# Bridging assessment and learning: a cognitive diagnostic analysis of a large-scale Spanish proficiency test

Mengmeng Wang
*Beijing Foreign Studies University*

Received: 19/09/2021 / Accepted: 25/01/2022
DOI: https://doi.org/10.30827/portalin.vi40.15930
ISSN paper edition: 1697-7467, ISSN digital edition: 2695-8244

**ABSTRACT:** The literature has highlighted the importance of Cognitive Diagnostic Assessment (CDA) of large-scale, high-stakes language tests to assess the individualized strengths and weaknesses of every learner. However, there are relatively few studies on the accuracy and applicability of feedback information in follow-up teaching and learning practices. Using both the diagnostic results of 1933 test takers derived from the Generalized Deterministic Inputs, Noisy and Gate (G-DINA) model of a national Spanish test (EEE) and qualitative data from the test takers' literature reviews in their bachelor thesis drafts, the precision of the diagnostic feedback was examined to verify its usefulness for the improvement of academic reading. The results showed that the G-DINA model had an appropriate model fit to the test performance data, and that the CDA is able to identify the specific skill profiles of each test taker, which were not always consistent with the scores provided by classical test analysis. The triangulation of the diagnostic reports and the literature reviews from the learners' thesis drafts, clearly showed that CDA used for a large-scale test can assess reading skills accurately and the feedback is valuable for improving the future academic reading and thesis revision.
**Key words:** Cognitive Diagnostic Assessment, Large-scale Test, Spanish, Individualized Learning, Reading.

**Enlazando evaluación y aprendizaje: análisis diagnóstico cognitivo de una prueba de español a gran escala**

**RESUMEN:** En el contexto de las pruebas a gran escala, previos estudios han destacado la importancia de la Evaluación Diagnóstica Cognitiva con el propósito de proveer las fortalezas y debilidades de cada alumno. Sin embargo, son escasos los estudios cuyo objetivo es verificar la precisión de los resultados y la viabilidad de enlazarlos con el futuro aprendizaje y enseñanza. Se ha empleado diagnosis cognitiva con el Modelo Generalizado de Entrada Determinista, Ruido y Puerta para analizar el puntaje de 1933 participantes en una prueba nacional de español (EEE). Se han analizado los datos cualitativos de la revisión de literatura en sus borradores de trabajo de grado para corroborar la exactitud de los resultados diagnósticos y su uso para mejorar la lectura académica. Los resultados indican que el modelo se ajusta a la prueba y permite determinar el perfil cognitivo de cada participante, lo que no siempre es viable en los análisis tradicionales de la prueba. La triangulación y análisis de los informes diagnósticos y las revisiones de literatura en los trabajos del grado, ha puesto de manifiesto que la diagnosis cognitiva permite una evaluación más precisa de las habilidades de lectura de los estudiantes y que la información obtenida puede ser de gran utilidad para futuras lecturas académicas y revisión de tesis.
**Palabras clave**: Evaluación Diagnóstica Cognitiva, Prueba a gran escala, Español, Aprendizaje Individualizado, Lectura.

## 1. Introduction

In the field of educational and psychological testing, the Classical Test Theory (CTT) and Item Response Theory (IRT) have traditionally provided useful accounts of how to validly and reliably determine the scores of the test takers by comparing and listing them in order of certain attributes. According to de la Torre and Minchen (2014: 89), these theories "are linked by their common goal of determining the extent to which students possess the proficiency or trait of interest," which is essential to decision-making in large-scale assessments. Stakeholders like policy-makers, administrators and teachers might be satisfied if the decisions made from these scores, such as judging the quality of education, determining granted privileges as admission or graduation, and identifying justification for allocating resources, are convincing (Abu-Alhija 2007). However, summative scores or rankings are not always meaningful and beneficial, especially when teachers or students are more concerned about effective remedial methods for follow-up individualized instruction or learning. This is particularly noteworthy in regard to large-scale assessments, for which some of the test takers and educators are learning or teaching about the test with the purpose of striving for better rankings, but losing sight of achieving real growth and achievements. The function of diagnosing individualized strengths and weaknesses and providing valuable remedial suggestions has not been achieved. As Mislevy (1989: 11) remarked: "Tracking students as they progress opens the door to finer grained 'micro-level' decisions to enhance learning along the way". Therefore, providing detailed feedback rather than summative scores is crucial to stakeholders who have formative needs, such as students who are unable to interpret the scores and make reasonable planning, or teachers who are confused about how to cater to students with diverse proficiency levels in the classroom.

Since it was reported in the 1980s, Cognitive Diagnostic Assessment (CDA) has attracted great interest, within the field of educational assessment (Haertel, 1989; Tatsuoka & Tatsuoka,1997; de la Torre, 2008; de la Torre *et al.*, 2010; de la Torre & Minchen, 2014; Wang and Qiu 2019; Wu *et al.* 2020; Chin 2021). Follow-up empirical studies in teaching and learning practice have provided evidence of the effectiveness of retrofitting large-scale test results, such as those from NAEP (de la Torre, 2008), SAT (Gierl *et al.*, 2009) and PISA (Wu *et al.* 2020). Previous studies have emphasized the need for research on language assessments. For instance, Lee and Sawaki (2009) pointed out the importance of applying CDA models to extract fine-grained information on linguistic knowledge states and skill mastery levels with the purpose of determining individual strengths and disadvantages.

