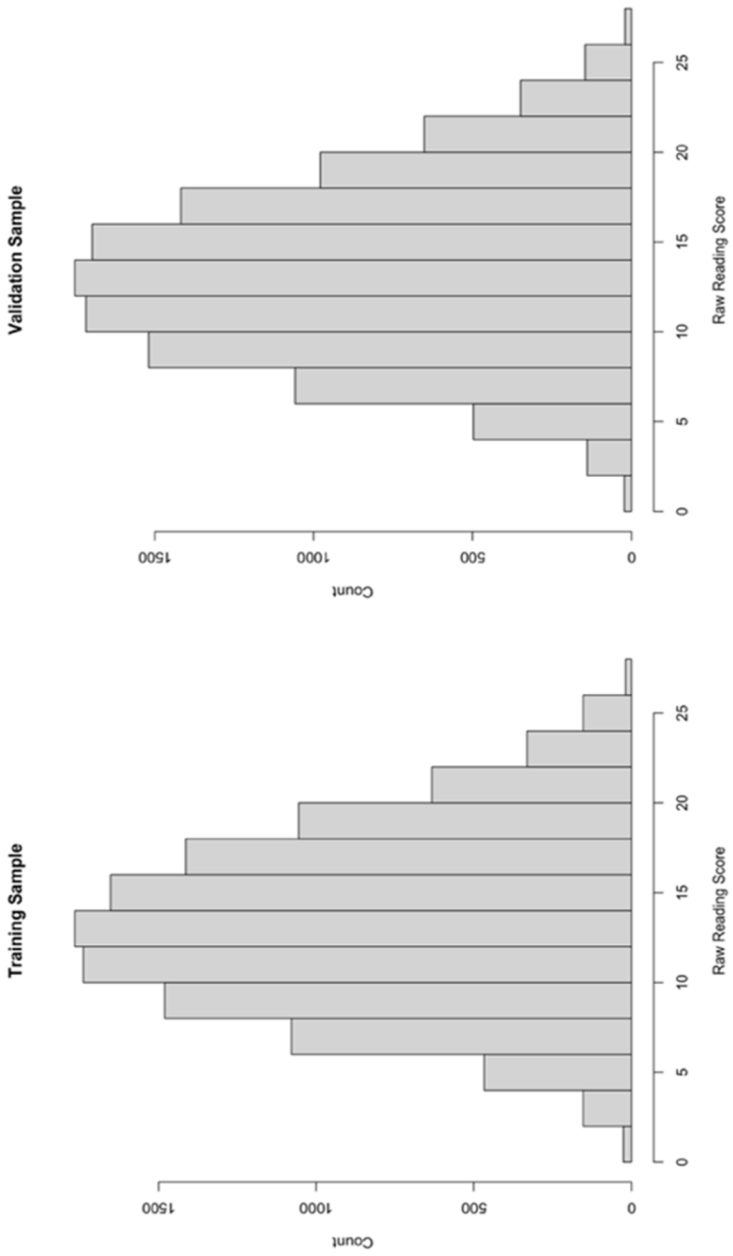

Fig. 1 Raw score distribution for the training and validation samples

Table 2 The final *Q*-matrix based on the standards and functions

Items	Attributes							
	LoSI	LoMA	LoLA	LoSC	LoSS	Recount	Explain	Argue
1	1	0	0	0	0	1	0	0
2	1	0	0	0	0	1	0	0
3	1	0	0	0	0	0	0	1
4	0	0	1	0	0	1	0	0
5	0	0	1	0	0	1	0	0
6	0	0	1	0	0	0	0	1
7	0	1	0	0	0	1	0	0
8	0	1	0	0	0	0	1	0
9	0	1	0	0	0	1	0	0
10	0	0	0	1	0	0	1	0
11	0	0	0	1	0	0	1	0
12	0	0	0	1	0	0	1	0
13	0	0	0	0	1	1	0	0
14	0	0	0	0	1	0	1	0
15	0	0	0	0	1	0	0	1
16	0	0	1	0	0	1	0	0
17	0	0	1	0	0	0	1	0
18	0	0	1	0	0	0	1	0
19	0	1	0	0	0	1	0	0
20	0	1	0	0	0	0	1	0
21	0	1	0	0	0	0	1	0
22	0	0	0	1	0	0	1	0
23	0	0	0	1	0	0	1	0
24	0	0	0	1	0	0	0	1
25	0	0	0	0	1	1	0	0
26	0	0	0	0	1	0	1	0
27	0	0	0	0	1	0	1	0

LoSI social instructional language, *LoMA* language related to math, *LoLA* language related to language arts, *LoSC* language related to science, *LoSS* language related to social studies

Table 3 Bivariate frequency of attributes in the final standards-based *Q*-matrix

	Recount	Explain	Argue
LoSI	2	–	1
LoMA	3	3	–
LoLA	3	2	1
LoSC	-	5	1
LoSS	2	3	1

Argue-Language of Math. This was expected as *Argument* is about opinions, while math items were concerned with reading passages about math concepts. Similarly, science items were about reading passages related to cycles or hypotheses that aligned better with either the *Explain* or *Argue* attributes.

The expert-defined Q-matrix

A total of six attributes were defined initially by three SMEs for this *Q*-matrix. The rest of the panel did not recommend any additional attributes.

- *Vocabulary* is related to understanding the keywords and phrases, recognizing synonyms, and paraphrasing. Because vocabulary knowledge might superficially apply to most items on a language test, the attribute was intended specifically for items requiring knowledge of difficult, content-specific, technical vocabulary.
- *Grammar* entails processing compound sentences. This includes understanding pronoun references, conjunctions, and other cohesive elements. Like the *Vocabulary* attribute, *Grammar* might apply to most items because baseline grammar knowledge is necessary to comprehend texts. This attribute is considered when extracting meaning from sentence structure is deemed necessary.
- *Explicit information and details* involve deriving specific details from the text and comprehending explicit information. It requires scanning the text for transparent details and matching it with the correct answer.
- *Inference* requires understanding implicit information and making inferences. Given the grade level of ELs, inferences can be low-level. For example, students might need to associate the information with an example situation. What distinguishes this attribute from the previous one is the transparency of the information.
- *Summary* represents integrating information from adjacent sentences or different parts of the text (i.e., across paragraphs or cells of a chart) to make meaning. In some situations, this attribute requires understanding the gist, summary, or rhetorical relations.
- *Sequences and processes* are related to understanding the description of processes, cycles, and sequential language.

