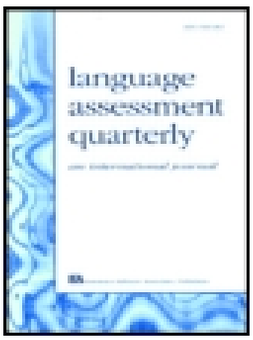

Language Assessment Quarterly

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

Q-Matrix Construction: Defining the Link Between Constructs and Test Items in Large-Scale Reading and Listening Comprehension Assessments

Yasuyo Sawaki , Hae-Jin Kim & Claudia Gentile

To cite this article: Yasuyo Sawaki , Hae-Jin Kim & Claudia Gentile (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, DOI: 10.1080/15434300902801917

To link to this article: https://doi.org/10.1080/15434300902801917

Language Assessment Quarterly, 6: 190–209, 2009

ISSN: 1543-4303 print / 1543-4311 online
DOI: 10.1080/15434300902801917

Routledge
Taylor & Francis Group

Q-Matrix Construction: Defining the Link Between Constructs and Test Items in Large-Scale Reading and Listening Comprehension Assessments

Yasuyo Sawaki
Waseda University

Hae-Jin Kim
Educational Testing Service, Princeton, NJ

Claudia Gentile
Mathematica Policy Research, Inc, Princeton, NJ

In cognitive diagnosis a Q-matrix (Tatsuoka, 1983, 1990), which is an incidence matrix that defines the relationships between test items and constructs of interest, has great impact on the nature of performance feedback that can be provided to score users. The purpose of the present study was to identify meaningful skill coding categories that reflect core language skills and processes assessed in the Reading and Listening sections of the Test of English as a Foreign Language™ Internet-based Test (TOEFL® iBT). The study was conducted as part of a research activity to explore the possibility of developing a detailed score report for low-stakes use by taking a cognitive diagnosis approach. Content experts conducted a test content analysis to develop draft Q-matrices, while measurement experts empirically analyzed the draft Q-matrices with examinee performance data using a cognitively diagnostic psychometric model called the fusion model (DiBello, Stout, & Roussos, 1995; Hartz, 2002). The draft Q-matrices were refined by repeating fusion model analysis and revision of skill definitions and item coding. This resulted in a set of Q-matrices that represented substantively meaningful score reporting categories of a suitable grain size for score reporting, while maintaining an acceptable level of examinee classification consistency.

INTRODUCTION

Cognitive diagnosis is a diagnostic assessment approach developed by integrating cognitive psychology and psychometrics. A primary purpose of cognitive diagnosis is to provide test score users with fine-grained information about learner test performance, so that individual learners' strengths and weaknesses can be identified to facilitate instruction. In cognitive diagnosis, performance feedback is provided at a level that is more detailed than the level at which scores are

Correspondence should be sent to Yasuyo Sawaki, Faculty of Education and Integrated Arts and Sciences, Waseda University, 1-6-1 Nishi Waseda, Shinjuku, Toyko 169-8050, Japan. E-mail: ysawaki@waseda.jp

typically reported in assessments designed for other purposes. For example, a second language listening comprehension test used for selection purposes might report a single score to represent an examinee's overall listening ability level. In contrast, a listening test used in the context of cognitive diagnosis might provide a score profile on multiple skills or processes representing different aspects of listening comprehension.

A first step in cognitive diagnosis is to define target attributes that serve as the basis for developing score reporting categories in a score report. An attribute "refers to anything that affects performance on a task: either a task characteristic, or any of the knowledge, skills or abilities necessary to complete the task" (Buck & Tatsuoka, 1998, p. 121). Defining score reporting categories involves (a) identifying a list of key attributes of interest that are useful in understanding learners' strengths and weaknesses and (b) devising detailed item coding rules, so that test items can be coded for the target attributes consistently across test forms and across those who code test items. Once target attributes are defined this way, individual test items are coded for the attributes to develop a Q-matrix (Tatsuoka, 1983, 1990), an incidence matrix that specifies the relationships between individual test items and target attributes (see Table 1 in the overview article by Lee and Sawaki in this special issue for an example). Then, examinee item response data are analyzed with the Q-matrix to estimate individual examinees' knowledge states on the target attributes. Finally, a score report is prepared by translating the empirical analysis results into a form and language that is accessible to the score user.

The idea of defining attributes that test items tap, which serves as the starting point for developing a Q-matrix in cognitive diagnosis, is not new in language assessment. For decades language assessment researchers have used various approaches to identifying specific attributes that L2 reading and listening comprehension items assess. For example, quite a few previous authors employed an approach focusing on surface task characteristics. This approach was used in previous item difficulty modeling studies (Bachman, Davidson, & Milanovic, 1996; Bachman, Davidson, Ryan, & Choi, 1995; Carr, 2006; Clapham, 1996; Freedle & Kostin, 1993, 1999; Nissan, DeVincenzi, & Tang, 1996) and in previous applications of rule-space methodology to language assessment (e.g., Buck & Tatsuoka, 1998; Buck, Tatsuoka, & Kostin, 1997; Buck, Tatsuoka, Kostin, & Phelps, 1997). Other approaches included those based on subskills, where target constructs of interest in language tests are identified according to theoretical taxonomies of language ability (e.g., Alderson, 1990a, 1990b; Alderson & Lukmani, 1989) and others based on skills and processes that learners report in introspective studies of language test-taking processes (e.g., Jang, 2005; Kasai, 1997; Scott, 1998).

Despite the breadth of the approaches to item coding taken previously in L2 comprehension assessment earlier, it is fair to say that the state of the knowledge in the field as to what aspects of language ability or test characteristics need to be identified as attributes in a Q-matrix for diagnostic purposes is quite limited for two reasons. The first is a variety of conceptual and empirical difficulties in identifying meaningful subconstructs reported in previous studies. For example, the field currently lacks consensus as to what constructs, that is, subskills, comprise second language reading and listening abilities (Alderson, 2000; Buck, 2001). Supporting evidence for this position includes previous studies that reported the difficulty expert judges had in agreeing what skills test items assessed (e.g., Alderson 1990a, 1990b; Alderson & Lukmani, 1989; Bachman et al., 1996; Lumley, 1993) and a controversy over the number of psychometrically distinct factors that can be identified within L2 comprehension (e.g., D. H. Rost, 1993; Sawaki, Stricker, & Oranje, 2008). Moreover, introspective studies of learners' test-taking

processes have shown that learners do not necessarily engage in the processes that test developers intend to test. Thus, an item might be testing an intended skill for some learners but not for others (Alderson, 1990b; Jang, 2005; Wijgh, 1996). Alderson (2000) stated that the difficulty of identifying subskills is not a problem when, for example, a reading test is constructed to cover a wide range of skills and processes for score reporting at a global level. However, this becomes a problem, he pointed out, when a claim is made to have developed a diagnostic assessment that identifies strengths and weaknesses of learner performance (Alderson, 2000, pp. 305–306).