The literature since the mid-1990s has focused on CDA in the field of language testing (Sheehan, *et al.*, 1993; Buck &Tatsuoka, 1995; Buck, *et al.*, 1997; Jang, 2009; von Davier, 2008; Sawaki, *et al.*, 2010; Li & Suen, 2013; Kim, 2015; Chen & Chen, 2016; Ranjbaran & Alavi, 2017; Yi, 2017; Min & He, 2021; Toprak & Cakir, 2021). Almost all of the previous studies have focused on assessing English language knowledge and skills. Very few studies (Jang, 2009), which have been limited to survey or interview studies, have investigated the accuracy and effect of diagnostic feedback; thus, the evidence for this issue is still inconclusive. For instance, in a study investigating L1 literacy, Sheehan *et al.* (1993) reported that CDA can be used to successfully diagnose skills and classify students into different types according to their skill profiles. In a study of L2 listening test, Buck and Tatsuoka (1995) found that CDA could be used to identify the mastery level of knowledge, skills and their interactions to distinguish Japanese learners of English into specific types. In a follow-up study of L2 reading, Buck, Tatsuoka and Kostin (1997) examined whether CDA can be used to analyze micro-skills on the TOEIC test. These early studies validated the feasibility of CDA for English language assessments, which laid the first stone for the development of a variety of models. The study by von Davier (2008) offered empirical analysis of L2 reading and listening. The author used CDA to analyze the performance in TOEFL test. Sawaki *et al.* (2010) and Min and He (2021) studied the use of CDA to classify skills on the TOEFL iBT and NETS reading and listening tests. In recent studies, Kim (2015), Ranjbaran and Alavi (2017) and Toprak and Cakir (2021) reported that in EFL reading tests, CDA is a useful tool to diagnose attributes, which can be a valuable reference for teachers. Data from these studies suggest that CDA can diagnose the individual mastery pattern of each learner. Through the retrofitting of large-scale language assessments, diagnostic purposes can also be successfully achieved. However, much uncertainty still exists about the accuracy of these results for follow-up teaching and learning practices. It should also be noted that the feedback process and the impact of these results on teaching and learning remain unclear.

To better understand teachers' and learners' views on diagnostic feedback, Jang's (2009) key study analyzed L2 reading performance in Next Generation TOEFL. The results indicated that test performance was successfully analyzed using CDA. Questionnaires and interviews were also adopted for teachers and students to provide feedback. The majority of them praised the usefulness of the diagnosis. However, the teachers noted that the effect also relied on other variables such as pedagogical approaches. It is still not known how to make the feedback compatible with teaching and learning and how effective these remedial measures are.

Overall, despite these promising results, further research should be undertaken to confirm the accuracy and applicability of the cognitive diagnostic feedback provided by retrofitting large-scale language assessment. Moreover, previous studies have tended to focus on English as L2 rather than other languages. This study therefore aimed to examine whether CDA can provide precise and fine-grained feedback and compatible suggestions for improvements in teaching and learning Spanish. Hence, two assumptions were made:

CDA can accurately estimate the reading performance of large-scale, high-stakes proficiency test of Spanish.

The feedback provided by CDA is applicable for Spanish instruction and learning practice.

## 2. Cognitive Diagnostic Assessment

CDA are cognitively grounded, diagnostic procedures, which aim to provide formative diagnostic feedback through fine-grained reporting of test takers' mastery profiles (Buck & Tatsuoka, 1998; Gierl, 2007; Jang, 2005; Lee & Sawaki, 2009; Min & He, 2021).

CDA consists of four main steps. First, it is necessary to identify the attributes required to successfully answer each item of the assessment. Empirical evidence of cognitive operations involved in the thinking process is collected through inspection of the assessment syllabus, a think-aloud protocol, or expert analysis of assessment content. Second, according to the specification of attributes, an arrangement of numbers in a pattern from top to bottom and from left to right, namely a Q-matrix, is built. This kind of coding scheme shows all attributes required to correctly answer the item. In the third step, the data analysis is carried out with an appropriate Cognitive Diagnostic Model (CDM). According to Lee and Sawaki (2009: 178), CDMs are "essentially classification algorithms and thus focus on classifying examinees into latent classes based on attribute mastery patterns." A reasonable classification should be grounded on the model-data fit. Finally, the analysis provides results including general mastery probability of the attributes, classification of the attribute profiles and individualized patterns. The strengths, weaknesses and remedial strategies are eventually reported.

An example will be given to illustrate how fine-grained feedback is obtained in CDA. It is generally accepted that an item such as "Which of the following might be the best title?*"* in the reading section mainly measures the skill of summarizing. Accordingly, in a traditional test analysis, a wrong answer indicates incompetency in this specific skill. However, when the test taker tries to answer, other skills, such as skimming and scanning the introduction and concluding sentences to find the main aim, or deducing the passage content according to the possible title options, are likely involved. Therefore, an incorrect answer is not sufficient to prove the mastery level of these skills. One test taker who fails in skimming, another one who doesn't make a right deduction, or even a third one who is unable to apply both methods, may give the same response. In this sense, CDA is able to provide more specific results because it measures all the skills related to a single item and analyzes the mastery probability of certain skills by synthesizing all of the results of the relevant items. In this way, diversified and individualized feedbacks will be provided to the test takers who score the same.

