International Journal of PSYCHOLOGICAL RESEARCH

RESEARCH ARTICLE

Vol 15, N° 2
https://revistas.usb.edu.co/index.php/IJPR
ISSN 2011-2084
E-ISSN 2011-7922

Assessing Language Skills Using Diagnostic Classification Models: An Example Using a Language Instrument

Evaluación de las habilidades lingüísticas mediante modelos de clasificación diagnóstica: un ejemplo de uso de un instrumento lingüístico

Georgios D. Sideridis[1,2⋆], Ioannis Tsaousis[3], Khaleel Al-Harbi[4].

[1] *Boston Children's Hospital, Harvard Medical School, Boston, Massachusetts, USA.*

[2] *National and Kapodistrian University of Athens, Navarinou 13A, Athens, Greece.*

[3] *National and Kapodistrian University of Athens, Department of Psychology, Athens, Greece.*

[4] *Education and Training Evaluation Commission (ETEC), Riyadh, Saudi Arabia.*

Abstract.
The primary purpose of the present study is to inform and illustrate, using examples, the use of Diagnostic Classification Models (DCMs) for the assessment of skills and competencies in cognition and academic achievement. A secondary purpose is to compare and contrast traditional and contemporary psychometrics for the measurement of skills and competencies. DCMs are described along the lines of other psychometric models within the Confirmatory Factor Analysis tradition such as the bifactor model and the known mixture models that are utilized to classify individuals into subgroups. The inclusion of interaction terms and constraints along with its confirmatory nature enables DCMs to accurately assess the possession of skills and competencies. The above is illustrated using an empirical dataset from Saudi Arabia ($n = 2642$), in which language skills are evaluated on how they conform to known levels of competency based on the CEFR (Council of Europe, 2001) using the English Proficiency Test (EPT).

Resumen.
El propósito principal del presente estudio fue informar e ilustrar, mediante ejemplos, el uso de Modelos de Clasificación Diagnóstica (DCM) para la evaluación de habilidades y competencias en cognición y rendimiento académico. Un propósito secundario fue comparar y contrastar la psicometría tradicional y contemporánea para la medición de habilidades y competencias. Los DCM se describen siguiendo las líneas de otros modelos psicométricos dentro de la tradición del Análisis Factorial Confirmatorio, como el modelo bifactor y los conocidos modelos mixtos que se utilizan para clasificar a los individuos en subgrupos. La inclusión de términos y restricciones de interacción junto con su naturaleza confirmatoria permite a los DCM evaluar con precisión la posesión de habilidades y competencias. Lo anterior se ilustra utilizando un conjunto de datos empíricos de Arabia Saudita ($n = 2642$), que evalúan cómo las habilidades lingüísticas se ajustan a los niveles de competencia conocidos, basados en el MCER (Council of Europe, 2001).

Keywords.
Cognitive Diagnostic Models, Diagnostic Classification Models, CEFR, bifactor models, language.

Palabras Clave.
Modelos de diagnóstico cognitivo; Modelos de clasificación diagnóstica; MCER; Modelos bifactor; Lenguaje.

OPEN ACCESS

Manuscript received: 24-10-2021
Revised: 13-07-2022
Accepted: 20-07-2022

⋆**Corresponding author**:
Georgios D. Sideridis.
Email: Georgios.sideridis@childrens.harvard.edu

Declaration of data availability: All relevant data are within the article, as well as the information support files.

Conflict of interests: The authors have declared that there is no conflict of interest.

How to Cite:
Sideridis, G. D., Tsaousis, I., & Al-Harbi, K. (2022). Assessing Language Skills Using Diagnostic Classification Models: An Example Using a Language Instrument. *International Journal of Psychologycal Research*, *15*(2), 94–104. https://doi.org/10.21500/20112084.5657

1. Introduction

Traditional and contemporary psychometrics deal with the ordering of individuals across a single latent or more continuum of skills and competencies. However, these models fail to describe a person's strengths and weaknesses or fine-grained competencies. Thus, a series of models have been developed termed Cognitive Diagnostic models (CDMs) or Diagnostic Classification Models (DCMs). The additional goal of these models, beyond rank-ordering individuals, is the classification of mastery and non-mastery individuals on specific attributes tapping single or multiple traits (Liu et al., 2018). The methodology has been utilized across a range of skills and competencies documenting the potential benefits of the procedure (Alexander et al., 2016; Gorin & Embretson, 2006; Jang, 2009; Kaya & Leite, 2017; McGill et al., 2016). The present paper is organized along the following axes: (a) it describes the logic and reasoning of Diagnostic Classification Models (DCMs) in relation to other known models and (b) it presents an applied example of the use of DCM for the assessment of language skills and competencies (Rupp & Templin, 2008; Sessoms & Henson, 2018) concerning the CEFR framework (e.g., Alderson, 2007).