The identified attributes were proved adequate and rigorous as they concurred with the skill and task descriptions in the item blueprints. In the blueprints, vocabulary requirements ranged from general to specialized vocabulary for items. In a similar vein, some items and their reading stimuli were prescribed to include simple sentences and modifiers, while others contained complex sentences with multiple clauses and a variety of modifiers. Other attributes also appeared directly or indirectly in item blueprints. For example, items were related to identifying, inferring, interpreting, summarizing, and sequencing characteristics, or details from the reading texts. It is worth noting that three SMEs did not see the blueprint descriptions while specifying the attributes and they were used by the researcher to confirm the attributes later.

Due to space limitations, the original codings of SMEs were not provided here. Table 4 presents the initial expert-defined *Q*-matrix where attributes specified by four or more raters were kept for an item. It must be noted that some variability occurred in the original mappings, as anticipated, due to the complexity of the reading construct. However, the confidence ratings in Table 5 showed SMEs were assured of their attribute choices. This can hint at the robustness, as well as clarity of attributes and the overall task. It was observed that SMEs were less certain about their coding when an additional attribute was needed for an item (i.e., complex items). As Fig. 2 shows

Table 4 The initial and final expert-defined *Q*-matrix after empirical validation

Items	Initial *Q*-matrix						Validated *Q*-matrix					
	VOC	GRM	EXP	INF	SUM	SEQ	VOC	GRM	EXP	INF	SUM	SEQ
1	0	0	1	0	0	0	0	0	1	0	0	0
2	0	0	1	1	0	0	0	0	1	1	0	0
3	1	0	1	0	0	0	1	0	1	0	0	0
4	0	0	1	0	0	0	0	0	1	0	0	0
5	1	0	0	0	1	0	1	0	0	0	1	0
6	1	0	1	0	0	0	1	0	1	0	0	0
7	1	0	1	0	0	0	1	0	1	0	0	0
8	0	0	1	0	0	1	0	0	1	0	0	1
9	0	1	1	0	0	1	0	1	1	0	0	1
10	0	0	0	0	0	1	0	0	0	0	0	1
11	0	0	0	1	1	0	0	0	0	1	1	0
12	1	0	1	0	0	0	1	0	1	0	0	0
13	1	1	0	0	0	0	1	1	0	0	0	0
14	0	0	0	0	1	0	0	0	0	0	1	0
15	1	1	0	1	0	0	1	1	0	1	0	0
16	0	0	1	0	0	1	0	0	1	0	0	1
17	0	0	0	0	1	1	0	0	0	0	1	1
[a]18	0	0	1	1	1	0	0	0	0	1	1	0
19	0	0	1	0	0	0	0	0	1	0	0	0
20	0	0	1	0	0	1	0	0	1	0	0	1
21	0	0	0	1	1	0	0	0	0	1	1	0
22	0	0	1	0	0	1	0	0	1	0	0	1
[a]23	1	0	1	0	1	0	1	0	0	0	1	0
24	0	0	0	1	0	0	0	0	0	1	0	0
25	1	0	0	0	1	0	1	0	0	0	1	0
26	1	0	0	0	0	0	1	0	0	0	0	0
27	1	0	0	1	0	0	1	0	0	1	0	0
Total	*11*	*3*	*15*	*7*	*8*	*7*	*11*	*3*	*13*	*7*	*8*	*7*

VOC vocabulary, *GRM* grammar, *EXP* explicit information and details, *INF* inference, *SUM* summary, *SEQ* sequences and process

[a] Denotes the refined items based on the GDI

Table 5 Confidence ratings for attribute-item mapping

	SME 1	SME 2	SME 3	SME 4	SME 5	SME 6	SME 7	Items
Mean	4.22	4.04	3.48	4.22	3.56	4.00	4.63	4.02
sd	0.89	0.85	1.25	0.80	0.93	0.48	0.56	0.35

variability in their confidence was greater for items measuring *Language of Math*. SMEs were also most assured when coding items related to *Social Instructional Language*. In addition, the agreement rate for the individual attributes in Table 6 ranged between 0.22 and 0.66, meaning there was fair to substantial levels of agreement. There was more variability concerning the selection of *Grammar* and *Extracting Explicit Information*, yet raters substantially agreed on Sequences. The variability was also projected as the SME group was relatively large and no overall group discussion

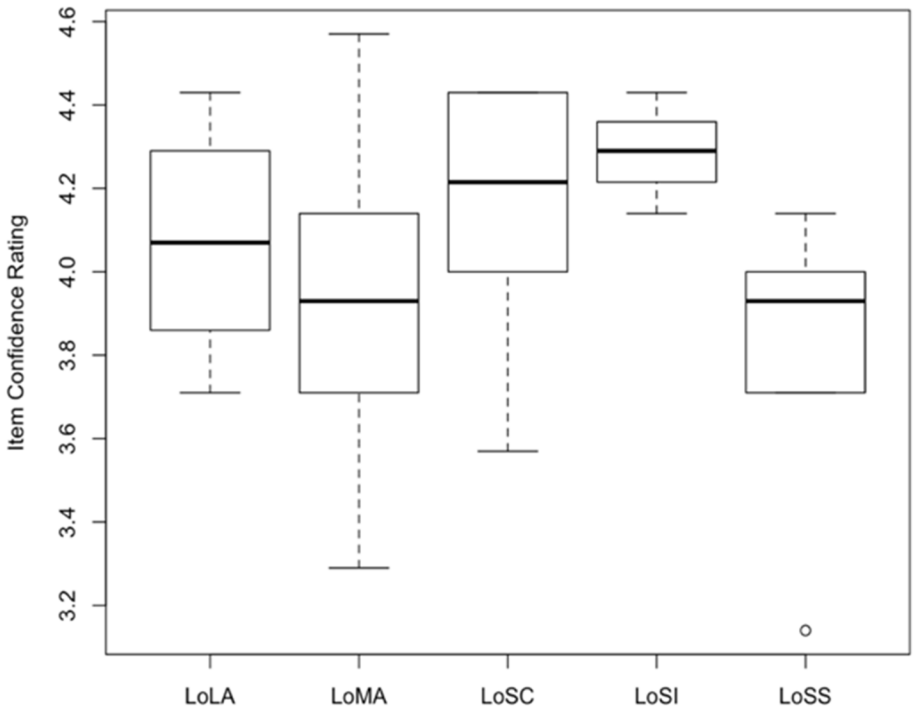

Fig. 2 Distribution of confidence ratings across content areas

Table 6 Attribute-level agreement rate among SMEs

	All SMEs ($N = 7$)			SME 1, 2, 3		
	Fleiss kappa	***z* statistic**	***p* value**	**Fleiss kappa**	***z* statistic**	***p* value**
VOC	0.381	9.074	0	0.604	5.44	0
GRM	0.216	5.151	0	0.777	6.99	0
EXP	0.234	5.564	0	0.54	4.858	0
INF	0.416	9.914	0	0.673	6.053	0
SUM	0.287	6.832	0	0.533	4.794	0
SEQ	0.664	15.817	0	0.871	7.843	0
Average	0.366	–	–	0.666	–	–