## 3. Methodology

### 3.1. Participants

A total of 2046 Chinese test takers at 53 universities in China took EEE test (*Examen de Especialidad Española* in Spanish) by early 2016. The test takers consisted of 483 males and 1563 females, aged 21 to 22. This gender distribution is caused by the "very high proportion of women majoring in language studies" (Zheng & Liu, 2015: 284). All of the test takers were Chinese undergraduates majoring in Spanish. Their reasons for taking the test varied: planned for a graduate project, to become licensed to practice a profession related to teaching Spanish, and to evidence proficiency to the future employers.

The reasons for taking EEE test takers as the subjects were as follows: On the one hand, among the large-scale tests of Spanish in China, the construct of the EEE is consistent with what it's supposed to be measure in CDA. On the other hand, although the results are published in the last semester, they are still very valuable, especially for test takers who are writing or revising their BA thesis and their advisors who intend to supervise the drafts. The remedial functions of the test can be fulfilled for many test takers who are planning for graduate projects or education jobs. According to Zheng and Liu (2015: 288), among the EEE test takers, 24.3% of them planned for graduate study, and 2% of them expected to become teachers; therefore the feedbacks would improve their future academic reading and writing.

Test takers who did not give any response were excluded because the blank examination papers may be related to absence or misunderstandings of the instructions, instead of a lack of certain skills. A cohort of 1933 test takers was finally selected from the candidates from 51 universities. The sample was representative with respect to gender and age; 19.19% of the participants were males and 80.81% were females, aged 21 to 22.

### 3.2. Instruments

The analysis was carried out using two instruments: the reading section of the EEE test and the Generalized Deterministic Inputs, Noisy and Gate (G-DINA) model of CDA.

#### *3.2.1. EEE test*

The EEE test is a nationwide, standardized, large-scale test for Spanish majors, administered on behalf of the Higher Education Department, Ministry of Education of People' s Republic of China. The first undergraduate Spanish major program was initiated in China in 1952. The number of related institutions grew rapidly to 60 in 2015. A similar growth pattern can be identified for the number of Spanish major undergraduates, which increased to 14000 in 2015 (Zheng & Liu, 2015: 2). This context has required an increasing need for decisions including program exit, admission, scholarship selection, promotion, etc. To fulfil these demands, in 1999, the EEE was administered nationwide. The purpose is to measure the Spanish language proficiency of Chinese undergraduates majoring in Spanish Language and Literature and to examine whether they meet the required levels as specified in the National College Spanish Teaching Syllabus for Spanish Majors (NACFLT, 2000). At present in China, the EEE is the only large-scale assessment of Spanish students approved for undergraduate achievement and academic purposes, and it is commonly accepted as a proof of proficiency for employment and promotion. The EEE test has two levels: level 4 for the foundation stage and level 8 for the advanced stage. In 2015, approximately 4327 students took the EEE level 4 and the EEE level 8 was administered to about 2049 students.

In each administration, the test measures language knowledge and skills in different sections, ranging from listening, reading, writing, translation, language and world knowledge. In the current study, the 2016 version of the EEE level 8 was adopted. Analysis of the validity and reliability was conducted. The results showed that the internal consistency coefficient was 0.912, the test-retest coefficient was 0.903, and the criterion-related validity coefficient was 0.922.

The reading section adopted in this study included 4 passages. It comprised 20 items measuring 5 skills: comprehending vocabulary and syntactic structures, skimming and scanning, interpreting explicit information, deducing ideas and summarizing ideas. The test included four-option multiple choice questions and the total score was 20 points.

#### *3.2.2. G-DINA model*

There are several models for investigating the cognitive diagnosis of L2 reading performance. In this study, the G-DINA model developed by de la Torre (2011) was chosen. As indicated by de la Torre (2011: 196), the G-DINA model is "an interpretable model based on the identity link function and represents one of the many alternative general CDM formulations." Thus, the G-DINA model allows flexible transformation and comparisons among different models. "The required attributes for item j can be represented by the reduced vector $a^*_{\alpha ij}$ = $(\alpha_{l1},\cdots,\alpha_{lk^*_j})$', where l+1,...$2^{\kappa j}$ and $2^{\kappa j}$ represent the number of unique attribute patterns. The probability that examinees with reduced attribute vector $\alpha^*_{lj}$ will answer item j correctly is expressed as P($X_j$=1|$\alpha^*_{lj}$)=P($\alpha^*_{lj}$)" (Chen & Chen, 2016). As de la Torre (2011: 95) indicated, the item response function is as follows.

$$P\left(\alpha^*_{ij}\right) = \delta_{j0} + \sum_{k-1}^{k^*_j} \delta_{jk}\alpha_{lk} + \sum_{k'=k+1}^{k^*_j} \sum_{k-1}^{k^*_j-1} \delta_{jkk'}\alpha_{lk}\alpha_{lk'} \ldots + \delta_{j12}\cdots k^*_j \prod_{k-1}^{k^*_j} \alpha_{lk}$$

The G-DINA model was chosen because it has attractive features. First, it is suitable for dichotomously scored items. Another popular model for this kind of scoring is Deterministic Inputs, Noisy and Gate (DINA) model. As indicated by de la Torre (2009: 117), the probability of examiee *i* with the skills verctor answering item *j* correctly is given by:

$$P_j(\boldsymbol{\alpha}_i) = P(X_{ij} = 1|\boldsymbol{\alpha}_i) = g_j^{1-\eta_{ij}}(1-s_j)^{\eta_{ij}}$$

Second, the compensatory nature of the G-DINA model offers another advantage. To correctly answer the items, the mastery of all the attributes is not necessary. Some mastered attributes may compensate for non-mastered ones. In this case, the DINA model is not appropriate because, with this model, only the test takers who master all the attributes can give the right answer.