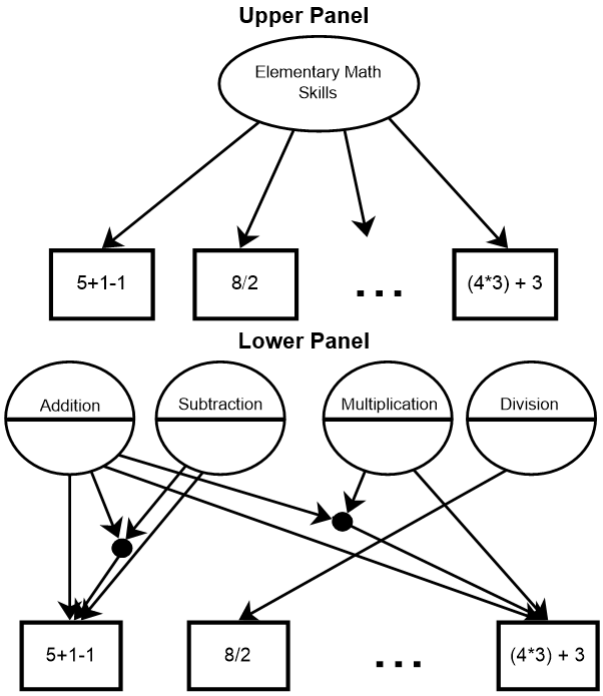
2. Diagnostic Classification Models (DCM): Description

Based on traditional modeling approaches, a person's score is comprised of a raw estimate and a standard estimate that describes the person's score in relation to the rest of the population (in normative instruments) using continuous or categorical approaches. For example, as shown in Figure 1, upper panel, a person's score could be the summed estimate of exercises designed to assess basic math skills. Using the proposed Diagnostic Classification Modeling (DCM) approach (Jurich & Bradshaw, 2014; Templin & Bradshaw, 2013; Templin & Hoffman, 2013), as shown in the lower panel of Figure 1, the competencies required to achieve the basic math exercise $5+1-1$ are both addition and subtraction. Consequently, estimation of the competencies of addition, subtraction, multiplication, and division requires the estimation of both main effects and interactions in exercises that involve multiple competencies. As a result, the conclusion derived from DCMs is one that a person is either proficient or not, in addition, subtraction, etc., but does require work on e.g., division (for an excellent discussion on DCMs see Kunina-Habenicht et al., 2009). On the other hand, the results from traditional analytical approaches (classical or contemporary) would provide information on placement only such as the person being in the 60^{th} percentile in math or having a pass/fail score in response to a categorical classification system.

Traditional modeling approaches involved the factor model, exploratory in the old days, and confirmatory

Figure 1

The logic of diagnostic classification models



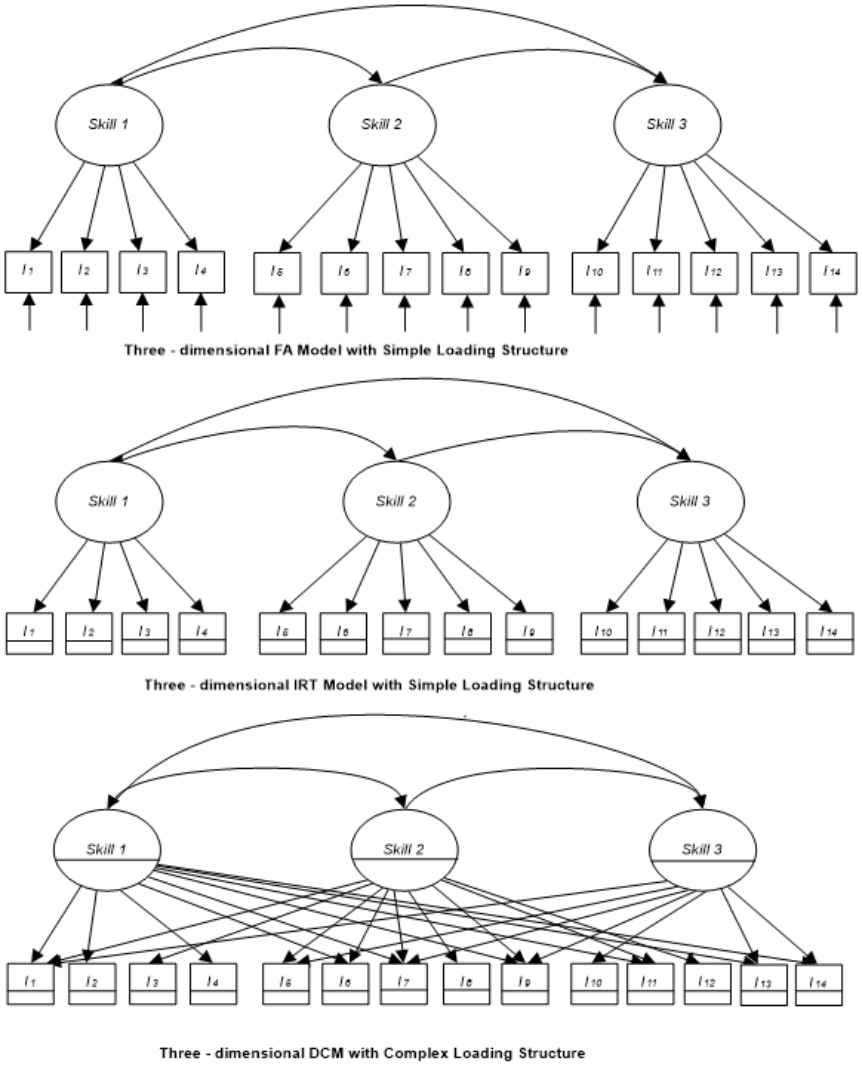
later (see Figure 2), upper panel, depicting a 3-factor correlated model of three intercorrelated skills. The middle panel of Figure 2 displays the same 3-skills structure by use of Item Response Theory (IRT) with the boxed (item estimates) containing information on item difficulty levels (crossed line shows intercepts) in addition to their link with the latent factors (slopes). The bottom panel of Figure 2 shows a complex structure, in which items define more than one skill and competencies and the circled latent variable estimates use a split line to denote threshold estimates of categorical variables denoting skill acquisition or not. This model resembles the exploratory structural equation modeling approach, recently put forth by Asparouhov and Muthen (2009).