< 0.00 = poor, < 0.20 = slight, 0.21–0.40 = fair, 0.41–0.60 = moderate, 0.61–0.80 = substantial, 0.81–.1.00 = perfect (Landis & Koch, 1977, p. 165)

was held. Kappa was also computed for SMEs 1, 2, and 3 who established the attributes and discussed their mappings. The findings demonstrate that they most likely benefited from their small group discussion.

The empirical validation of this initial Q-matrix via the GDI method suggested updating attribute-item relations for some items. Fig. 3 displays the Mesa plots (Ma, 2019) based on GDI. In these plots, PVAF (y-axis) was plotted for different attribute combinations on the x-axis, The original q-vector specified by the experts was marked with a red dot.

For instance, for item 15, 3 attributes were specified initially (attributes 1, 2, 4). Among these, attribute 2 was relevant but not enough. There was a noticeable leap when attribute 1 was specified. However, attribute 4 did not add above and beyond and could be

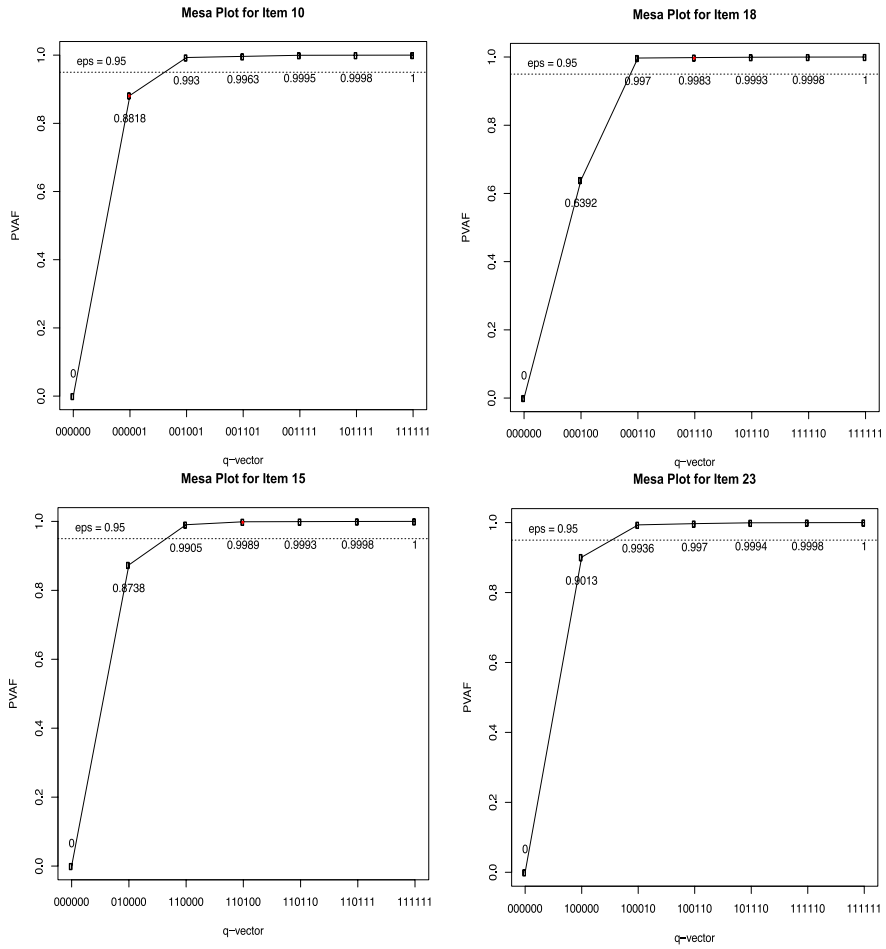

Fig. 3 Mesa plots of the four items flagged by the GDI

omitted as the plateau effect was evident. A similar pattern is apparent for items 18 and 23. For item 18, attributes 4 and 5, and for item 23, attributes 1 and 5 met PVAF > 0.95 and two attributes were sufficient. A different pattern emerged for item 10. Adding attribute 3 would increase PVAF by 0.11 points and was tenable based on the GDI.

Substantive evidence (e.g., blueprints, expert rationale) was also sought to determine the plausibility of these suggestions. Of the recommended changes, only two were applicable. For item 15, deleting attribute 4 (*Inference*) was not reasonable because all SMEs matched it to the item. The item blueprint also evinced that the item was designed to assess inferencing skills. For item 10, it was not plausible to add attribute 3 (*Explicit Information*) as the attribute would be specious. The item required an understanding of the whole process rather than specific details of the process also shown in blueprint descriptions (i.e., sequence sentences). On the other hand, the omission of *Explicit Information* from items 18 and 23 could be supported. For item 18, two SMEs who coded *Explicit Information* described a synthesis process in their rationale. The blueprint also referred to inference and synthesis skills only. Similarly, item 23 blueprint, suggested this item requires understanding part of a large process, which conveys that ELs need to comprehend the whole process.