In the item difficulty modeling studies just described, some task characteristics, predictive of item difficulty or discrimination, were identified. However, this does not necessarily mean that such item characteristics can directly be adapted to diagnostic score reporting. As pointed out by Bachman (2002), the results are difficult to generalize because of the diversity of the approaches taken to define and quantify the variables investigated by the previous researchers. Moreover, these studies were exploratory in nature, because the primary goal was to identify constructs and test task characteristics that predict the items' difficulty and/or discrimination. Thus, some of these studies were conducted with a large pool of test items (e.g., Carr, 2006; Freedle & Kostin, 1993, 1999; Nissan et al., 1996). This sharply contrasts with the development of score reporting categories in a cognitive diagnosis context, where the researcher is forced to work with a limited number of items available in a test form.

A second reason for the lack of guidance in the literature for how one might go about identifying meaningful attributes for learner diagnosis is the general paucity of previous research on diagnostic language assessment, as pointed out by authors such as Alderson (2005) and Alderson and Huhta (2005). It is true that few published studies were available on language assessments specifically designed for diagnosis until very recently, when studies on the development of the DIALANG, a diagnostic self-assessment in 14 European languages, began to be reported at conferences and in publications in the late 1990s. However, some pioneering work on extracting diagnostic information from existing nondiagnostic language tests has been around for a while. Some notable examples include Buck and his associates' work on the rule space methodology (RSM), an early classification algorithm that inspired the development of many currently available psychometric models for cognitive diagnosis (e.g., Buck & Tatsuoka, 1998; Buck, Tatsuoka, & Kostin, 1997; Buck, Tatsuoka, Kostin, et al., 1997).

The authors of these pioneering studies pointed out the potential of the RSM for providing learners with diagnostic information (e.g., see Buck, Tatsuoka, & Kostin, 1997, for a sample diagnostic score report based on their RSM analysis of the TOEIC® Reading section). However, these studies have yet to demonstrate the full potential of the framework for applications to learner skill diagnosis in language assessment. The nature of the attributes identified in these studies were closely linked to test task characteristics ("a 'nuts and bolts' level of attribute definition"; Buck & Tatsuoka, 1998, p. 125). The reasonable rater agreement on item coding reported by these authors (e.g., Buck, Tatsuoka, Kostin, et al., 1997) showed promise of this approach for better understanding the relationship between test task characteristics and item response patterns. Nevertheless, the conceptual difficulties associated with "translating" these attributes to meaningful score reporting categories based on target abilities still remains. Moreover, the discussions in these studies focused primarily on the test item analysis for identification of useful attributes, whereas the actual learner profiles, such as the prevalence of different attribute mastery profiles, were not explored in detail. For these reasons, it was not entirely clear how this

approach might be useful to understand various learner profiles and better link assessment to instruction.[1]

Although the current state of the field in terms of diagnostic language assessment described above makes identifying meaningful diagnostic score reporting categories a challenge, learners and their teachers would benefit enormously from receiving detailed performance feedback that informs further learning. Toward this end, a team of content experts and measurement specialists collaborated in a research study to explore the possibility of developing an enhanced score report for the Test of English as a Foreign Language™ Internet-based Test (TOEFL® iBT) Reading and Listening sections by identifying potential score reporting categories. Currently two types of score reports are available for TOEFL iBT. The first is the TOEFL iBT Official Score Report. This report provides the TOEFL iBT total and section scores and is used by designated institutions for high-stakes decision making about candidates. The other is the TOEFL iBT Examinee Score Report that examinees receive. Based on results of two scale-anchoring studies, this report provides not only the TOEFL iBT total and section scores but also a description of typical performance of examinees on different sections at three different ability levels (see Gomez, Noah, Schedl, Wright, & Yolkut, 2007, for details). In this approach, however, all examinees in the same ability group receive the same feedback. For this reason, the description of typical performance of examinees at a given ability level may not necessarily match strengths and weaknesses of each examinee. Thus, exploring the possibility of employing a cognitive diagnosis approach for enhancing the examine score report would move us a step forward in the direction of providing individualized performance feedback tailored to the needs of individual examinees.

Having this context as the background, our study addressed three specific goals. The first was to develop score reporting categories focusing primarily on key constructs. Although key task characteristics that are deemed to affect test performance were taken into consideration for devising detailed scoring rules for each skill, the proposed score reporting categories do not highlight task characteristics. This is because providing performance feedback focusing exclusively on features of test tasks that the learner completed during the test might offer a rather shortsighted view of language ability that lacks generalizability to nonassessment language use tasks and contexts. This might potentially lead to unwanted encouragement of test preparation focusing too much on how to "get by" items that have certain task features.

Second, despite the proposed low-stakes context just described, the goal was to identify score reporting categories that maintain an acceptable level of examinee classification reliability across test forms. A basic requirement to achieve this goal would be to identify core language skills and processes that effectively differentiate high-performing examinees from low-performing examinees in this population. At the same time, the score reporting categories should be consistently supported by the limited number of items available in the TOEFL iBT Reading and Listening sections across forms. Thus, the key was to define appropriate score reporting categories at a right grain size while maintaining an acceptable level of examinee classification consistency.

Finally, an important question that we attempted to address in this study was whether it is possible to extract detailed diagnostic information about examinee test performance from an existing test that is not specifically designed for diagnosis. Given that previous language assessment studies have indicated the conceptual difficulties in disentangling various skills that contribute

[1]Note, however, that more recent applications of rule space methodology to subject-area assessment report learner score profiles (e.g., Birenbaum, Tatsuoka, & Yamada, 2004).

to language comprehension, it was of particular interest to see whether employing a psychometric model for cognitive diagnosis called the fusion model (Hartz, 2002) as an integral part of the coding scheme development facilitates the process.

METHOD

Participants in the Skill Category Development

A team of content experts served as developers of skill categories that reflect core language skills and processes for successful performance on the TOEFL iBT Reading and Listening sections. The content experts included six Educational Testing Service staff with experience in applied linguistics research and English as a Second Language/English as a Foreign Language teaching. Three of them were TOEFL iBT assessment development specialists, whereas the remaining three were language assessment researchers not directly involved in TOEFL test development. At the beginning, the three researchers conducted initial conceptualization and draft coding category development work. Once a draft list of skills was identified for the TOEFL iBT Reading and Listening sections, the three researchers and the assessment development specialists worked as a team to refine the coding scheme and develop Q-matrices. Throughout the project, this team worked closely with measurement experts at Educational Testing Service, who conducted the fusion model analysis, for discussion of various substantive and empirical issues.

Test Forms and Data

Reading and Listening items in four TOEFL iBT test forms were analyzed. Two forms were TOEFL iBT prototype test forms published as Tests 1 and 2 in the LanguEdge Courseware. The others were two TOEFL iBT test forms administered as part of a field study for the new test conducted in 2003 and 2004. Each Listening test form comprised two conversations and four academic lectures of approximately 4 to 6 min in length. Each conversation set contained 5 items, whereas each academic lecture set contained 6 items, with 34 multiple-choice items per test form. Each Reading form consisted of 38 to 40 multiple-choice items based on three academic texts of approximately 700 words in length. Each text was followed by 12 to 14 items. All items in the Reading and Listening test forms were scored dichotomously, except a few partial-credit items worth more than 1 point in some of the Listening test forms and in each Reading form.