3. Statistical Properties of Diagnostic Classification Models (DCMs)

Cognitive diagnostic models have recently received increased attention with applications across various disciplines (see Gierl et al., 2010; Tu et al., 2017; Xie, 2017; Walker et al., 2018). The first step in the development of DCMs is the creation of the Q-matrix which shows which items define which skill(s) (Chen et al., 2015; Köhn & Chiu, 2018; Liu et al., 2017; Bradshaw, 2016; Madison & Bradshaw, 2015). Table 1 shows a Q-matrix of a portion of the English Proficiency Test (EPT) measure, in which 8 items were aligned to each one of the A1 and A2 skills, as based on the Com-

Figure 2

Traditional and contemporary models for the assessment of skills and competencies. Horizontal lines within boxes reflect thresholds of categorical variables



mon European Framework (CEFR) (see Alderson, 2007; Hasselgreen, 2013; Little, 2007; Kusseling & Longsdale, 2013). These items are dichotomously scored. More information about the instrument can be found here: (https://etec.gov.sa/EN/PRODUCTSANDSERVICES/QIYAS/EDUCATION/EPT/Pages/default.aspx).

As shown in Table 1, item i19 defines only A1 skills whereas items i7, i18, i21, and i14 only A2 skills. Last, items i6, i1, and i20 define both A1 and A2 skills. The next step in the creation of a DCM model is to define parameter values for each item. As shown in Table 2, item i19 was defined to assess only the A1 skill. Consequently, it has an intercept parameter $\lambda_{i,0}$ and a main effect for skill A1 termed $\lambda_{i,1,(1)}$ (and zero terms for skill A2 and the interaction between the two skills). Item 6, which defined both A1 and A2 skills, contains an intercept term $\lambda_{i,0,}$ a main effect for the A1 skill $\lambda_{i,1,(1)}$, a

Table 1

Q-Matrix of 8 items belonging to 2 attributes in reading competency using the EPT as based on the CEFR framework

Items	A1	A2
i19	1	0
i6	1	1
i1	1	1
i20	1	1
i7	0	1
i18	0	1
i21	0	1
i14	0	1

Note. Item numbers reflect actual item numbers of EPT measure.

Table 2

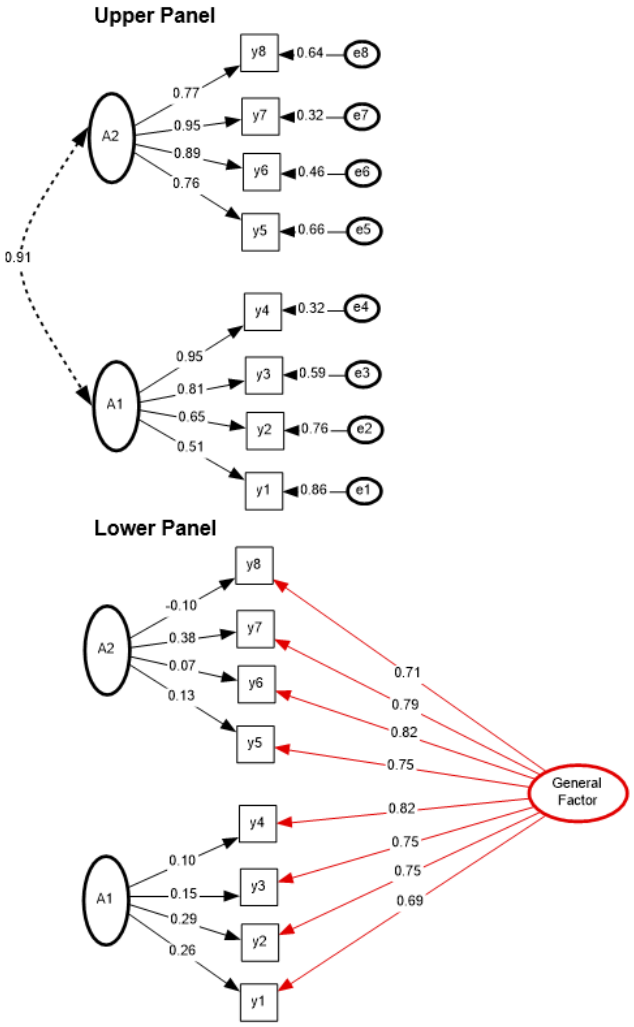
DCM Parameter Values for reading as a second language