Finally, according to previous research such as de la Torre and Douglas's (2004) study, the accuracy of the G-DINA model is high. According to Li et al.'s (2016: 13) study, "G-DINA is expected to produce the best absolute fit and hence classification accuracy". Thus, the G-DINA model can be useful for gaining insights through accurate diagnostic information.

### 3.3. Procedure

The process of the current study was divided into the following 4 steps.

In the first step, the EEE test guidelines were analyzed to determine the specific reading skills assessed, and the results showed that 5 skills were involved: comprehending vocabulary and syntactic structures, interpreting explicit information, skimming and scanning, deducing ideas and summarizing ideas. On the other hand, the guidelines also indicated the target skill mainly assessed by each item. 5 experts coded the reading skills required for each item. There were several shared characteristics of these experts: they were all aged 35-45, had 5 or more years teaching reading courses, and had a post-graduate qualification. Two of them were also expert EEE test raters. The methodology in Chen and Chen's study (2016) was adopted. A list of the 5 skills was provided. Afterwards, the experts selected the skills required to answer each item correctly. The skills for which consensus of more than three experts was reached were added to the Q-matrix.

In the second step, through the integration of the results of the first step, it was possible to define an appropriate Q-matrix as displayed in the following table. In this table, 1 indicates that the item aims to measure the skill, and 0 indicates that it does not.

**Table 1.** *Q-matrix*

| Item No. | Skill1 Comprehending | Skill2 Interpreting | Skill3 Skimming/Scanning | Skill4 Deducing | Skill5 Summarizing |
|---|---|---|---|---|---|
| 1 | 1 | 0 | 0 | 1 | 0 |
| 2 | 1 | 1 | 0 | 0 | 0 |
| 3 | 1 | 0 | 1 | 0 | 0 |
| 4 | 1 | 0 | 1 | 1 | 0 |
| 5 | 0 | 1 | 0 | 1 | 1 |
| 6 | 1 | 0 | 0 | 1 | 0 |
| 7 | 1 | 1 | 0 | 0 | 0 |
| 8 | 1 | 0 | 1 | 0 | 0 |
| 9 | 1 | 0 | 1 | 1 | 0 |
| 10 | 0 | 1 | 0 | 1 | 1 |
| 11 | 1 | 0 | 0 | 1 | 0 |
| 12 | 1 | 1 | 0 | 0 | 0 |
| 13 | 1 | 0 | 1 | 0 | 0 |
| 14 | 1 | 0 | 1 | 1 | 0 |
| 15 | 0 | 1 | 0 | 1 | 1 |
| 16 | 1 | 1 | 0 | 1 | 1 |
| 17 | 1 | 1 | 0 | 1 | 1 |
| 18 | 1 | 1 | 0 | 1 | 1 |
| 19 | 1 | 1 | 0 | 1 | 1 |
| 20 | 1 | 1 | 0 | 1 | 1 |

In the third step, a diagnostic analysis was conducted using the G-DINA model and the specified Q-matrix. In the fourth step, the diagnostic analysis provides feedback. To prove the accuracy and applicability of the results, via the case study method, the parts related to academic reading in the students' BA thesis drafts were also examined.

## 4. Results and discussion

### 4.1. Model fit

Several fit statistics were considered, with the purpose of demonstrating good model fit, namely: Akaike's information criterion (AIC), the sample size adjusted AIC (AICc) and the Bayesian information criterion (BIC), all of which are useful in selecting the correct model. The smaller the indices are, the better the model fit. The AIC, AICc, BIC of the

G-DINA model were 47737.15, 47774, and 48739.08, respectively. As mentioned above, the DINA model is not suitable for this study and the DINA model fit parameters provide more evidence for this assertion. The relative fit indices of G-DINA were better than those of DINA (AIC=48006.45, AICc=48009.85, BIC=48318.16). In addition, G-DINA model's least information criterion was also preferable. The max($X^2$) was 9.93, which indicated a significant model fit (p=0.31). However, the max($X^2$) of the DINA model was 26.19, and the model misfit was significant (p=0.00). These evidence provided positive evidence that substantiated the quality of the G-DINA model.

### 4.2. General skill mastery probability

General skill mastery probability refers to the probability that all the participants would master a certain skill. Table 2 shows the general mastery probability of the five skills for all the participants. This table shows that 62% of all the test takers mastered the skill of interpreting explicit information. Compared with other skills, interpreting showed the best performance. Approximately 54% and 46% of all the participants mastered skimming/scanning and summarizing ideas. A total of 42% and 43% of the test takers mastered deducing and comprehending, respectively.

According to Brown (2000), comprehending and interpreting explicit information are micro-skills, while skimming and scanning, deducing and summarizing are macro-skills. The average mastery of micro-skills was better than that of macro-skills, which means that the test takers accomplished the reading tasks better in a bottom-up process, such as understanding the meaning and explaining meanings, rather than through a top-down process, such as glancing through the passage to identify or retrieve the details, drawing conclusions by reasoning or giving a brief statement of the main point.