Examinee item response data from more than 3,000 examinees who completed the LanguEdge test forms in 2002 as well as those from more than 3,000 examinees who participated in a field study in 2003–2004 were available to this study. Among the field study participants, 441 completed both test forms. As described by Chapelle, Enright, and Jamieson (2008), the LanguEdge and field study samples were reasonably representative of the TOEFL test taker population in terms of native country of origin and native languages. With regard to language ability level, the LanguEdge sample was comparable to the operational TOEFL test taker population based on their TOEFL Paper-based Test total scores (Chapelle et al., 2008). In contrast, the 2003–2004 field study sample scored roughly half a standard deviation below the operational TOEFL population on the TOEFL CBT Reading and Listening sections, resulting in mean differences indicating medium effect sizes (Cohen's, 1988, *d* values of .54 for Listening and .59 for Reading).

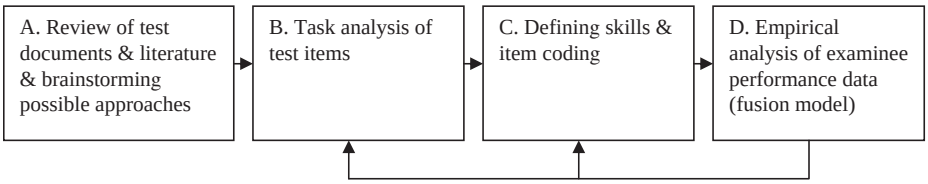

FIGURE 1 General steps taken for defining skills and developing Q-matrices for the TOEFL iBT Test Reading and Listening sections.

These findings suggest that the field test sample represented a lower ability group compared to the operational TOEFL CBT population (see Chapelle et al., 2008, for more details about the pilot test and the field test samples).

Procedure

The process of defining target skills and developing Q-matrices for the four test forms was iterative. The process was a variation of the first three steps of CDA described in the overview paper by Lee and Sawaki (this issue), where the steps of defining attributes (skills), Q-matrix construction and data analysis were repeated in multiple rounds. Figure 1 illustrates the overall procedure.

During this iterative process, the team first developed draft skills lists and coding rules for construction of Q-matrices for the LanguEdge test forms. After refining the initial skills lists and the Q-matrices based on empirical analysis results of the examinee performance data for the LanguEdge test forms, the team applied the skills lists to the development of Q-matrices for the field study test forms. The results reported in the subsequent sections focus on the analysis of the field study test forms only.

Test content analysis, defining skills, and developing draft Q-matrices. The team of content specialists first engaged in a review of key test documents related to TOEFL iBT and brainstormed possible approaches to defining a list of target skills. The test documents reviewed included the TOEFL 2000 Framework papers (Bejar, Douglas, Jamieson, Nissan, & Turner, 2000; Enright et al., 2000), which provided an initial conceptualization of the new assessment, as well as test specifications based on Evidence Centered Design, a general guiding framework adopted for the TOEFL iBT development.[2] In particular, the team paid attention to the test specifications to ensure that various features of particular test items reviewed in this study for identifying skills and coding rules indeed reflected key design features that generalize across different test forms. At the same time, relevant literature in language assessment was reviewed, and preliminary task analysis of test items on the LanguEdge test forms was conducted. Then a few possible approaches to the development of skills lists were laid out. An important decision made at this stage was to develop lists of target skills for the Reading and Listening sections that elaborate on the broad constructs defined in the current TOEFL iBT test specifications. Based on the TOEFL 2000 Framework papers for the Reading and Listening sections earlier, the current TOEFL iBT test

[2]For an introduction to evidence centered design in the context of language assessment, see Mislevy, Steinberg, and Almond (2002).

specifications define three constructs for the Reading section (Basic Comprehension, Inferencing, and Reading to Learn) and three for the Listening section (Basic Understanding, Pragmatic Understanding, and Connecting Information). These constructs as defined in these documents are mutually exclusive in nature in that each item is classified into one of the three categories within each section.

Although this three-part classification of items is useful for understanding what broad academic reading or listening construct an item is designed to tap, this system alone would not provide a full cognitive model of task performance, or fine-grained skills and processes that need to be executed to answer a given item correctly. A preliminary task analysis of individual items revealed that arriving at a correct answer to each item requires the examinee to complete various "subtasks." For example, a Connecting Information item in the Listening section can require the candidate to identify different pieces of relevant information stated in different parts of the text first and then connect them to understand the relationships among them. Moreover, this task analysis also showed a considerable range of skills and processes that are shared across items classified into different categories. On one hand, some pieces of the relevant information that need to be identified in a Connecting Information item might be explicitly stated in a localized portion of the text, which is similar to a characteristic of a typical Basic Understanding item. On the other hand, a Basic Understanding item might have a characteristic of connecting ideas. For example, when the gist of a conversation or lecture is neither explicitly stated nor presented in a localized portion of the text, the examinee needs to connect information across the text to understand the gist. Thus, the team felt it necessary to adequately capture such overlaps of skills and processes across items. Given the limited number of items available in each TOEFL iBT Reading or Listening test form, this does not necessarily lead to a dramatic increase of the number of skill categories. However, the degree of overlap of skills and processes across different categories of items needs to be reflected in the coding rules in such a way that a given item can be coded for more than one skill category as needed.

Based on the consideration of these key issues, a decision was made to employ a task analysis as the primary vehicle for developing skill coding categories. This is essentially a type of expert content analysis of test tasks that is often employed in cognitive diagnosis analyses of educational tests. The three content experts independently responded to each test item and described the skills involved in each reading or listening item, by taking an exploratory approach, to answer the question, "What skills and processes are required in order for a learner to answer this question correctly?" Then the team met to discuss the results the individual members brought and come to a consensus on the key skills required to answer each item correctly by discussion. Based on the results of this extensive content analysis, draft lists of reading and listening skills and item coding rules were developed, and draft Q-matrices were prepared by assigning a coding of 1 to an item if the skill was thought to be required to answer the item correctly.[3]

Fusion model analysis. Once the draft Q-matrices were prepared, the examinee response data (scored item responses) and the Q-matrices were analyzed together by using the fusion model (Hartz, 2002).[4] An empirical analysis based on the fusion model relates observed item response patterns to a set of attributes of interest in order to estimate individual examinees'

[3]The team took a narrow definition of the attributes by identifying only those that were deemed required to answer an item correctly, in part reflecting the nature of the fusion model, which assumed conjunctive relationships among the skills being modeled (see the overview paper by Lee and Sawaki in this issue about the discussion of conjunctive skills as an attribute structure).

[4]For a more technical discussion of this model, see von Davier, DiBello, and Yamamoto (2006).

knowledge states on the target attributes. An advantage of employing the fusion model over some other psychometric models for cognitive diagnosis is the availability of detailed information not only about individual examinees' attribute mastery states but also about the functioning of items for differentiating examinees at different attribute mastery states. The information about items is particularly useful for evaluating the appropriateness of the item coding in the Q-matrix, as discussed next.