Items	Intercept $\lambda_{i,0}$	Main Effect α_1 $\lambda_{i,1,(1)}$	Main Effect α_2 $\lambda_{i,1,(2)}$	2-Way Interaction $\lambda_{i,2,(1,2)}$
i19	1	1	0	0
i6	1	1	1	1
i1	1	1	1	1
i20	1	1	1	1
i7	1	0	1	0
i18	1	0	1	0
i21	1	0	1	0
i14	1	0	1	0

Note. 1=selected, 0=not selected.

Figure 3

Two-factor correlated model for the assessment of A1 and A2 skills of the EPT instrument (upper panel) and bifactor model extending the 2-factor correlated model with the inclusion of a general factor and two specific ones



main effect for the A2 skill $\lambda_{i,1,(2)}$, and an interaction term for A1 and A2 $\lambda_{i,2,(1,2)}$. These terms are estimated within a confirmatory latent class model using the terms in Table 3, for estimating the outcomes in each one of the latent classes (Rupp & Templin, 2008). The classes in Table 3 define each one of the four possible outcomes: the C1 class shows individuals who do not possess any of the A1 or A2 skills; the C2 class shows the presence of individuals who possess the A2 skill in the absence of A1 (a potentially undesirable finding); class C3 shows a subgroup of individuals who achieve A1 levels of proficiency; last, C4 participants are those who possess both A1 and A2 attributes. As shown in Table 3, each item contains an intercept term and then a slope if it defines any of the two skills, as well as an interaction term when that item defines both skills. For example, item 1 has an intercept term $\lambda_{3,0}$, an intercept and slope terms when defining the A1 $\lambda_{3,0}+\lambda_{3,1,(1)}$ or A2 $\lambda_{3,0}+\lambda_{3,1,(2)}$ skills, and an intercept, two linear slopes, and an interaction term $\lambda_{3,0}+\lambda_{3,1,(2)}+\lambda_{3,1,(1)}+\lambda_{3,2,(1,2)}$ when defining acquisition of both skills.

4. Diagnostic Classification Models (DCMs): An Applied Example

As described above, a DCM was fit to the data from a language instrument (for the acquisition of English as a second language) for the assessment of A1 and A2 skills based on the CEFR framework of languages. The English Proficiency Test (EPT https://etec.gov.sa/EN/PRODUCTSANDSERVICES/QIYAS/EDUCATION/EPT/Pages/default.aspx) targets at determining English language competency for individuals wishing to join academic programs taught in English. It is part of a battery of tests related to university admission and is comprised of 80 multiple-choice questions assessing three domains, namely, language structure (40 items) reading comprehension (20 Items), and written analysis (20 Items).

Participants were 2642 examinees who took on the English Proficiency Test (EPT), measure as part of their English competency exam. The specific items were de-