Table 7 Bivariate frequency of attributes in the final expert-defined *Q*-matrix

	VOC	GRM	EXP	INF
GRM	1			
EXP	5	1		
INF	2	–	1	
SUM	3	–	–	3
SEQ	–	–	5	–

Table 8 Comparison between the expert-defined and standards-based *Q*-matrices

	Standards-based *Q*-matrix	Expert-defined *Q*-matrix
Relative fit		
-LL	**− 195658.4**	− 196322.85
AIC	**391606.701**	392893.709
BIC	**392678.286**	393810.099
Absolute fit		
$M\chi^2$	75.290	129.163
MADcor	0.012	0.018
SRMSR	0.016	0.024
MADres	0.281	0.392
MADQ3	0.024	0.023
Mean RMESEA	0.016	0.023

When the GDI was applied again to the data, it did not suggest further recommendations beyond the modifications in the first step. Thus, the *Q*-matrix was finalized as presented in Table 4. *Grammar* was the least frequent attribute with 3 items. *Vocabulary* and *Explicit Information* were the most frequent attributes with 11 and 13 items respectively. Seven items had a simple structure, meaning they were associated with a single attribute. Except for *Grammar*, all attributes were measured alone at least one time. As shown in Table 7 the attributes were also coupled with each other at least one time, except for *Vocabulary-Sequences*, *Inference-Sequences*, and *Explicit Information-Summary*. In summary, the structure of the initial *Q*-matrix was fair despite retrofitting.

Research question 2: An evaluation of the expert-defined and standards-based *Q*-matrices regarding model fit

All fit indices obtained from the LCDM estimation for the evaluation of standards-based and expert-defined *Q*-matrices are presented in Table 8. The standards-based *Q*-matrix with lower AIC, BIC, and log-likelihood values fit slightly better than the expert-defined *Q*-matrix. It must be highlighted that more attributes were associated with the standards-based matrix. The number of model parameters increase with the number of attributes which might also result in better model fit. On the other hand, the penalty is greater in BIC with additional parameters and it was lower for the standards-based *Q*-matrix. In addition, both matrices were acceptable based on absolute fit as all indices approached 0 except for MADres. In general, the absolute fit indices were lower for

Table 9 Tetrachoric correlations among attributes in the standards-based *Q*-matrix

	LoSI	LoMA	LoLA	LoSC	LoSS	Recount	Explain
LoMA	0.751						
LoLA	0.867	0.795					
LoSC	0.798	0.877	0.738				
LoSS	0.768	0.895	**0.957**	0.846			
Recount	0.053	− 0.066	0.234	0.135	0.168		
Explain	0.163	− 0.088	0.336	0.052	0.222	**0.999**	
Argue	0.006	− 0.033	0.179	0.082	0.129	**0.981**	**0.981**

Correlations higher than 0.90 are bolded. The proportion of masters for recount, explain, and argue was 47%, 41%, and 46%. LoSS and LoLA were mastered by 37% and 52% respectively

the standards-based *Q*-matrix in comparison to the expert-defined *Q*-matrix. However, tetrachoric correlation coefficients among attributes of the standards-based *Q*-matrix in Table 9 revealed that the expert-defined *Q*-matrix was more suitable. The coefficients indicate correlations of attribute mastery in the target population and they were high in the standards-based *Q*-matrix.

Specifically, correlations among the attributes related to the standards ranged between 0.75 and 0.96. On the other hand, the attributes associated with language functions were slightly correlated or not correlated with the attributes related to standards with a range of − 0.07–0.34. The highest correlations were between *Explain*, *Recount*, and *Language ofLanguage of Arts*, and *Explain* and *Language of Social Studies*. However, even those were weak correlations. This was not surprising, as standards represent language related to the subject areas, while functions signify language processes students need to respond correctly. To give an example, knowing math language does not suggest the student should master arguments as well. Furthermore, attributes related to functions were perfectly correlated with each other at 0.98–0.99. The proportions of students who mastered these attributes were also similar. Forty-seven percent of ELs mastered *Recount*, 41% mastered *Explain*, and 46% mastered *Argue*. The high correlations and similar proportions of masters suggest that these attributes are highly associated with each other. If an EL masters *Explain*, then they are also masters of *Recount* and *Argue*. These finding hints that language functions cannot be represented as separate attributes. This pattern was also reflected in the class proportions. More than half of the students were classified in the profile where none of the attributes (20%), just the standards (13%), just the functions (12%), or all attributes were mastered (11%). However, merging functions and representing them as a single attribute was not possible. It would reduce the standards-based *Q*-matrix to a simple structure because each item measures only one standard. Given the inseparability of attributes related to functions in the standards-based *Q*-matrix and a relatively large number of classes (2^8) that is not practical for reporting, the expert-defined *Q*-matrix was adopted as the final *Q*-matrix.

Discussion

This study explored the development of two alternative *Q*-matrices within DCM methodology, a standards-based and an expert-defined *Q*-matrix, and evaluated their fit to represent the L2 reading construct in a K–12 ELP assessment. The expert-defined *Q*-matrix was found more appropriate to model student responses. In this section, a discussion of findings related to the *Q*-matrix development and selection processes, an evaluation of the attributes from the theoretical and instructional perspective, and the study limitations and recommendations are presented.

Q-matrix development and selection between alternative *Q*-matrices

The standards-based *Q*-matrix provided a practical approach to *Q*-matrix development as no experts were involved in the process. Test and item blueprints were adequate to build the *Q*-matrix which made the processing time and resource involvement more efficient. On the other hand, the structure of the standards-based *Q*-matrix was less desirable because it compiled two simple structure dimensions together. However, despite this property, it showed a better fit than the expert-defined *Q*-matrix based on absolute and relative fit indices. This was contrary to the expectation. The standards-based matrix included more attributes and model parameters, which might have accounted for the better fit. Nevertheless, the expert-defined *Q*-matrix was selected as the final matrix for the study, because the attributes related to the language functions in the standards-based matrix were perfectly correlated with each other and a similar proportions of ELs achieved these attributes. This finding shows when alternative *Q*-matrices are employed, an inspection of fit indices might not be sufficient. A more holistic model inspection is necessary for choosing the final *Q*-matrix. However, the study findings should not be interpreted as the attributes in the standards-based *Q*-matrix are less related to the items or are not germane to the performance of the ELs. It has the advantage of being familiar to the potential users of the diagnostic information. Yet, the expert-defined *Q*-matrix can be more suitable for diagnostic information (i.e., better differentiation of the attributes) and feasible for reporting (i.e., 64 vs. 256 classes). In addition, despite having more dimensions, the attributes themselves in the standard-based *Q*-matrix were more broadly defined, an expected issue debated by some researchers (Li & Suen, 2013; Leighton & Gierl, 2007). This broadness in definitions might also have caused the inseparability of attributes related to the functions.