A computer program, Arpeggio Version 2.0 (Hartz, 2002), was employed. Separate analyses were conducted for each section of each form. Among various types of information available in a fusion model analysis is the functioning of items for classification of examinees to different skill mastery states.[5] The team focused specifically on two types of information. The first type included two fusion model item parameters, π^*_i and $r^*_{i,k}$, which provide diagnostic information about the functioning of individual items. These item parameters were employed to evaluate the appropriateness of the particular coding assigned to each item in the Q-matrix:

> π^*_i = "a conditional item difficulty parameter" (von Davier et al., 2006), that is, the probability for masters of all skills required by a given item *i* as specified in the Q matrix to correctly apply all the required skills when answering item *i*. The higher the value, the easier item *i* is for masters of all skills required by the item.
>
> r^*_{ik} = a discrimination parameter of item *i* for attribute *k* (Hartz, 2002). The lower the value, the higher the discrimination power of item *i* between masters vs. non-masters of attribute *k*. In this study the r^*_{ik} value of .90 was used as a rule of thumb for an indication of sufficient discriminatory power.

Another type of information employed was the consistency of examinee classification across test forms for the test–retest sample (N = 441) collected as part of the TOEFL iBT field study. Following the procedure to investigate Test–retest Consistency Rates (TCR) proposed by Zhang, DiBello, Puhan, Henson, and Templin (2006), two types of information were closely examined.[6] First, Cohen's kappa, an index of classification agreement adjusted for agreement by chance, was obtained for each skill to evaluate the extent to which individual examinees were consistently classified as masters or nonmasters for each skill across the two test forms. Second, the number of skills for which individual examinees' master versus nonmaster classification agreed across the two test forms was noted.

Q-matrix revision. The draft skills lists, coding rules, and Q-matrices for the two field test forms were discussed with assessment development and measurement specialists in multiple meetings. In these meetings, test items and coding assigned to them were reviewed one by one, along with estimates of π^*_i and r^*_{ik}. Although no hard cut-off values were set, typically the discussions started with identification of an item with an r^*_{ik} of over .90, which indicates that the item did not effectively distinguish between masters and nonmasters of each skill. The team discussed potential reasons that might be causing the low discrimination by an extensive item content analysis and used this information to revise the codes assigned to the given item as

[5]Different methods are available in cognitive diagnosis models to classify examines to different skill mastery states. This study employed a method called expected a priori method, where a probability of mastery was obtained for each skill separately. Then, for each skill, examinees with the skill mastery probabilities of .5 or above were classified as masters of the skill, whereas those with skill mastery probabilities of below .5 were classified as nonmasters.

[6]The information required for calculation of TCRs was obtained from Arpeggio and other ancillary programs. For more details, see Zhang et al. (2006).

needed. Discussions typically centered on the coding rules (i.e., when a coding of 1 can be assigned to an item).

Because TCR did not pinpoint which code for a specific item might require reconsideration it was used primarily to evaluate the appropriateness of the Q-matrix at the global level. For example, TCR analysis results from multiple fusion model analysis runs with different skill structures were examined to determine whether an entire skill needed to be retained or omitted from the skills list for the Listening section in particular, as described in detail next. Based on discussions of various substantive issues and the Arpeggio analysis results, a revised Q matrix was devised and used in the next round of fusion model analysis.

RESULTS

Reading Section

The following list shows six skills that were initially identified as potential score reporting categories for the Reading section:

Skill 1: Understanding Word Meaning
Skill 2: Identifying Information: Search & Match
Skill 3: Understanding Information within Sentences
Skill 4: Understanding and Connecting Information within a Paragraph
Skill 5: Understanding and Connecting Information across Paragraphs
Skill 6: Understanding Relative Importance of Information and Relationships among Ideas

Skill 1 was reserved primarily for a portion of Basic Understanding items designed to assess understanding vocabulary in context, whereas a small number of other items that required understanding of technical or difficult vocabulary to arrive at a correct answer were also coded for this skill. Skill 2 involved identifying information when a lexical and/or syntactic overlap existed between the key (correct answer) and the text. In this case, the task for a candidate was to locate information in the text and match the answer in the key. Skills 3 through 5 were similar to one another in that the information needed to answer an item correctly was implicit in the text, or answering an item involved making some degree of inference. However, these skills were distinguished among themselves by the size of the search area involved to respond to an item. This decision was made based on the result of the content analysis, which suggested that differing amounts of text a test taker has to process might help distinguish among different levels of reading ability. Finally, Skill 6 was distinguished from Skills 2 through 5 in that it required not only understanding of relevant key information but also evaluating the relative importance of information and relationships among ideas presented in text.

While the team proceeded with the refinement of the coding rules for each skill based on multiple rounds of discussions on various substantive issues as well as reviews of fusion model item parameters, a few major changes were made to the original list of skills. First, Skill 4 (Understanding and Connecting Information within a Paragraph) and Skill 5 (Understanding and Connecting Information across Paragraphs) were combined into one category called “Connecting Information.” This decision was made because the cognitive processes involved in connecting information beyond a sentence level seemed to be similar, whether or not the unit of text to

analyze was at the paragraph level or beyond the paragraph level. Second, Skills 2 and 3 were combined into one category called "Understanding Specific Information." The items included in the particular forms employed in this study indicated that search and match items typically required students to process information within a sentence, suggesting that these items would be coded for both skills. However, such a coding pattern, where Skill 2 coding almost always appears with Skill 3, might lead to statistical identification problems for Skill 2. A close examination of the items coded for these skills revealed that the nature of the skills could be better represented by redefining the category as Understanding Specific Information. Table 1 provides the definition of the four skills in the final skills list for the Reading section.

In total, across the two test forms, 8 to 20 items were coded for each reading skill. A small portion of the items (12 in Form A and 10 in Form B) were coded for two or three skills. The following is an example as to how individual items were coded based on the reading skills. The options marked with asterisks (*) in all upcoming examples indicate the keys (correct answers).

Opportunists and Competitors

(1) Growth, reproduction, and daily metabolism all require an organism to expend energy. The expenditure of energy is essentially a process of budgeting, just as finances are budgeted. If all of one's money is spent on clothes, there may be none left to buy food or go to the movies. Similarly, a plant or animal cannot squander all its energy on growing a big body if none would be left over for reproduction, for this is the surest way to extinction.

TABLE 1
Skills List for the TOEFL iBT Reading Section

Skills	*Definition*
Skill 1: Understanding word meaning	The ability to recognize and know the meaning of key words.
Skill 2: Understanding specific information	The ability to understand the key information requested in the item stem, search the text and locate the sentence that contains the answer. Answering the question involves matching information in the text that has similar/synonymous wording to the key or making local inferences about information within the sentence.
Skill 3: Connecting information	The ability to understand the key information requested in the item stem, search the text and locate the paragraph or group of sentences that contains the answer. However, the key information is not stated in exactly the same words in the text as in the item. Instead, answering the question involves connecting two or more ideas or pieces of information in a paragraph or across paragraphs.
Skill 4: Synthesizing & organizing information	The ability to synthesize and organize information across parts of the text. This often involves understanding the relative importance of information, (i.e., distinguishing between main and supporting ideas). It also may involve understanding the relationship among the ideas in the text (i.e., understanding rhetorical structures such as comparison/contrast, cause and effect, temporal order, problem/solution).