**Table 2.** *Skill mastery probability*

| Skill | Skill mastery probability |
|---|---|
| Comprehending vocabulary and syntactic structures | 0.43 |
| Interpreting explicit information | 0.62 |
| Skimming and scanning | 0.54 |
| Deducing ideas | 0.42 |
| Summarizing ideas | 0.46 |

### 4.3. Classification of skill profiles

Table 3 presents an overview of the groups clustered according to their specific skill profiles. As shown in the first column, the test takers were classified into 32 latent groups. Data in the second column illustrate the representative skill profile of each group. The number 0 indicates incompetency in the skill and number 1, indicates mastery. The order of the skills represented in the profiles is as follows: comprehending, interpreting explicit information, skimming and scanning, deducing ideas and summarizing ideas. For example, in the third row, 01110 shows a skill profile with mastery in interpreting explicit information, skimming/scanning and deducing ideas, but incompetency of comprehending and summarizing ideas. The third column shows the number of cases clustered in the group. As shown in Table 3, skill profiles 11111 and 00000 had the highest frequencies. These test takers represented one-quarter of all the participants. The frequencies of profiles 01110, 11110, 01000 were also high. Test takers with these 5 profiles represented more than a half of all the participants. Notably, the results show that 10.67% of the test takers could not master none of the skills. Among the half of test takers, the most prominent weaknesses were summarizing ideas and comprehending.

**Table 3.** *Skill profile classes*

| Latent class | Skill profile | Frequency | Percentage |
|---|---|---|---|
| 1 | 11111 | 215 | 11.13% |
| 2 | 00000 | 206 | 10.67% |
| 3 | 01110 | 190 | 9.84% |
| 4 | 11110 | 175 | 9.06% |
| 5 | 01000 | 148 | 7.66% |
| 6 | 00001 | 128 | 6.63% |
| 7 | 00101 | 114 | 5.90% |
| 8 | 10001 | 107 | 5.54% |
| 9 | 01100 | 67 | 3.47% |
| 10 | 01101 | 63 | 3.26% |
| 11 | 11010 | 63 | 3.26% |
| 12 | 11000 | 63 | 3.26% |
| 13 | 01111 | 58 | 3.00% |
| 14 | 11001 | 50 | 2.59% |
| 15 | 10000 | 43 | 2.23% |
| 16 | 00100 | 38 | 1.97% |
| 17 | 01001 | 29 | 1.50% |
| 18 | 11101 | 27 | 1.40% |
| 19 | 10111 | 27 | 1.40% |
| 20 | 10101 | 24 | 1.24% |
| 21 | 01010 | 17 | 0.88% |
| 22 | 11011 | 16 | 0.83% |
| 23 | 00110 | 16 | 0.83% |

| Latent class | Skill profile | Frequency | Percentage |
|---|---|---|---|
| 24 | 00111 | 15 | 0.78% |
| 25 | 11100 | 7 | 0.36% |
| 26 | 10110 | 7 | 0.36% |
| 27 | 10010 | 6 | 0.31% |
| 28 | 10011 | 5 | 0.26% |
| 29 | 00010 | 4 | 0.21% |
| 30 | 01010 | 3 | 0.16% |
| 31 | 01011 | 1 | 0.05% |
| 32 | 00011 | 1 | 0.05% |

### 4.4. Skill profiles of university learners

Cases of test takers from a university were provided to verify the individualized skill profiles. In the third column of Table 4, number 0 and 1 have the same meanings as shown in Table 3 and the sequence of the skills remains the same.

As shown in Table 4, 39 test takers in one of the universities could be clustered into 9 groups. The data showed that 23 learners mastered all of the skills. The numbers of test takers with incompetence in comprehending, interpreting, skimming/scanning, deducing and summarizing were 2, 2, 2, 4 and 3, respectively. Therefore, at this university the test takers' major strength was related to micro-skills, and a handful of learners' weaknesses were macro-skills, which was inconsistent with the nationwide results. The mastery level of all the skills was superior to the national average level, especially with respect to comprehending, interpreting and skimming/scanning, therefore the pedagogical or learning goals of reading for this university would be expected to be different from the national standards.

**Table 4.** *Profile classes of a university*

| Latent class | Number of students | Skill profile |
|---|---|---|
| 1 | 23 | 11111 |
| 2 | 5 | 11110 |
| 3 | 3 | 10111 |
| 4 | 2 | 11011 |
| 5 | 2 | 11100 |
| 6 | 1 | 00101 |
| 7 | 1 | 11000 |
| 8 | 1 | 01111 |
| 9 | 1 | 11101 |

### 4.5. Case studies: accuracy and applicability of the diagnostic results

The above evidence was insufficient to claim that the diagnostic results were accurate and remedial in real practice. Case studies of the learners' BA thesis drafts were also conducted. The accuracy would be supported if the same skill deficiencies were identified in the literature review parts related to the references as were identified in the diagnostic reports. In this way, the results would be useful for the advisors and the learners to better review the previous literature.