In the development of the expert-defined *Q*-matrix, a relatively large group of SMEs with seven experts collaborated. Typically, four–five experts were involved in similar studies. The composition of the SME panel was also distinctive, as compared to those described in the published literature. The current panel included test developers and all panelists had language testing and L2 reading content experience. Weir et al. (1990) argue that for the sake of consistency of decisions about reading skills, experienced experts from testers and linguists should be selected. Because experts were not only familiar with the construct domain but also with measurement concepts, they were provided with statistical information for the items and the distractors. The group was tacitly rely on such information when selecting attributes. For example, one an SME explained they decided to select vocabulary for an item because the most picked distractor with technical vocabulary was creating confusion for ELs. Thus, statistical information might

be helpful for the correct specification of the *Q*-matrix in retrofitting studies. Although rater agreement results showed a fair amount of agreement among SMEs–a result consistent with previous research (e.g., 0.31–0.38 among 4–5 raters in Jang, 2009; Li & Suen, 2013; Kim, 2015), the study hints that the agreement rate can be increased with training and discussion. Three SMEs who specified attributes had small group meetings and there was substantial agreement among them.

The study also employed an empirical *Q*-matrix validation method and an elaborate design to cross-validate modifications for integrity for the expert-defined *Q*-matrix. The GDI method proposed changes for four items and three of them were over-specification issues. Previous studies also reported over-specification as an issue, albeit for more items (e.g., 8 items in Kim, 2015; 7 items in Ravand, 2016). Language experts might favor to adding rather than missing skills. It must be noted that fewer modifications do not speak to a perfectly specified *Q*-matrix. Some misspecifications might still be present within, which the empirical approach did not catch. The true *Q*-matrix is not known since the attributes, as explained earlier, were retrofitted. The final *Q*-matrix was the most optimal to account for 95% of the variance between the masters and non-masters of the attributes. It must be acknowledged that when the cutoff (i.e., PVAF) was changed, seven modifications were proposed by the GDI method. However, the additional suggestions were not applicable for substantive reasons. The *Q*-matrix validation is not only an iterative process but also entails a holistic approach. For instance, Nájera et al. (2019) proposed a relationship between PVAF cutoff and item discrimination.

It was challengingGiven the retrofitting design, it was not possible to meet the ideal *Q*-matrix conditions as specified in other studies in terms combining/separating attributes (Madison & Bradshaw, 2015) or having enough items for each attribute (Deonovic et al., 2019; Jang, 2009). *Grammar, as an example,* was a sparse attribute measured by only three items and it was not separated from other attributes. Results regarding *Grammar* should be treated cautiously. Liu et al. (2018) recommend maintaining such attributes for the completeness of the *Q*-matrix but disregarding them in interpretation. The rest of the attributes were associated with 7–13 items. All attributes were measured together, except for *Sequences* with *Vocabulary* and *Inference*. Some of the desired conditions were met for the expert-defined *Q*-matrix, but results can still be improved in the future. Thus, the findings of the study are informative for the test design. It will be beneficial to compare the *Q*-matrix structure with the test developer's intentions when building forms. For instance, more items were associated with identifying details. It must be confirmed whether this was intentional or consistent with coverage of the skill. In conclusion, the present results speak to the feasibility and utility of the *Q*-matrix approach with a retrofitting design.

An evaluation of the underlying attributes in the selected *Q*-matrix

Six attributes were represented in the selected final expert-defined *Q*-matrix, which were *Vocabulary,Grammar*, *Explicit Information and Details*, Inference, *Summary*, and *Sequences*. The grain size can be regarded as being proper when the number of attributes in similar studies (i.e., 3–10) is taken into consideration. The attributes identified for the test also aligned with the definition of the academic L2 reading construct. In addition to the knowledge of vocabulary and grammar that are deemed necessary for reading

comprehension (e.g., Grabe, 1991; Koda, 2007), the attributes were related to language functions. Indeed, K–12 academic language is operationalized as language functions by some language researchers (Wolf & Faulkner-Bond, 2016) such as sequencing, summarizing, inferencing, synthesizing, retelling, and describing (e.g., Sato, 2007 in Frantz et al., 2014, p. 442). The attributes specified were also connected with the process of reading. For example, Koda (2007) suggests that reading comprehension involves drawing information from the text and processing it by integrating, synthesizing, and using prior knowledge. The attributes were also akin to the attributes in some earlier DCM studies concerned with the college-level L2 reading construct (e.g., Li & Suen, 2013; Sawaki et al., 2009). This was expected because the assessment focused on academic reading at a lower level though. In other words, despite sharing similar processes, the complexity of the attributes varies at different levels. Moreover, the attributes defined by the experts matched the task specifications of the test developer, which provides further evidence of their rigor.

It is also worth pointing out that because vocabulary and syntax are critical for comprehension (Grabe, 1991; Harding et al., 2015; Koda, 2007), they might relate to all items on a reading test. This concern was also raised by two SMEs, who later acknowledged that *Vocabulary* and *Grammar* were more relevant to specific items on the form, or that some items required more than baseline grammar and vocabulary knowledge. The item blueprints confirmed that some items differed, with respect to the requirement of these dimensions (e.g., understanding of complex sentence structure, and technical vocabulary). It is also suggested that K–12 academic English is characterized by complex structures, embedded sentences, various phrases, conjunctions, etc. (Frantz et al., 2014). Fostering awareness of academic language is regarded as being effective. Giving feedback on how ELs performed on these attributes could initiate such awareness. Thus, the two attributes were kept in the study. The usefulness of the attributes was also considered when selecting them, and because they are consistent with the L2 reading process and academic language, the attributes specified in this study can be helpful for teachers.