In paragraph 1, the author explains the concept of energy expenditure by

A. identifying types of organisms that became extinct
B. comparing the scientific concept to a familiar human experience*
C. arguing that most organisms conserve rather than expend energy
D. describing the processes of growth, reproduction, and metabolism

This item was coded for both Skills 3 and 4 in Table 1 because a test taker needs to connect information across sentences as well as understand the rhetorical structure to answer the question correctly. The fusion model analysis produced a π^*_i value of .82 for this item. This means that the probability for masters of Skills 3 and 4 to apply both skills correctly for answering this item was .82. Moreover, the r^*_{ik} values for Skills 3 and 4 were .60 and .72, respectively, indicating that this item sufficiently discriminated between the masters and nonmasters of these skills.

It is worth noting that the language ability of the population played an important role in defining the coding rules. There were cases where it was clear that a particular skill was necessary to answer a given item correctly, but it was reasonable to assume that a majority of the TOEFL test takers had already mastered the skill. In such cases a given item might not effectively discriminate low level performers from high level performers for the particular skill. For example, even though understanding vocabulary is necessary, the previous sample item was not coded for Skill 1. The key words that test takers needed to understand to correctly answer this item were associated with a familiar human experience (e.g., *expenditure, budget, finance, money, spend, clothes*), which were deemed to be already sufficiently familiar to the TOEFL population. Thus, coding for Skill 1 was not deemed necessary.

Listening Section

The following are the initial coding categories identified for the Listening section:

Skill 1: Understanding vocabulary
Skill 2: Understanding overall topic/gist
Skill 3: Understanding important information
Skill 4: Understanding structure (rhetorical, discourse)
Skill 5: Making inferences

Skill 1 was created to report on test takers' understanding of vocabulary knowledge because it is the most fundamental skill needed in comprehension. Skill 2 represented the ability to understand overall topic or gist, whereas Skill 3 involved the ability to understand main points and supporting details. Skill 4 tapped into the ability to understand rhetorical patterns as well as turn-taking patterns in discourse. Skill 5 involved the ability to understand implied meaning across the text.

Revision of the initial skills list for the Listening section proceeded with multiple rounds of discussion on item coding rules and fusion model analysis results. Some changes were made to the initial skills list. First, Skill 1 was dropped entirely, as discussed in detail next. Second, Skills 2 and 3 were renamed Understanding General Information and Understanding Specific Information in the final skills list, respectively, to better highlight the contrast between these two skills. Third, the definitions of Skills 4 and 5 were refined. An extensive discussion of the skill

definitions and item coding rules revealed that both skills involved different types of inferencing, while this point was unclear in the original skill definitions. For this reason, these skills were redefined in the final skills list as "Understanding Text Structure & Speaker Intention" and "Connecting Ideas," respectively.

A primary issue of concern throughout the process, however, was whether to retain Skill 1 (Understanding Vocabulary) in the list, and if so, how to code items on this skill. In the test specifications and related documents on the listening section, vocabulary knowledge is considered as a skill underlying comprehension, rather than a target construct directly assessed in the TOEFL iBT Listening section. Thus, unlike the Reading section, there are no items specifically designed to assess vocabulary knowledge. However, the task analysis identified the critical importance of online processing of key vocabulary.

Accordingly, the team developed three different versions of the Q-matrices for the two field test forms and examined the impact of different coding rules on examinee classification consistency. One approach was to "prune" the item coding based on the r^*_{ik} value. All items were coded for Skill 1 for an initial fusion model analysis run. Then, revised Q-matrices were prepared by dropping Skill 1 coding for any items for which the r^*_{ik} values obtained in the initial run was greater than .90. A second approach was based on expert judgment. In this approach a given item was coded for Skill 1 only when the area of stimulus that needed to be understood as well as the stem and options involved difficult or infrequent vocabulary. In a third approach Skill 1 was entirely dropped from the Q-matrices. This approach is consistent with the test specifications that did not specifically define vocabulary knowledge as one of the target constructs.

Based on these three versions of the Q-matrices, TCR analyses were conducted using the two field test forms. Table 2 shows that the Cohen's kappa values were all in the .50s and .60s, suggesting moderate to substantial levels of examinee classification consistency across the two test forms for all three approaches (Landis & Koch, 1977).

Furthermore, the team examined the frequencies of the test–retest examinees whose master versus nonmaster classification results agreed between the two field test forms based on the three Q-matrices. Table 3 shows, for each of the three Q-matrix construction approaches, the

TABLE 2
Cohen's Kappa Values for Examinee Classification Agreement Between Field Test Forms A and B by Q-Matrix Construction Approach

	Q-Matrix Construction Approach		
	$r^*_{i,k}$ *Prune*	*Expert Judgment*	*Drop Skill 1*
Skill 1	.60	.58	N/A
Skill 2	.61	.57	.55
Skill 3	.68	.63	.62
Skill 4	.65	.69	.68
Skill 5	.65	.53	.53

TABLE 3
Frequencies of Examinees Whose Classification Results Agreed Between Field Test Forms A and B by Q-Matrix Construction Approach

Q-Matrix Construction Approach	*No. of Skills with Classification Agreement*					
	0	*1*	*2*	*3*	*4*	*5*
$r^*_{i,k}$ prune	9 (2.0%)	26 (5.9%)	18 (4.1%)	30 (6.8%)	97 (22.0%)	261 (59.2%)
Expert Judgment	18 (4.1%)	18 (4.1%)	28 (6.3%)	50 (11.3%)	80 (18.1%)	247 (56.0%)
Drop Skill 1	32 (7.3%)	24 (5.4%)	36 (8.2%)	78 (17.7%)	271 (61.5%)	

proportion of examinees for whom master versus nonmaster classification agreed for only one skill, two skills, and so on, between the two test forms.

The results in Table 3 indicate that the proportions of examinees for whom their classification agreed for all skills or all but one skill between the two test forms were comparable between the r^*_{ik} pruning method (81.2%) and the Skill 1 drop method (79.2%), whereas the proportions were relatively low for the expert judgment method (74.1%). Thus, the aforementioned TCR analysis results suggest that additional analysis or item coding work required in the r^*_{ik} pruning method and the expert judgment method, respectively, did not lead to a marked improvement in the resulting test–retest examinee classification consistency. For this reason, a decision was made to drop Understanding Vocabulary from the listening skills list. Table 4 shows the revised, final skills list.