Table 3

DCM Kernels for each item and each one of the latent classes in Reading

	$\mathbf{C_1}$	$\mathbf{C_2}$	$\mathbf{C_3}$	$\mathbf{C_4}$
$\alpha_{\mathbf{C}}$	$[\mathbf{0},\mathbf{0}]$	$[\mathbf{0},\mathbf{1}]$	$[\mathbf{1},\mathbf{0}]$	$[\mathbf{1},\mathbf{1}]$
1. i19	$\lambda_{1,0}$	$\lambda_{1,0}$	$\lambda_{1,0}+\lambda_{1,1,(1)}$	$\lambda_{1,0}+\lambda_{1,1,(1)}$
2. i6	$\lambda_{2,0}$	$\lambda_{2,0}+\lambda_{2,1,(2)}$	$\lambda_{2,0}+\lambda_{2,1,(1)}$	$\lambda_{2,0}+\lambda_{2,1,(2)}+\lambda_{2,1,(1)}+\lambda_{2,2,(1,2)}$
3. i1	$\lambda_{3,0}$	$\lambda_{3,0}+\lambda_{3,1,(2)}$	$\lambda_{3,0}+\lambda_{3,1,(1)}$	$\lambda_{3,0}+\lambda_{3,1,(2)}+\lambda_{3,1,(1)}+\lambda_{3,2,(1,2)}$
4. i20	$\lambda_{4,0}$	$\lambda_{4,0}+\lambda_{4,1,(2)}$	$\lambda_{4,0}+\lambda_{4,1,(1)}$	$\lambda_{4,0}+\lambda_{4,1,(2)}+\lambda_{4,1,(1)}+\lambda_{4,2,(1,2)}$
5. i7	$\lambda_{5,0}$	$\lambda_{5,0}+\lambda_{5,1,(2)}$	$\lambda_{5,0}$	$\lambda_{5,0}+\lambda_{5,1,(2)}$
6. i18	$\lambda_{6,0}$	$\lambda_{6,0}+\lambda_{6,1,(2)}$	$\lambda_{6,0}$	$\lambda_{6,0}+\lambda_{6,1,(2)}$
7. i21	$\lambda_{7,0}$	$\lambda_{7,0}+\lambda_{7,1,(2)}$	$\lambda_{7,0}$	$\lambda_{7,0}+\lambda_{7,1,(2)}$
8. i14	$\lambda_{8,0}$	$\lambda_{8,0}+\lambda_{8,1,(2)}$	$\lambda_{8,0}$	$\lambda_{8,0}+\lambda_{8,1,(2)}$

Note. $\lambda_{1,0}$=intercept of item 1. The c-term denotes latent class.

signed to assess two attributes or skills, namely, A1 and A2 as per the CEFR framework. The hypothesis put forth was that there would be a distinct group possessing A1 skills, a group having both A1 and A2 skills, and, more interestingly, would examine the presence of a group that does not possess any of the two skills, called pre-A1 level, which has been observed in certain cultures. Data were analyzed using the Q-matrix in Table 1, the parameters in Table 2, and the LCDM cognitive Kernel functions in Table 3 (see also DiBello et al., 2015). Furthermore, data were also analyzed using a variable-based approach and the confirmatory factor models of 2-factor correlated, and bifactor models (see Figure 3). All models were analyzed using Mplus 8.7. An annotated syntax file using Mplus that was used for the DCM model in Figure 4 is shown in Appendix A.

As Figure 3 shows, both a 2-factor correlated model and a bifactor model provided a very good model fit using both absolute and relative criteria [2-factor model: $\chi^2(19) = 36.566$, $p = .009$, CFI=.999, TLI=.998; RMSEA=.019; bifactor model: $\chi^2(12) = 20.160$, $p = .064$, CFI=.999, TLI=.998; RMSEA=.016]. Specifically, the 2-factor correlated model showed significant factor loadings for each indicator on the respective latent construct. Furthermore, the bifactor model provided a superior model fit with the general factor being a dominant factor and the two specific factors losing most of the explanatory power. Subsequently, factor scores reflecting person's abilities were saved for further scrutiny.

For comparative purposes, data were also analyzed using a 2-class and a 3-class exploratory model with no constraints on the latent class formation, thus, subgroups emerged in an exploratory fashion. Last, when data were analyzed by use of the DCM model, results indicated the presence of three distinct subgroups: a group having neither A1 and A2 skills, termed pre-A1 level group, which comprised 830 participants, representing 31.4% of the samples' participants. A group having A2 skills in the absence of A1 was not observed, as expected, having zero participants. Two groups with A1 and A2 skills reflected 399 and 1,413 participants, which accounted for 15.1% and 53.5% of the participants. These results are shown in Figure 4 with the pre-A1 class showing response probabilities less than 50% throughout, A1 participants being successful on only the 4 specific A1 items, and A2 individuals being successful with a probability of success greater than 50% on all eight language items. Table 4 presents model comparisons across several competing models. As shown in the table and based on information criteria, the best model fit was linked to a 3-class exploratory model. However, this model was not interpretable with regard to the measurement of specific skills and competencies, that is, A1 and A2 levels, because there was a class with mixed skills that are against the logic of mastery put forth by DCM models. Consequently, the 3-class exploratory model was not deemed appropriate. From the remaining models, a superior model fit emerged for the DCM model with 3 skills, including the interesting pattern of [0,0], suggesting the absence of minimum levels of A1 skills. A similar model fit was observed by the 2-class exploratory model and the bifactor models that have a close resemblance to the DCM model but were inferior in model fit. Last, the worst fit was observed by the 2-factor correlated model.