Limitations and recommedations

The study, similar to all empirical investigations has limitations. Discussions among SMEs were not feasible. Researchers working in the area are encouraged to plan for group discussions. At minimum, the research plan should solicit documented rationale for SME choices. In addition, despite not being asked to, some SMEs indicated their thoughts about interactions among attributes. This SME input converged with the LCDM estimation. Rather than just asking SMEs to map attributes to test items, other input such as SMEs' thoughts about the interaction between attributes can be requested in future studies. This input can be compared with statistical analysis to confirm attribute relations and model selection. Future implementations can also benefit from refining the *Q*-matrix with other sources of information. As suggested by Li and Suen (2013) the present SME panel may not have captured all salient processes that test-takers employed. Matthews (1990) argues that skills and strategies employed by learners might vary, and they "interrelate differently" for learners (p. 515). Input from ELs, in the form of cognitive surveys, can help verify the attributes specified in the study. Such input can also

reveal the existence of other strategies like distractor elimination brought up by some SMEs in this study. Cognitive surveys with ELs can also help us better understand the influence of content knowledge on solving test items. For some items reading processes might vary depending on the level of content knowledge. This would result in a different *Q*-matrix specification. Due to the absence of such information, the current study worked on the assumption that processes are the same across all ELs. However, ELs themselves are a heterogonous group. Future studies can integrate background variables or proficiency levels when building *Q*-matrices. Another group of stakeholders that can provide beneficial input for the *Q*-matrix in future studies is the teachers. They observe the reading processes of students every day and can provide valuable insights into the types of skills they use. The involvement of teachers can also increase their understanding of the construct and improve their interpretations of the results (Wolf et al., 2016).

Implications and conclusions

This study provided further evidence for the representation of the L2 reading construct and its divisibility when specifying the attributes and developing the *Q*-matrix. However, as the study implies specifying attributes and the *Q*-matrix is an arduous undertaking. L2 reading maintains its obscurity and not all the aspects are fully understood (Sawaki et al., 2009). Thus, developing alternative *Q*-matrices might offer a better strategy to better represent the construct especially when attributes are specified **ad hoc**. The study also suggests that despite the practicality of standards or test blueprints for *Q*-matrix development an alternative matrix might be more suitable for the application. Different sources of information, a careful composition of the expert panel, and rounds of discussion are recommended to facilitate the *Q*-matrix development process. Another implication is that a holistic approach should be undertaken when selecting from alternative matrices or validating a *Q*-matrix. Considering the specificity, the divisibility of the attributes, and the practicality of reporting the expert-defined *Q*-matrix was more appropriate for this study. In addition to the detailed feedback, attributes and the *Q*-matrix structure is useful for test construction purposes such as verifying dimensions and coverage.

Abbreviations

DCMs	Diagnostic classification models
ELP	English language proficiency
ELs	English learners
L2	Second language
TOEFL	Test of English as a Foreign Language
MELAB	Michigan English Language Assessment Battery

Acknowledgements
I would like to thank all subject matter experts who participated in the study for their generous time.

Author's contributions
The author assumes individual responsibility for the study conception, design, data collection, analysis, and manuscript preparation. The author read and approved the final manuscript.

Authors' information
Meltem Yumsek is a psychometrician currently working for the Ministry of Education, Turkey. She obtained her Ph.D. degree from the University of North Carolina Greensboro in educational research and measurement. Her research interests include measuring language ability, diagnostic measurement, validity and validation.

Funding
The author did not receive funding to complete this study.

Availability of data and materials
The dataset used in the study is not publicly available and obtained from WIDA with permission.

Declarations

Competing interests
The author declares no competing interests.

Received: 13 September 2022 Accepted: 24 December 2022
Published online: 24 January 2023

References

Akaike, H. (1987). Factor analysis and AIC. *Psychometrika*, *52(3)*, 317-332.
Alderson, J. C. (1990). Testing reading comprehension skills (part one). *Reading in a Foreign Language*, *6*(2), 425–438.
Alderson, J. C. (2000). *Assessing reading*. Cambridge University Press.
Alderson, J. C., & Lukmani, Y. (1989). Cognition and reading: Cognitive levels as embodied in test questions. *Reading in a Foreign Language*, *5*(2), 253–270.
Brunfaut, T., & McCray, G. (2015). *Looking into test-takers' cognitive processes whilst completing reading tasks: A mixed-method eye-tracking and stimulated recall study (ARAGs Research Report)*. British Council https://www.britishcouncil.org/sites/default/files/brunfaut_and_mccray_report_final_0.pdf.
Buck, G., Tatsuoka, K., & Kostin, I. (1997). The subskills of reading: Rule-space analysis of a multiple-choice test of second language reading comprehension. *Language Learning*, *47*(3), 423–466. https://doi.org/10.1111/0023-8333.00016.
Chalhoub-Deville, M. (2016). Validity theory: Reform policies, accountability testing, and consequences. *Language Testing*, *33*(4), 453–472. https://doi.org/10.1177/0265532215593312.
Chen, W., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics*, *22*(3), 265–289. https://doi.org/10.2307/1165285.
Davis, F. B. (1968). Research in comprehension in reading. *Reading Research Quarterly*, *3*, 499–545.
de la Torre, J. (2009). A cognitive diagnosis model for cognitively based multiple-choice options. *Applied Psychological Measurement*, *33*(3), 163–183. https://doi.org/10.1177/0146621608320520.
de la Torre, J., & Chiu, C. Y. (2016). A general method of empirical q-matrix validation. *Psychometrika*, *81*(2), 253–273. https://doi.org/10.1007/s11336-015-9467-8.
de la Torre, J., Hong, Y., & Deng, W. (2010). Factors affecting the item parameter estimation and classification accuracy of the dina model. *Journal of Educational Measurement*, *47*(2), 227–249. https://doi.org/10.1111/j.1745-3984.2010.00110.x.
de la Torre, J., & Minchen, N. (2014). Cognitively diagnostic assessments and the cognitive diagnosis model framework. *Psicología Educativa*, *20*(2), 89–97. https://doi.org/10.1016/j.pse.2014.11.001.
de la Torre, J., & Minchen, N. D. (2019). The g-dina model framework. In M. von Davier, & Y. S. Lee (Eds.), *Handbook of diagnostic classification models*, (pp. 155–170). Springer.
Deonovic, B., Chopade, P., Yudelson, M., de la Torre, J., & von Davier, A. (2019). Application of cognitive diagnostic models to learning and assessment systems. In M. von Davier, & Y. S. Lee (Eds.), *Handbook of diagnostic classification models*, (pp. 461–488). Springer.
Deville, C., & Chalhoub-Deville, M. (2011). Accountability assessment under no child left behind: Agenda, practice, and future. *Language Testing*, *28*(3), 307–321. https://doi.org/10.1177/0265532211400876.
DiBello, L. V., Roussos, L. A., & Stout, W. (2007). Review of cognitively diagnostic assessment and a summary of psychometric models. In C. R. Rao, & S. Sinharay (Eds.), *Handbook of statistics*, (vol. 26, pp. 970–1030). Elsevier.
Every Student Succeeds Act [ESSA], 20 U.S.C. § 6301 (2015). https://www.congress.gov/bill/114th-congress/senate-bill/1177
Frantz, R. S., Bailey, A. L., Starr, L., & Perea, L. (2014). Measuring academic language proficiency in school-age English language proficiency assessments under new college and career readiness standards in the United States. *Language Assessment Quarterly*, *11*(4), 432–457. https://doi.org/10.1080/15434303.2014.959123.
Gierl, M. J., & Cui, Y. (2008). Defining characteristics of diagnostic classification models and the problem of retrofitting in cognitive diagnostic assessment. *Measurement Interdisciplinary Research and Perspectives*, *6*(4), 263–268. https://doi.org/10.1080/15366360802497762.
Grabe, W. (1991). Current developments in second language reading research. *TESOL Quarterly*, *25*(3), 375–406. https://doi.org/10.2307/3586977.
Haberman, S. J. (2008). When can subscores have value? *Journal of Educational and Behavioral Statistics*, *33*(2), 204–229. https://doi.org/10.3102/10769986073026.
Haberman, S. J., & von Davier, M. (2007). Some notes on models for cognitively based skills diagnosis. In C. R. Rao, & S. Sinharay (Eds.), *Handbook of statistics*, (vol. 26, pp. 1031–1038). Elsevier.
Haertel, E. H. (1999). Validity arguments for high-stakes testing: In search of the evidence. *Educational Measurement: Issues and Practice*, *18*(4), 5–9. https://doi.org/10.1111/j.1745-3992.1999.tb00276.x.
Haertel, E. H., & Herman, J. L. (2005). A historical perspective on validity arguments for accountability testing. *Yearbook of the National Society for the Study of Education*, *104*(2), 1–34. https://doi.org/10.1177/016146810510701401.
Harding, L., Alderson, J. C., & Brunfaut, T. (2015). Diagnostic assessment of reading and listening in a second or foreign language: Elaborating on diagnostic principles. *Language Testing*, *32*(3), 317–336. https://doi.org/10.1177/0265532214564505.