In total, across the two test forms, 5 to 16 items were coded for each listening skill. Although the majority of the items were coded for only one skill, a small portion of the items (10 in Form A

TABLE 4
Skills List for the TOEFL iBT Listening Section

Skills	*Definition*
Skill 1: Understanding general information	The ability to understand general information or the main point of the lecture or conversation.
Skill 2: Understanding specific information	The ability to understand (to refer to notes, to remember) the details and/or supporting ideas of the lecture or conversation; to understand the ideas that are salient enough to remember or to take notes about (such as the important points).
Skill 3: Understanding text structure & speaker intention	The ability to recognize the rhetorical patterns of the text (e.g., the use of cause and effect); the structure of spoken conversations (e.g., turn taking); the rhetorical purpose (e.g., the reason the speaker told a story), or author's stance (e.g., the emotional state of the speaker).
Skill 4: Connecting ideas	The ability to make appropriate inferences and make connections across the text. The ability to understand meaning beyond the actual words spoken (when a relevant point is not explicitly stated in the text).

and 3 in Form B) were coded for two skills. Next are two examples illustrating how each item was coded.

(Narrator) Listen to part of a conversation in a library.

(Woman) Hi. Can I help you?

(Man) Yeah, I'm looking for a reference book.

(Woman) OK. Do you know the title?

(Man) Well, that's the thing. I'm not exactly sure what I'm looking for. I need uh, information on European demographics.

(Woman) OK, do you just need population statistics, like, total population, male–female . . . real basics for demographics?

(Man) Yeah. Population, literacy rate, uh, let's see . . . life expectancy by gender, like if women tend to live longer than men . . . things like that.

(Woman) OK, well, I-I'm pretty sure you can get most—if not all—of those statistics from an atlas. I can tell you where to find one in the reference section.

(Man) Yeah, but I'm kind of looking for it by city, not by country and the atlas I saw . . .

(Woman) Uh huh . . . I see. . . .

(Man) Well, do you know if there are any other reference books I can use for this? To find the statistics by city?

(Woman) City, you say. Any particular part of Europe? Eastern, western . . . southern?

(Man) No. Pretty much all across Europe.

[Conversation continues]

1. What is an example the man gives of the kind of information he needs about European cities?

   A. Their climate
   B. Their geographic size
   C. How long people live *
   D. What languages people speak

Item 1 was coded for Skill 2 only. This is because the specific information test takers need to understand is contained in a specific utterance of the male speaker, "Yeah. Population, literacy rate, uh, let's see . . . life expectancy by gender, like if women tend to live longer than men . . . things like that." Moreover, the way the speaker structured this utterance made the relevant information required to answer this item explicit. First, the use of a rising intonation for the first two examples (*population* and *literacy rate*) signaled that the next item, *life expectancy by gender,* was part of the list. Next, by the use of the word *like* the speaker indicated his intention to elaborate on this direct paraphrase of the correct option further. The π^*_i value for this item obtained from the fusion model analysis was .89, indicating that the probability for masters of Skill 2 to correctly apply the skill for answering this item was .89. The r^*_{ik} for Skill 2 was .61, suggesting that this item sufficiently discriminated between the masters and non-masters of Skill 2.

2. What does the man imply about the atlas he looked at?

   A. It does not list population statistics by city.*
   B. It does not list population statistics by country.
   C. It contains information about Europe that is out of date.
   D. It lacks information on southern Europe.

Item 2 was coded for Skills 2 and 4. This item required test takers to understand multiple turns in the discourse: "Yeah, but I'm kind of looking for it by city, not by country and the atlas I saw . . ." / "Uh huh . . . I see. . . ." / "Well, do you know if there are any other reference books I can use for this? To find the statistics by city?" Through the use of a distinct intonation pattern and multiple attempts to get across his point in different words in multiple turns, the speaker implies that the atlas does not list population statistics by city. The item not only requires test takers to understand specific information in each turn but also requires them to connect information across turns. The π^*_i value of .96 for this item indicates that the probability for masters of Skills 2 and 4 to apply both skills correctly when answering this item was .96. Moreover, the r^*_{ik} values for Skills 2 and 4 were .80 and .82, respectively, suggesting that this item sufficiently discriminated between the masters and nonmasters of these skills.

Examinee skill classification reliability based on the final Q-matrices. The examinee classification reliability based on the final Q-matrices for the field test Reading and Listening test forms were examined and reported by Zhang et al. (2006). The various types of analyses conducted by Zhang et al. (2006) included a TCR analysis similar to the one reported for the Listening section. For example, the Cohen's kappa values obtained to investigate examinee classification for the individual skills ranged from .54 to .69 for the Listening section and from .50 to .62 for the Reading section, suggesting moderate to substantial levels of examinee classification agreements on each skill across the two test forms. Moreover, when the number of skills on which examinee classification results agreed across the forms was examined, the results matched for all or all but one skill for the Listening section for 79.6% of the examinees, whereas the proportion was slightly lower for the Reading section, at 76.2%.

DISCUSSION AND CONCLUSION

The purpose of the present study was to identify meaningful skill coding categories that reflect core language skills and processes assessed in the Reading and Listening sections of the TOEFL iBT. The study was conducted as part of a research activity to explore the possibility of developing a detailed score report for low-stakes use by taking a cognitive diagnosis approach. The process of developing score reporting categories and constructing Q-matrices based on the score reporting categories was iterative, featured by a combined use of expert test content analysis and empirical analyses of examinee performance data based on the fusion model. This allowed the team to make informed decisions, taking account of various substantive and psychometric issues, throughout the development of the skills lists, coding rules, and Q-matrices for the TOEFL iBT Reading and Listening sections. A set of resulting Q-matrices represented substantively meaningful score reporting categories of a suitable grain size for score reporting while maintaining an acceptable level of examinee classification consistency across two test forms. Some key issues for further consideration that emerged as the team went through the process are discussed in detail next.

Identifying Attributes: Other Possibilities

It is worth noting that using a cognitive diagnosis approach can result in many different types of skill definitions, coding rules, and Q-matrices. For example, as previously mentioned, this study identified skills that were closely related to key constructs. However, it is possible to define skills more broadly to include other types of attributes provided that they offer useful diagnostic information for learners. For instance, a plausible alternative for the TOEFL iBT Listening section might be to model some key constructs with language use contexts. Each TOEFL iBT Listening test form consists of two conversation sets and four lecture sets. When these text types are defined as part of the score reporting categories, learners can receive performance feedback on target subconstructs as well as on their comprehension levels of conversations that take place in various campus situations and academic lectures. This approach may be feasible based on previous corpus linguistics and listening comprehension research that suggested the potential link among linguistic text complexity, text type, and ESL learners' listening comprehension level (e.g., Biber, Conrad, Reppen, Byrd, & Helt, 2002; Read, 2002; Shohamy & Inbar, 1991).

A given skill can be defined at different grain sizes as well. For example, there is a notable difference in the number of skills identified between this study and Jang's (2005) study, despite the fact that both studies developed skills lists for the Reading section of the TOEFL iBT. Jang's study was based on LanguEdge, namely, TOEFL iBT prototype test forms. In contrast, this study focused primarily on TOEFL iBT field study test forms, whereas LanguEdge test forms were used only for preliminary stages of the coding scheme development work. Jang defined nine skills for the LanguEdge Reading section, whereas this study identified only four skills for the TOEFL iBT Reading section. Moreover, Jang identified two vocabulary skills, one with and the other without the use of context clues. As shown in the differences across items in the extent to which her verbal protocol study participants utilized immediate contexts in Jang's study, it is reasonable to assume that there is more than one possible way to answer a given vocabulary item correctly. Suppose, for example, that the task for the examinee is to provide the definition of the word *squander* that appears in the last sentence of the Opportunists and Competitors paragraph presented previously. For those who know the definition of the word, this is a straightforward vocabulary item not requiring much processing of the surrounding text. However, for those who are not familiar with this word, it becomes a completely different task. These examinees would require comprehension of the surrounding text, and then fill the "gap" with an option that fits the context.