According to the results in row 2 of Table 4, learner S1's skill profile was 11110, which indicates the only weakness was summarizing. In the draft of S1's BA thesis, evidence was found to verify this incompetence. The following summary of the conversation between Gabriel García Márquez and *Subcomandante Marcos* described in S1's thesis can be used to illustrate this issue:

> *En 2001, la revista colombiana Cambio con García Márquez le hizo una entrevista. En lo siguiente están unas partes de la entrevista:*
>
> – *¿Todavía, en medio de todos esos rollos, tiene tiempo para leer?*
>
> – *Sí porque, si no..., ¿qué hacemos? En los ejércitos de antes, el militar aprovechaba el tiempo para limpiar su arma y rehacerse de parque. En este caso, como nuestras armas son las palabras, tenemos que estar pendientes de nuestro arsenal a cada momento…*
>
> – *Todo lo que dice, la forma en que lo dice y el contenido, demuestran una formación literaria muy seria y muy antigua. ¿Cómo se hizo y de dónde salió?*
>
> – *De una u otra forma adquirimos la conciencia del lenguaje como una forma no de comunicarnos sino de construir algo. Como si fuera un placer más que un deber. Cuando viene la etapa de las catacumbas, frente a los intelectuales burgueses, la palabra no es lo más valorado. Queda relegado a un segundo plano. Es cuando llegamos a las comunidades indígenas, cuando el lenguaje llega como una catapulta. Te das cuenta de que te faltan palabras para expresar muchas cosas y eso obliga a un trabajo sobre el lenguaje. Volver una y otra vez sobre las palabras para armarlas y desarmarlas… Se nota que Marcos tiene un afán por la literatura y conoce muy bien el poder de palabras. Y también sabe muy bien cómo manejar las palabras.*

Evidently, the last paragraph was a short summary of the interview, in which García Márquez asked if *Subcomandante Marcos* had time available to read and how could literacy be cultivated. Marcos stated that, reading was indispensable because words were their weapons, especially when they needed to communicate with aboriginal people. Therefore, he treated reading as duty or responsibility. However, as S1 summarized, Marcos was eager to read and desired to read the literature. This was not consistent with the statements in the interview. However, S1 gave a good summary emphasizing the forces of language. There was a partial mistake in the summary because S1 added unrelated background knowledge. It was inappropriate to conclude from this interview that Marcos was eager to read literature, even though this was definitely true according to his biographic information. The diagnostic findings seemed consistent with the summary part. Thus, the CDA results provided to S1, who was at that time revising the literature review part, would indicate his/her weaknesses in summarizing, and would suggest the needs for remedial strategies.

Other cases may similarly indicate the accuracy of the results. Another interesting result of the skill profiles was the identification of incompetence in comprehending vocabulary/ syntactic structures. Row 9 of Table 4 shows that S2's profile was 01111. Evidence was found also in S2's thesis draft. In the following example, S2 mentioned the features of women's speech described by Robin Lakoff (1980) and tried to ascertain this assumption in a case study of the conversations in *La casa de los espíritus*, the debut novel of Isabel Allende. The following provides an example.

> *Lo siguiente es una lista de los rasgos del "lenguaje femenino" resumida por Lakoff (1980):…Las mujeres utilizan más diminutivos y eufemismo que los hombres… He aquí las palabras diminutivas y su distribución en las conversaciones de La casa de los espíritus. Los hombres utilizan: hijita (4 veces), señorita (4 veces), maldito,a (4 vece), chiquillo(2 veces), casita (1 vez). Mientras que las mujeres usan: hijita (1 vez), señorita(1 vez), viejitos (2 veces), chiquilla (2 veces), perrito (1 vez), Clarita (1 vez), Miguelito (1 vez), Angelito (1 vez). Es obvio que los personajes femeninos utilizan más diminutivos que los masculinos, sea en el número o en los tipos. Los diminutivos se usan como matiz de tamaño pequeño o de poca importancia, o bien para dar expresión de cariño o afecto. El primero uso es despreciativo mientras que el segundo, apreciativo. A las mujeres del libro les gusta llamar a la gente con la forma diminutiva. "Hijita", "señorita", "chiquilla", "Angelito", "Clarita", "Miguelito" son vocativos y aparecen 15 veces en total. Esto indica que ellas muestran más cariño en el trato social y pretenden mantener buenas relaciones con los demás. Sin embargo, el uso de los hombres no es así. Utilizan "maldito" o "maldita" cuatro veces para insultar algo o a alguien.*

According to S2, the results proved that female interlocutors used more kinds of diminutives than males. The word *maldito* is not a diminutive, but the past participle of the verb *maldecir*, which means "curse" in English. S2 wrongly assumed that the words with the representative suffix were definitely diminutive. He/she misunderstood the irregular past participle *maldito*, as "*–ito*" is not necessarily a suffix added to a word to show affection or convey the smallness of the object.

The qualitative data from the thesis draft showed S2's incompetency in comprehending. The accuracy of the CDA results was further proven by the triangulation method. In this case, the diagnostic results would prompt S2 and the advisor to check and correct the misunderstandings in the review part.

In addition to the learners with profiles with incompetence in only one skill, there were 3 learners who failed to master two or more skills. As shown in row 6 of Table 4, S3's weaknesses consisted of comprehending, interpreting explicit information and deducing implicit ideas. These results were also triangulated with qualitative information. The following example provides some evidences.