Henson, R. A. (2009). Diagnostic classification models: Thoughts and future directions. *Measurement: Interdisciplinary Research and Perspectives*, *7*(1), 34–36. https://doi.org/10.1080/15366360802715395.
Henson, R. A., Templin, J. L., & Willse, J. T. (2009). Defining a family of cognitive diagnosis models using log-linear models with latent variables. *Psychometrika*, *74*(2), 191–210. https://doi.org/10.1007/s11336-008-9089-5.
Hudson, T. (1996). *Assessing second language academic reading from a communicative competence perspective: Relevance for TOEFL 2000*. Educational Testing Service https://www.ets.org/Media/Research/pdf/RM-96-06.pdf.
Huff, K., & Goodman, D. P. (2007). The demand for cognitive diagnostic assessment. In J. P. Leighton, & M. J. Gierl (Eds.), *Cognitive diagnostic assessment for education*, (pp. 19–60). Cambridge University Press.
Jang, E. E. (2009). Demystifying a q-matrix for making diagnostic inferences about L2 reading skills. *Language Assessment Quarterly*, *6*(3), 210–238. https://doi.org/10.1080/15434300903071817.
Kang, C., Yang, Y., & Zeng, P. (2019). Q-Matrix refinement based on item fit statistic rmsea. *Applied Psychological Measurement*, *43*(7), 527–542. https://doi.org/10.1177/0146621618813104.
Kim, A. A., Kondo, A., Blair, A., Mancilla, L., Chapman, M., & Wilmes, C. (2016). *Interpretation and use of K-12 language proficiency assessment score reports: Perspectives of educators and parents*. University of Wisconsin-Madison https://wcer.wisc.edu/docs/working-papers/Working_Paper_No_2016_8.pdf.
Kim, A. Y. (2009). Investigating second language reading components: Reading for different types of meaning. *Teachers College, Columbia University Working Papers in TESOL & Applied Linguistics*, *9*(2), 1–28.
Kim, A. Y. (2015). Exploring ways to provide diagnostic feedback with an ESL placement test: Cognitive diagnostic assessment of L2 reading ability. *Language Testing*, *32*(2), 227–258. https://doi.org/10.1177/0265532214558457.
Koda, K. (2007). Reading and language learning: Crosslinguistic constraints on second language reading development. *Language learning*, *57*(1), 1–44. https://doi.org/10.1111/j.1467-9922.2007.00411.x.
Koda, K. (2012). How to do research on second language reading. In A. Mackey, & S. M. Gass (Eds.), *Research methods in second language acquisition: A practical guide*, (pp. 158–179). Blackwell Publishing.
Kunina-Habenicht, O., Rupp, A. A., & Wilhelm, O. (2012). The impact of model misspecification on parameter estimation and item-fit assessment in log-linear diagnostic classification models. *Journal of Educational Measurement*, *49*(1), 59–81. https://doi.org/10.1111/j.1745-3984.2011.00160.x.
Kunnan, A. J., & Jang, E. E. (2009). Diagnostic feedback in language assessment. In M. H. Long, & C. J. Doughty (Eds.), *The handbook of language teaching*, (pp. 610–627). Wiley Blackwell.
Landis, J. R., & Koch, G. G. (1977). An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. *Biometrics*, 363–374. https://doi.org/10.2307/2529786.
Lei, P. W., & Li, H. (2016). Performance of fit indices in choosing correct cognitive diagnostic models and q-matrices. *Applied Psychological Measurement*, *40*(6), 405–417. https://doi.org/10.1177/0146621616647954.
Leighton, J. P., & Gierl, M. J. (2007). Why cognitive diagnostic assessment. In J. P. Leighton, & M. J. Gierl (Eds.), *Cognitive diagnostic assessment for education: Theory and applications*, (pp. 9–18). Cambridge University Press.
Li, H., Hunter, C. V., & Lei, P. W. (2016). The selection of cognitive diagnostic models for a reading comprehension test. *Language Testing*, *33*(3), 391–409. https://doi.org/10.1177/026553221559084.
Li, H., & Suen, H. K. (2013). Constructing and validating a q-matrix for cognitive diagnostic analyses of a reading test. *Educational Assessment*, *18*(1), 1–25. https://doi.org/10.1080/10627197.2013.761522.
Liu, R., Huggins-Manley, A. C., & Bulut, O. (2018). Retrofitting diagnostic classification models to responses from irt-based assessment forms. *Educational and Psychological Measurement*, *78*(3), 357–383. https://doi.org/10.1177/0013164416685599.
Lopez, A. (2019). Empowering K-12 teachers to make better use of high-stakes summative elp assessments [Paper presentation]. In *Annual meeting of Language Testing Research Colloquium, Atlanta, GA*.
Ma, W. (2019). Cognitive diagnosis modeling using the gdina r package. In M. von Davier, & Y. S. Lee (Eds.), *Handbook of diagnostic classification models*, (pp. 593–601). Springer.
Madison, M. J., & Bradshaw, L. P. (2015). The effects of q-matrix design on classification accuracy in the log-linear cognitive diagnosis model. *Educational and Psychological Measurement*, *75*(3), 491–511. https://doi.org/10.1177/0013164414539162.
Matthews, M. (1990). Skill taxonomies and problems for the testing of reading. *Reading in a Foreign Language*, *7*(1), 511–517.
Maydeu-Olivares, A. (2013). Goodness-of-fit assessment of item response theory models. *Measurement: Interdisciplinary Research and Perspectives*, *11*(3), 71–101. https://doi.org/10.1080/15366367.2013.831680.
McDonald, R. P., & Mok, M. M. C. (1995). Goodness of fit in item response models. *Multivariate Behavioral Research*, *30*, 23–40. https://doi.org/10.1207/s15327906mbr3001_2.
Nájera, P., Sorrel, M. A., & Abad, F. J. (2019). Reconsidering cutoff points in the general method of empirical q-matrix validation. *Educational and Psychological Measurement*, *79*(4), 727–753. https://doi.org/10.1177/0013164418822700.
Ravand, H. (2016). Application of a cognitive diagnostic model to a high-stakes reading comprehension test. *Journal of Psychoeducational Assessment*, *34*(8), 782–799.
Reid, A. M., Hoeve, K., & Henson, R. A. (2018). Fitting a diagnostic assessment to standards-defined skills versus expert-defined skills [Paper presentation]. In *Annual meeting of National Council on Measurement in Education, New York, NY*.
Rost, D. H. (1993). Assessing different components of reading comprehension: Fact or fiction? *Language Testing*, *10*(1), 79–92. https://doi.org/10.1177/026553229301000105.
Rupp, A., Templin, J., & Henson, R. A. (2010). *Diagnostic measurement: Theory, methods, and applications*. Guilford Press.
Rupp, A. A., & Templin, J. L. (2011). Unique characteristics of diagnostic classification models: A comprehensive review of the current state-of-the-art. *Measurement*, *6*(4), 219–262. https://doi.org/10.1080/15366360802490866.
Sawaki, Y., Kim, H. J., & Gentile, C. (2009). Q-matrix construction: Defining the link between constructs and test items in large-scale reading and listening comprehension assessments. *Language Assessment Quarterly*, *6*(3), 190–209. https://doi.org/10.1080/15434300902801917.
Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6*, 461–464.
Sessoms, J., & Henson, R. A. (2018). Applications of diagnostic classification models: A literature review and critical commentary. *Measurement*, *16*(1), 1–17. https://doi.org/10.1080/15366367.2018.14351.