The potential use of context clues to answer a vocabulary item such as this example was recognized during the task analysis conducted in this study. However, a decision was made not to take account of the use of context clues for the purpose of defining the skill categories in this study, primarily for two reasons. First, although using context clues can be part of the process of responding to a vocabulary item, it was deemed required only when the examinee is not sufficiently familiar with an appropriate definition of a word in question, and this occurred for most of the items.[7] Second, although devising two vocabulary skills might allow one to extract

[7]In other words, the team identified few vocabulary items that required both (a) knowing a definition of a word and (b) using context clues (e.g., an item on a word that has multiple meanings, so the examinee has to resort to the context to identify a definition appropriate in the particular context).

diagnostic information about examinee performance at a finer grain size, it might not be feasible given that only a small number of vocabulary items in the TOEFL iBT test forms, if at all, required the use of context as part of the test-taking process. This decision was made because, in the present large-scale assessment context, identifying stable coding categories that can consistently be supported by sufficient numbers of items for score reporting across test forms was of prime importance.

Limitations and Directions for Future Research

In the present study the content specialists were able to reach a reasonable level of consensus on the skill definitions and item coding in the Q-matrix by discussion. This process was facilitated by the availability of the fusion model item parameters, which provide diagnostic information about the functioning of items for examine classification to masters and nonmasters of the skills. The process of defining skills and coding rules as well as developing Q-matrices described in this article covers only some early stages of a cognitive diagnosis application. Two issues that are critical for validating the score reporting categories could not be addressed adequately. First, although this study examined the extent to which examinee skill mastery classification results agreed across two test forms, the sample used for the test–retest analysis was rather small. For this reason, the generalizability of the results across test forms need to be examined more extensively with a larger sample and a larger number of test forms in the future.

A second issue that could not be addressed in this study is whether learners' test-taking processes support the cognitive model of task performance as represented in the Q-matrix developed in this study. Learners' verbal protocols were not available to this study. One approach to validate the results of the present study in the future is to examine the extent to which the coding assigned to individual items adequately account for major steps and processes different learners report that they go through to complete the items. Given the variability of learners' test-taking processes reported in previous studies, this is expected to be the key to support the appropriateness of score reporting categories that are amenable to learner diagnosis.

This article described results from earlier stages of the Q-matrix construction process based on a cognitive diagnosis analysis. Extensive further research is necessary before consideration of this approach for potential use in operational language tests. Now that the coding schemes developed in this study based on a consensus approach is available, the next step would be to investigate the extent to which raters can reach a reasonable level of agreement when they independently code items for the attributes defined in the Q-matrix. It may be possible to achieve a reasonable level of rater agreement by implementing a systematic rater training process that incorporates detailed skill definitions and item coding rules such as those documented in this study (see Bachman et al., 1996; Lumley, 1993). Once items are coded, statistical information about item functioning such as the fusion model item parameters can inform adjudication of individual item coding as well as refinement of skill definitions and the global Q-matrix structure.

Another possible direction for future research is to scrutinize possible limitations of extracting diagnostic information from an existing nondiagnostic L2 comprehension assessment. The attributes identified in this study focused on the task completion process, that is, skills and processes that experts believed are required to answer each item correctly. Using these attributes as score reporting categories would provide learners with feedback as to *what aspects* of L2

comprehension they had trouble with. What these attributes cannot tell learners, however, are the *why*. One way to address this issue may be to look more deeply into various types of linguistic subprocesses that are thought to underlie L2 comprehension. Current theories of reading and listening suggest the importance of the efficiency of lower level linguistic processes for rapid and accurate reading and listening comprehension of text (for recent reviews, see Enright et al., 2000, and M. Rost, 2005). Thus, one might consider various types of linguistic knowledge and skills (e.g., knowledge of vocabulary, syntax, semantics, acoustics, efficiency of word recognition and sentence parsing) as viable attributes to be included in the Q-matrix. However, unless the test itself is designed to specifically assess these component skills, it is extremely difficult, if not impossible, to disentangle all these linguistic knowledge and processes because of the issue raised by Alderson (1990b): the highly overlapping, synergistic nature of the construct of L2 comprehension, which seems at least partly responsible for the lack of rater agreement and the variability of learners' test-taking processes reported in previous studies. No matter how sophisticated psychometric models become, it seems that the fundamental conceptual issue—the "elusiveness" of the L2 comprehension constructs—will continue to be a major challenge that language testers have to grapple with.

The difficulties associated with identifying distinct skills within L2 comprehension assessments might suggest the need for designing a true diagnostic assessment. This might require administering a comprehension measure with a set of other discrete measures specifically designed to assess areas of linguistic knowledge that are deemed to be useful for providing diagnostic feedback to learners. A line of inquiry that may inform this approach is research focusing on the relationships among L2 comprehension and component skills profiles (e.g., van Gelderen et al., 2004; Sawaki & Sabatini, 2008; Schoonen, Hulstijn, & Bossers, 1998). It will be of interest to see if future research on component skills, for example, can provide better diagnostic feedback to learners about the skills that contribute to their L2 comprehension.

ACKNOWLEDGEMENTS

An earlier version of this paper was presented at the 29th Language Testing Research Colloquium held in Barcelona, Spain, from June 9–11, 2007. The authors' special thanks go to Louis DiBello, Xiaoying Ma, Gautam Puhan, Lin Wang and Yanling Zhang for guidance on psychometric issues and conducting fusion model analyses, and Susan Nissan, Mary Schedl and Robert French for invaluable discussions with the authors about various test design issues. Any opinions expressed in this article are those of the authors and not necessarily of Educational Testing Service.

REFERENCES

Alderson, J. C. (1990a). Testing reading comprehension skills (Part One). *Reading in a Foreign Language, 6*(2), 425–438.

Alderson, J. C. (1990b). Testing reading comprehension skills (Part Two). *Reading in a Foreign Language, 7*(1), 465–503.

Alderson, J. C. (2000). *Assessing reading*. Cambridge, UK: Cambridge University Press.

Alderson, J. C. (2005). *Diagnosing foreign language proficiency: The interface between learning and assessment*. New York: Continuum.

Alderson, J. C., & Huhta, A. (2005). The development of a suite of computer-based diagnostic tests based on the Common European Framework. *Language Testing, 22*, 301–320.

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.

Bachman, L. F. (2002). Some reflections on task-based language performance assessment. *Language Testing, 19*(4), 453–476.