> *En este estudio, comparamos las formas de citación en las noticias de China y España. Usamos estilo directo (ED), estilo indirecto (EI) y estilo indirecto libre (EIL) para definir los tres tipos de citas. El ED reproduce las palabras, las repite literalmente. Se ponen las palabras del hablante entre comillas y no se cambian ni el tiempo verbal ni el pronombre demostrativo. Por otro lado, el EI no aparece entre comillas. Por eso, habrá cambios de forma de lo que ha dicho el hablante durante el proceso de citar en estilo indirecto. El EI no mantiene estable más que el contenido del discurso citado: es una interpretación del discurso citado y no su reproducción (Karam, 2006). El Estilo Indirecto Libre es un concepto un poco más complejo. Según Alcina y Blecua (1985) en su Gramática Española:*
> *-La expresión literaria y parcialmente la expresión hablada, en ocasiones, reproducen lo dicho por alguien sin acudir a verbos modales, empleando, característicasmente, las mismas trasposiciones verbales propias del estilo indirecto, por mera yuxtaposición al discurso del narrador o bien reproduciendo, sin más, las mismas palabras del enunciado que se traslada. Se conoce esta construcción con el nombre de estilo indirecto libre. El Estilo Directo se caracteriza por frases completas entre comillas, el Estilo Indirecto no suele estar entre comillas mientras el Estilo Indirecto Libre a veces tiene una parte entre comillas, pero se mezcla con la interpretación u opinión del autor.*

S3 aimed to compare and distinguish three concepts: direct speech, indirect speech and free indirect speech. Most of us have a general knowledge of the first two kinds of speech, therefore S3 only quoted two definitions of free indirect speech. S3 tried to explain the definitions and deduce the differences. Interpreting explicit information and deducing ideas are involved in this process of comprehending. The following part of Alcina and Blecua's (1985) definition is essential: *mera yuxtaposición* means a transposition of the statement without any quotation marks. However, according to S3, "*el Estilo Indirecto Libre a veces tiene una parte entre comillas.*" S3 failed to understand and interpret the meaning of free indirect speech. On the other hand, according to the definition of *Esbozo de una nueva gramática de la lengua española* (1973), what the author adds is the relative grammatical changes compared with the original statements in direct speech. The expressive elements are still retained. From these changes it is inappropriate to deduce that "*se mezcla con la interpretación u opinión del autor*". Whatever comments or statements the author adds to the discourse, it is not an essential difference among these speeches. Thus, S3 also failed to distinguish the three concepts by making inferences.

The data showed that, S3 misunderstood the definition and did not make accurate explanations or inferences of the differences. However, on the basis of what he/she understood, the summary was relatively clear. The remaining weaknesses were skills such as comprehending, interpreting and deducing.

## 5. Conclusion

Prior studies have noted the importance of using CDA in language assessments to provide a precise picture of the individualized strengths and weaknesses of every test taker. However, very little was found in the literature on the question of how to prove whether these diagnostic results are accurate, reliable and applicable for teaching and learning purposes.

The first aim of the present research was to verify whether applying CDA in a large-scale Spanish language test could accurately estimate individualized reading skill profiles. The results showed that the G-DINA model was appropriate for providing personalized diagnostic reports of the EEE test, which showed differences comparing with the overall performance. interestingly, the triangulating the diagnostic results and the learners' performance in reviewing literature in their BA thesis drafts showed that the weaknesses and strengths revealed in the review part were consistent with the findings. Thus, with respect to the first research question, CDA performed better because it reliably and accurately determined the skill profile of each learner, evidencing his/her mastery level of the skills.

Earlier findings have shown that some cognitive diagnostic reports are not completely linked with follow-up teaching and learning practice, thus an issue emerging from these findings was related to the applicability of diagnostic results. The current study aimed to evaluate the usefulness of the feedback information in teaching and learning contexts. The results of the case studies of thesis writing showed that the diagnostic reports could provide insights into remedial suggestions for reviewing and discussing the literature. If the student's diagnostic feedback indicates specific reading weaknesses, the teachers may advise and provide on what to do and what not to do before the students start reviewing the literature. For example, for students with incompetence in comprehending, their attention should be focused primarily on understanding the key terms and points of the literature. For students who fail to deduce ideas, argument mapping training could be conducted prior to review and discussion to improve inference-making. Similarly, students can also predict what mistakes they might make and try to be ready to handle them.

The diagnostic reports can guide the teachers to accurately determine and adjust their teaching purposes, changing general purposes targeted at a whole group into individualized purposes for each learner. In this way, education can be catered to all abilities. On the other hand, learners, taking in consideration their characteristics, can also continually adjust their learning aims, strategies, tasks and carry out self-assessments. The CDA contributes to a more formative process rather than a summative process.

A major limitation of this study was the method used to determine the Q-matrix which revealed the skills required to successfully answer each item. Further experimental investigations should be carried out with Delphi method to collect experts' ideas and qualitative analysis of test takers' verbal reports. Another issue that was not addressed in this study was whether verified diagnostic feedback will be valid and reliable in diversified follow-up learning and teaching practices and will have long-term effects. A greater focus on the personalized achievements could produce interesting findings that better explain the values of CDA.

## 6. References

Abu-Alhija, F.N. (2007). Large-scale Testing: Benefits and pitfalls. *Studies in Educational Evaluation, 33*: 50-68. https://doi.org/10.1016/j.stueduc.2007.01.005.

Brown, D.H. (2000). *Principles of Language Learning and Teaching*. New York: Longman.

Buck, G. & Tatsuoka, K.K. (1995). Investigation of the linguistic, cognitive and method attributes underlying test task preference: a pilot analysis using rule space methodology. Paper presented at the Language Testing Research Colloquium, Long Beach, CA.

Buck, G., Tatsuoka, K.K. & Kostin, I. (1997). The Subskills of Reading: Rule-space Analysis of a Multiple-choice Test of Second Language Reading Comprehension. *Language Learning*, *47*, 423–466. https://doi.org/10.1111/0023-8333.00016.