In an attempt to compare and contrast person-based estimates from the factor model and the DCM, scatterplots were created, as shown in Figure 5. The upper panel of the figure shows factor scores based on the 2-factor correlated model, which were related to the person-based estimates, based on the latent class representing pattern [1,0]. The relationship between the two estimates was only 0.487, which is at best modest. The lower panel of the figure shows the relationship between the general factor scores, from the bifactor model (reflecting ability estimates at A1 and A2) in relation to the latent class of the DCM representing pattern [1,1]. The relationship was .927, relating the two person-based estimates, which were very high. However, the scores at A1 level skill were very discrepant between the factor model and the DCM, showing disparate estimates of person ability. Furthermore, no comparison was available for

Table 4

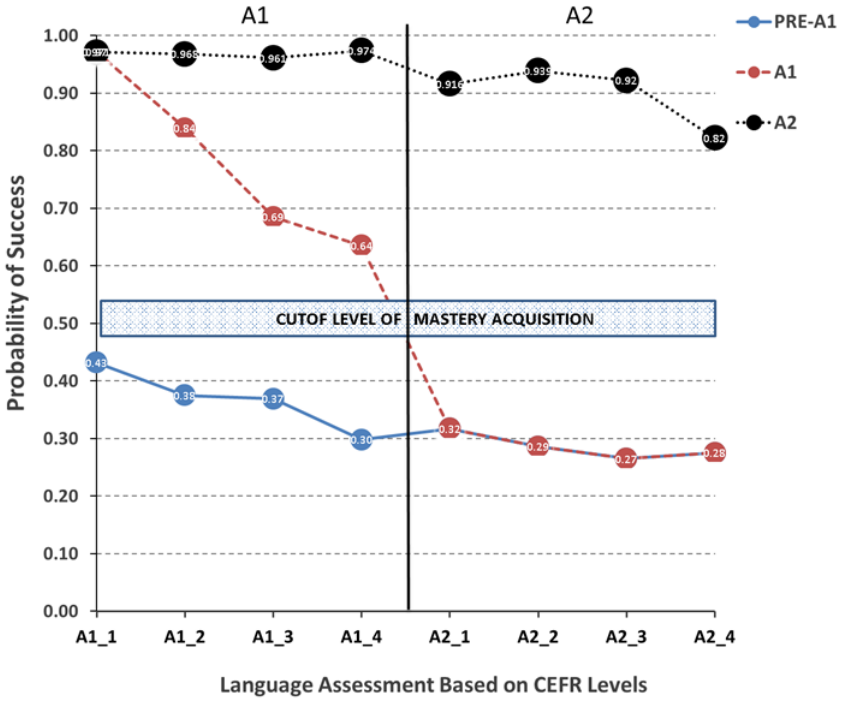
Model Fit Comparison Using Evaluative Criteria

Model Comparison	LL	Npar	AIC	BIC	SABIC	CAIC	AWE
2-Factor Correlated	−10548.79	17	21131.59	21231.53	21177.52	21248.53	21416.48
2-Class Exploratory	−10554.82	17	21143.65	21243.60	21189.58	21260.60	21428.54
Bifactor Model	−10540.11	24	21128.21	21269.31	21193.06	21293.31	21530.42
3-Class Exploratory	−10433.85	26	20919.69	21072.56	20989.95	21098.56	21355.42
DCM: 3 Skills	−10483.34	27	21020.68	21179.42	21093.64	21206.42	21473.16

Note. $\lambda_{1,0}$=intercept of item 1. The c-term denotes latent class.

Figure 4

Diagnostic cognitive model for the assessment of pre-A1, A1, and A2 skills and competencies of the EPT measure



the class lacking A1 skills (i.e., pattern [0,0]), as the factor model could not provide scores for such a subgroup.