Song, M. Y. (2008). Do divisible subskills exist in second language (L2) comprehension? A structural equation modeling approach. *Language Testing*, *25*(4), 435–464. https://doi.org/10.1177/0265532208094 2.
Tatsuoka, K. K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement*, *20*, 345–354. https://doi.org/10.1111/j.1745-3984.1983.tb00212.x.
Templin, J., & Bradshaw, L. (2014). Hierarchical diagnostic classification models: A family of models for estimating and testing attribute hierarchies. *Psychometrika*, *79*(2), 317–339. https://doi.org/10.1007/s11336-013-9362-0.
Templin, J., & Hoffman, L. (2013). Obtaining diagnostic classification model estimates using Mplus. *Educational Measurement: Issues and Practice*, *32*(2), 37–50. https://doi.org/10.1111/emip.12010.
Urquhart, A. H., & Weir, C. J. (1998). *Reading in a second language: Process, product and practice*. Routledge.
von Davier, M. (2006). *Multidimensional latent trait modelling (MDLTM) [Software pro- gram]*. Princeton: Educational Testing Service.
von Davier, M. (2008). A general diagnostic model applied to language testing data. *British Journal of Mathematical and Statistical Psychology*, *61*(2), 287–307. https://doi.org/10.1348/000711007X193957.
Weir, C. J. (2005). *Language testing and validation*. Palgrave McMillan.
Weir, C. J., Hughes, A., & Porter, D. (1990). Reading skills: Hierarchies implications relationships and identifiability. *Reading in a Foreign Language*, *7*(1), 505–510.
WIDA Consortium (2012). *2012 Amplification of the English Language Development Standards Kindergarten–Grade 12*. Board of Regents of the University of Wisconsin System https://wida.wisc.edu/sites/default/files/resource/2012-ELD-Standards.pdf.
WIDA Consortium (2016). *Can do descriptors: Key uses edition*. Board of Regents of the University of Wisconsin System https://wida.wisc.edu/sites/default/files/resource/CanDo-KeyUses-Gr-6-8.pdf.
Wolf, M. K., & Faulkner-Bond, M. (2016). Validating English language proficiency assessment uses for English learners: Academic language proficiency and content assessment performance. *Educational Measurement: Issues and Practice*, *35*(2), 6–18. https://doi.org/10.1111/emip.12105.
Wolf, M. K., Guzman-Orth, D., & Hauck, M. C. (2016). *Next-generation summative English language proficiency assessments for English learners: Priorities for policy and research*. Educational Testing Service https://files.eric.ed.gov/fulltext/EJ1124766.pdf.
Yang, X., & Embretson, S. E. (2007). Construct validity and cognitive diagnostic assessment. In J. P. Leighton, & M. J. Gierl (Eds.), *Cognitive diagnostic assessment for education: Theory and applications*, (pp. 119–145). Cambridge University Press.
Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement*, *8*(2), 125–145. https://doi.org/10.1177/014662168400800201.

Publisher's Note