Buck, G. & Tatsuoka, K.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. https://doi.org/10.1177/026553229801500201.

Chen, H. & Chen, J. (2016). Retrofitting Non-cognitive-diagnostic Reading Assessment Under the Generalized DINA Model Framework. *Language Assessment Quarterly*, *13*(3), 218-230. https://doi.org/10.1080/15434303.2016.1210610.

Chin, H., Chew, C., Lim, H.L. & Thien, L.M. (2021). Development and Validation of a Cognitive Diagnostic Assessment with Ordered Multiple-Choice Items for Addition of Time. *International Journal of Science and Mathematics Education*, (1), 137-157. https://doi.org/10.1007/s10763-021-10170-5.

De la Torre, J. & Douglas, J.A. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika*, *69*, 333-353. https://link.springer.com/article/10.1007/BF02295640.

De la Torre, J. (2008). An Empirically Based Method of Q-Matrix Validation for the DINA Model: Development and Applications. *Journal of Educational Measurement*, *45*(4), 343-362. https://doi.org/10.1111/j.1745-3984.2008.00069.x.

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. (2011). The Generalized DINA Model Framework. *Psychometrika, 76*(2): 179-199. https://link.springer.com/article/10.1007/s11336-011-9207-7.

De la Torre, J. & Minchen, N. (2014). Cognitively Diagnostic Assessments and the Cognitive Diagnosis Model Framework. *Psicoglogía Educativa*, 20, 89-97. https://doi.org/10.1016/j.pse.2014.11.001.

Haertel, E.H. (1989). Using restricted latent class models to map the skill structure of achievement items. *Journal of Educational Measurement*, *26*(4), 301-323. https://doi.org/10.1111/j.1745-3984.2009.00082.x.

Jang, E.E. (2009). Cognitive Diagnostic Assessment of L2 Reading Comprehension Ability: Validity Arguments for Fusion Model Application to *LanguEdge* Assessment. *Language Testing, 26*(1), 31-73. https://doi.org/10.1177/0265532208097336.

Kim, A. (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.

Lee, Y. & Sawaki, Y. (2009). Cognitive Diagnosis Approaches to Language Assessment: An Overview. *Language Assessment Quarterly*, 6, 172-189. https://doi.org/10.1177/0265532214558457.

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.

Li, H., Hunter, V.C. & Lei, P (2016). The Selection of Cognitive Diagnostic Models for a Reading Comprehension Test. *Language testing, 33*(3), 391-409. https://doi.org/10.1177/0265532215590848.

Min, S.& He, L.(2021). Developing individualized feedback for listening assessment: Combining standard setting and cognitive diagnostic assessment approaches. Language Testing, *38*(1), 1-27. https://doi.org/10.1177/0265532221995475.

Mislevy, R.J. (1989). Foundations of a new test theory. Educational Testing Service.

National Advisory Committee for Foreign Language Teaching. (1998). *National College Spanish Teaching Syllabus for Spanish Majors*. Shanghai: Shanghai Foreign Language Education Press.

Ranjbaran, F. & Alavi, S.M.(2017). Developing a reading comprehension test for cognitive diagnostic assessment: A RUM analysis. *Studies In Educational Evaluation*, *55*, 167-179. https://doi.org/10.1016/j.stueduc.2017.10.007.

Sawaki, Y., Kim, H. & 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*, 190–209. https://doi.org/10.1080/15434300902801917.

Tatsuoka, K.K. (1983). Rule-space: An approach for Dealing with Misconceptions Based on Item Response Theory. *Journal of Educational Measurement*, *20*(4), 345-354. https://doi.org/10.1111/j.1745-3984.1983.tb00212.x.

Tuprak, T. E. & Cakir, A. (2021). Examining the L2 Reading Comprehension Ability of Adult ELLs: Developing a Diagnostic Test within the Cognitive Diagnostic Assessment Framework. *Language Testing, 38*(1): 106-131. https://doi.org/10.1177/0265532220941470.

Von Davier, M. (2005). A General Diagnostic Model Applied to Language Testing Data (ETS Research Rep. No. RR-05-16). Princeton, NJ: Educational Testing Service.

Wang, W. & Qiu, X. (2019). Multilevel Modeling of Cognitive Diagnostic Assessment: The Multilevel DINA Example. *Applied Psychological Measurement*, *43*(1), 34-50. https://doi.org/10.1177/0146621618765713.

Wu, X. Wu, R. Chang, H. Kong, Q. & Zhang, Y. (2020). International Comparative Study on PISA Mathematics Achievement Test Based on Cognitive Diagnostic Models. Frontiers in *Psychology*. *11*,1-13. https://doi.org/10.3389/fpsyg.2020.02230.

Yi, Y. (2017). Probing the Relative Importance of Different Attributes in L2 Reading and Listening Comprehension Items: An Application of Cognitive Diagnostic Model. *Language Testing*, *34*(3), 337-355. https://doi.org/.

Zheng, S. & Liu, Y. (2015). A study of Spanish Education in Colleges and Universities in China. Beijing: Foreign Language Teaching and Research Press.

## ACKNOWLEDGEMENTS

The present study is part of the research and dissemination phase of the Project of Educational Planning in Beijing (N° CHDB2020173) and the Outstanding Talent Support Program of BFSU.