Bachman, L. F., Davidson, F., & Milanovic, M. (1996). The use of test method characteristics in the content analysis and design of EFL proficiency tests. *Language Testing, 13*, 125–150.

Bachman, L. F., Davidson, F., Ryan, K., & Choi, I.-C. (1995). *An investigation into the comparability of two tests of English as a foreign language: The Cambridge-TOEFL comparability study*. Cambridge, England: Cambridge University Press.

Bachman, L. F., & Palmer, A. (1996). *Language testing in practice*. Oxford, UK: Oxford University Press.

Bejar, I., Douglas, D., Jamieson, J., Nissan, S., & Turner, J. (2000). *TOEFL 2000 listening framework: A working paper* (TOEFL Monograph Series No. MS-21). Princeton, NJ: Educational Testing Service.

Biber, D., Conrad, S., Reppen, R., Byrd, P., & Helt, M. (2002). Speaking and writing in the university: A multidimensional comparison. *TESOL Quarterly, 36*(1), 9–48.

Birenbaum, M., Tatsuoka, C., & Yamada, T. (2004). Diagnostic assessment in TIMSS-R: Between countries and within country comparisons of eighth graders' mathematics performance. *Studies in Educational Evaluation, 30*, 151–173.

Buck, G. (2001). *Assessing listening.* Cambridge, UK: Cambridge University Press.

Buck, G., & Tatsuoka, K. (1998). Application of the rule-space procedure to language testing: examining attributes of a free response listening test. *Language Testing, 15*(2), 119–157.

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.

Buck, G., Tatsuoka, K., Kostin, I., & Phelps, M. (1997). The sub-skills of listening: Rule-space analysis of a multiple-choice test of second language listening comprehension. In A. Huhta, V. Kohonen, L. Kurki-Suonio, & S. Luoma (Eds.), *Current developments and alternatives in language assessment: Proceedings of LTRC 96* (pp. 589–624). Jyväskylä, Finland: University of Jyväskylä.

Carr, N. (2006). The factor structure of test task characteristics and examinee performance. *Language Testing, 23*(3), 269–289.

Chapelle, C., Enright, M. K., & Jamieson, J. (Eds.). (2008). *Building a validity argument for the Test of English as a Foreign Language*. New York: Taylor & Francis.

Clapham, C. (1996). *The development of IELTS: A study of the effect of background knowledge on reading comprehension.* Cambridge: University of Cambridge Local Examination Syndicate and Cambridge University Press.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.) Hillsdale, NJ: Erlbaum.

Enright, M. K., Grabe, W., Koda, K., Mosenthal, P., Mulcahy-Ernt, P. I., & Schedl, M. (2000). *TOEFL 2000 reading framework: A working paper* (TOEFL Monograph Series No. MS-17). Princeton, NJ: Educational Testing Service.

Freedle, R., & Kostin, I. (1993). The prediction of TOEFL reading item difficulty: Implications for construct validity. *Language Testing, 10*, 131–170.

Freedle, R., & Kostin, I. (1999). Does the text matter in a multiple-choice test of comprehension? The case for the construct validity of TOEFL's minitalks. *Language Testing, 16*(1), 2–35.

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

Hartz, S. M. (2002). *A Bayesian framework for the unified model for assessing cognitive abilities: Blending theory with practicality.* Unpublished doctoral dissertation, University of Illinois, Urbana–Champaign.

Jang, E. E. (2005). *A validity narrative: Effects of reading skills diagnosis on teaching and learning in the context of NG TOEFL*. Unpublished doctoral dissertation, University of Illinois, Urbana.

Kasai, M. (1997). *Application of the rule space model to the reading comprehension section of the Test of English as a Foreign Language (TOEFL)*. Unpublished doctoral dissertation. Urbana: University of Illinois.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 45*, 255–268.

Lumley, T. (1993). The notion of subskills in reading comprehension tests: an EAP example. *Language Testing 10*(3), 211–234.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2002). Design and analysis in task-based language assessment. *Language Testing, 19*(4), 477–496.

Nissan, S., DeVincenzi, F., & Tang, L. K. (1996). *An analysis of factors affecting the difficulty of dialogue items in TOEFL listening comprehension.* (TOEFL Research Rep. No. 51). Princeton, NJ: Educational Testing Service.

Read, J. (2002). The use of interactive input in EAP listening assessment. *Journal of English for Academic Purposes, 1*, 105–119.

Rost, D. H. (1993). Assessing different components of reading comprehension: Fact or fiction? *Language Testing 10*, 79–92.

Rost, M. (2005). L2 listening. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning* (pp. 503–527). Mahwah, NJ: Erlbaum.

Sawaki, Y., & Sabatini, J. P. (2008, March). *Component reading skill efficiency and reading comprehension for adult ESL/EFL learners*. Paper presented at the Annual Conference of the American Association for Applied Linguistics, Washington, DC.

Sawaki, Y., Stricker, L., & Oranje, A. (2008). *Factor structure of the TOEFL Internet-based Test (iBT): Exploration in a Field Trial Sample* (TOEFL iBT Research Report TOEFLiBT-04). Princeton, NJ: Educational Testing Service.

Schoonen, R., Hulstijn, J., & Bossers, B. (1998). Metacognitive and language-specific knowledge in native and foreign language reading comprehension: An empirical study among Dutch students in Grades 6, 8 and 10. *Language Learning, 48*(1), 71–106.

Scott, H. S. (1998). *Cognitive diagnosis perspectives of a second language reading test*. Unpublished doctoral dissertation, University of Illinois, Urbana.

Shohamy, E., & Inbar, O. (1991). Validation of listening comprehension tests: the effect of text and question type. *Language Testing, 8*(1), 23–40.

Tatsuoka, K. K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement, 20*, 345–354.

Tatsuoka, K. K. (1990). Toward an integration of item-response theory and cognitive error diagnosis. In N. Fredericksen, R. Glaser, A. Lesgold, & M. G. Shafto (Eds.), *Diagnostic monitoring of skill and knowledge acquisition* (pp. 453–488). Hillsdale, NJ: Erlbaum.

van Gelderen, A., Schoonen, R., de Gropper, K., Hulstijn, J., Simis, A., Snellings, P., et al. (2004). Linguistic knowledge, processing speed, and metacognitive knowledge in first- and second-language reading comprehension: A componential analysis. *Journal of Educational Psychology, 96*(1), 19–30.

von Davier, M., DiBello, L., & Yamamoto, K. (2006). *Reporting test outcomes using models for cognitive diagnosis* (ETS Research Rep. No. RR-06-28). Princeton, NJ: Educational Testing Service.

Wijgh, I. F. (1996). A communicative test in analysis: Strategies in reading authentic texts. In A. Cumming & R. Berwick (Eds.), *Validation in language testing* (pp. 154–170). Clevedon, UK: Multilingual Matters.

Zhang, Y-L., DiBello, L., Puhan, G., Henson, R., & Templin, J. (2006, April). *Estimating skills classification reliability of student profile scores for a large-scale international English language assessment*. Paper presented at the 2006 Annual Meeting and Training Sessions of the National Council on Measurement in Education, San Francisco, CA.