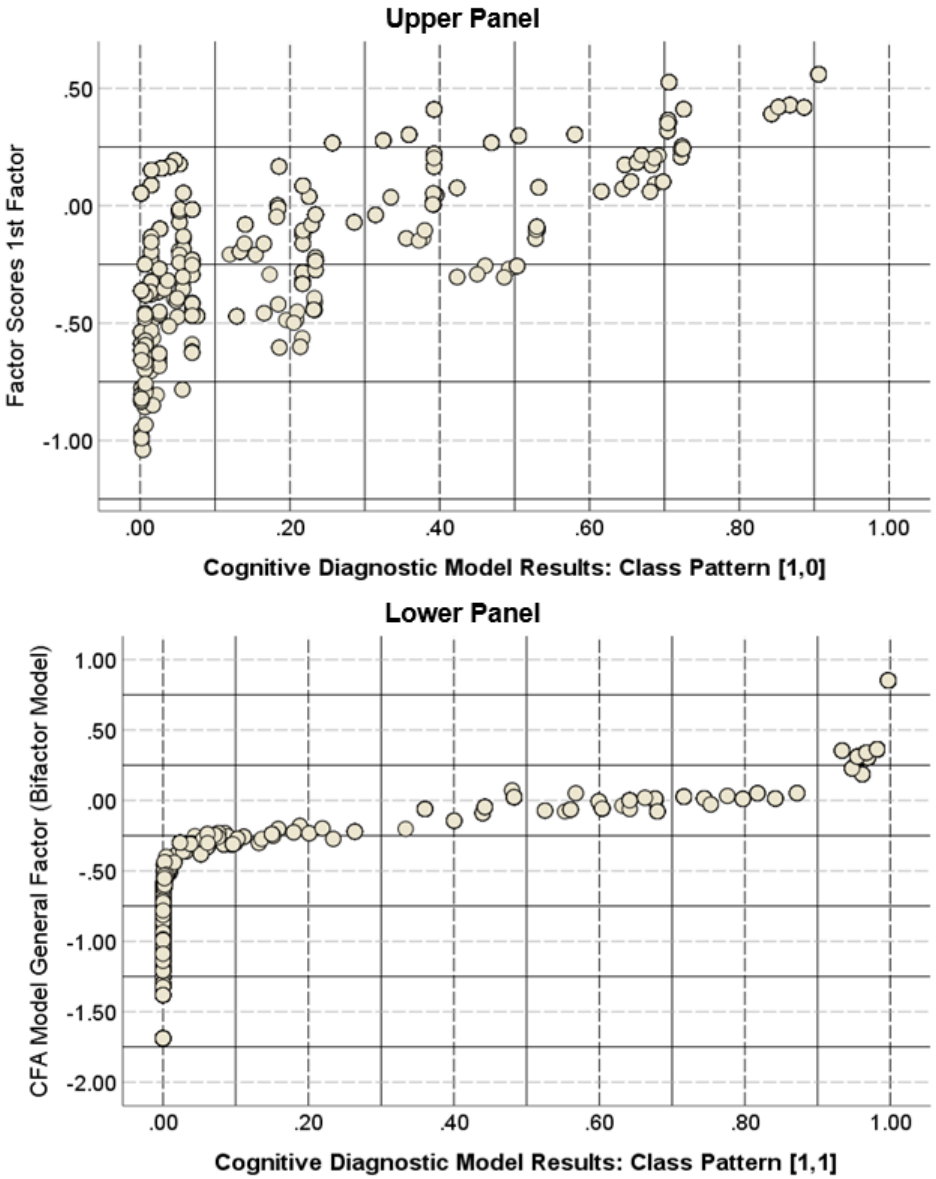
5. Conclusions Limitations and Recommendations for Future Research

The primary purpose of the present study was to inform and illustrate, using examples, the use of Diagnostic Classification Models (DCMs) for the assessment of skills and competencies in language skills and competencies. A secondary purpose is to compare and contrast traditional and contemporary psychometrics for the measurement of skills and competencies. The most important finding of the present application was that three distinct language skills groups were observed by use of the DCMs, including the recently observed pre-A1 group across various countries (e.g., Bower et al., 2017). The pre-A1 group was comprised of ample participants who did not possess the required level of A1 proficiency as delineated by the CEFR framework, certainly not in the Kingdom of Saudi Arabia.

The present findings, however, are very important, especially when contrasted with those of traditional methodologies such as the factor model (Gorsuch, 1983). Relationships between person-based estimates of ability and those from the DCM were modest, to say the least when looking at the A1 skill. Consequently, estimates of person skill acquisition by use of the factor model are clearly inappropriate in light of the advanced knowledge provided by the CDMs. A significant difference between the two is that the factor model addresses the question of degree of acquisition in total, and in the absence of requisite and other skills that extend beyond the person's level. All that information is included in the factor model and contributes to the estimation of persons' skills and competencies. In the DCMs, however, a skill is clearly defined as being dependent upon a specific set of competencies and excludes other com-

Figure 5

Comparisons between person estimates of ability as based on the factor model and the DCM model's estimates. Values on the y-axis are factor scores from the first factor (A1) of the CFA model (upper panel) and in the lower panel, factor scores from the general factor of the bifactor model in CFA



petencies that potentially confound the measurement of a person's abilities. For that reason, DCMs represent a more accurate estimate of a person's set of skills.

The present findings have several limitations. One of the potential limitations reflects a large number of available DCM models and the proper choice among them (Alexander et al., 2016; Bonifay & Cai, 2017; Bozard, 2010; Bradshaw & Madison, 2016; Bradshaw et al., 2014; Davier, 2009). A second limitation put forth by Raykov relates to the internal consistency estimates of latent subgroups reflecting specific skill levels (see Huang, 2017). A third potential limitation reflects accounting for complex structures and also the presence of covariates in the model that likely alter person's estimates of skills (Xia & Zheng, 2018). For example, in a measure of learning disabilities, how would a measure of IQ as a covariate will factor in the model? (See McGill et al., 2016). A fourth potential limitation reflects disparate opinions on what constitutes a proper measure of global fit in DCM models in light of challenges related to the number of response patterns and consequently the degrees of freedom, etc. (Hansen et al., 2016). A fifth limitation relates to the estimation of item discrimination parameters (Henson et al., 2018) and that of person-based estimates of fit (Emons et al., 2003). Last, issues

on the measurement of reliability of DCM Kernes have been raised (Templin & Bradshaw, 2013). Conversely, future directions may target at empirically investigating how to deal with these potential limitations, as well as a methodological extension such as the use of DCMs in Computerized Adaptive Testing (CAT) environments (Wang, 2013). For example, accounting for complex structures may involve simply modeling random effects to incorporating stratification weights.

References

Alderson, C. (2007). The CEFR and the need for more research. *The Modern Language Journal*, *91*, 659–663. https://doi.org/10.1111/j.1540-4781.2007.00627_4.x

Alexander, G. E., Satalich, T. A., Shankle, W. R., & Batchelder, W. H. (2016). A cognitive psychometric model for the psychodiagnostic assessment of memory-related deficits. *Psychological assessment*, *28*(3), 279. https://doi.org/10.1037/pas0000163

Asparouhov, T., & Muthén, B. (2009). Exploratory structural equation modeling. *Structural Equation Modeling*, *16*, 397–438. https://doi.org/10.1080/10705510903008204

Bonifay, W., & Cai, L. (2017). On the complexity of item response theory models. *Multivariate behavioral research*, *52*(4), 465–484. https://doi.org/10.1080/00273171.2017.1309262

Bower, J., Runnels, J., Rutson-Griffiths, A., Schmidt, R., Cook, G., Lehde, L., & Kodate, A. (2017). Aligning a Japanese university's English language curriculum and lesson plans to the CEFR-J. In F. O'Dwyer, M. Hunke, A. Imig, N. Nagai, N. Naganuma, & M. G. Schmidt (Eds.), *Critical, Constructive Assessment of CEFR-informed Language Teaching in Japan and Beyond* (pp. 176–225). Cambridge University Press.

Bozard, J. L. (2010). *Invariance testing in diagnostic classification models (Doctoral dissertation)*. The University of Georgia. https://getd.libs.uga.edu/pdfs/bozard_jennifer_l_201005_ma.pdf

Bradshaw, L., Izsák, A., Templin, J., & Jacobson, E. (2014). Diagnosing teachers' understandings of rational numbers: Building a multidimensional test within the diagnostic classification framework. *Educational measurement: Issues and practice*, *33*(1), 2–14. https://doi.org/10.1080/15305058.2015.1107076

Bradshaw, L. P., & Madison, M. J. (2016). Invariance properties for general diagnostic classification models. *International Journal of Testing*, *16*(2), 99–118. https://doi.org/10.1080/15305058.2015.1107076

Chen, Y., Liu, J., Xu, G., & Ying, Z. (2015). Statistical analysis of Q-matrix based diagnostic classification models. *Journal of the American Statistical Association*, *110*(510), 850–866. https://doi.org/10.1080/01621459.2014.934827

Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, teaching, assessment*. Cambridge University Press.

Davier, M. V. (2009). Some notes on the reinvention of latent structure models as diagnostic classification models. *Measurement: Interdisciplinary Research and Perspectives*, *7*(1), 67–74. https://doi.org/10.1080/15366360902799851

DiBello, L. V., Henson, R. A., & Stout, W. F. (2015). A family of generalized diagnostic classification models for multiple choice option-based scoring. *Applied Psychological Measurement*, *39*(1), 62–79. https://doi.org/10.1177%2F0146621614561315

Emons, W. H., Glas, C. A., Meijer, R. R., & Sijtsma, K. (2003). Person fit in order-restricted latent class models. *Applied psychological measurement*, *27*(6), 459–478. https://doi.org/10.1177%2F0146621603259270

Gierl, M. J., Alves, C., & Majeau, R. T. (2010). Using the attribute hierarchy method to make diagnostic inferences about examinees' knowledge and skills in mathematics: An operational implementation of cognitive diagnostic assessment. *International Journal of Testing*, *10*(4), 318–341. https://doi.org/10.1080/15305058.2010.509554

Gorin, J. S., & Embretson, S. E. (2006). Item difficulty modeling of paragraph comprehension items. *Applied Psychological Measurement*, *30*, 394–411. https://doi.org/10.1177/0146621606288554

Gorsuch, R. (1983). *Factor analysis*. Lawrence Erlbaum Associates.

Hansen, M., Cai, L., Monroe, S., & Li, Z. (2016). Limited-information goodness-of-fit testing of diagnostic classification item response models. *British Journal of Mathematical and Statistical Psychology*, *69*(3), 225–252. https://doi.org/10.1111/bmsp.12074

Hasselgreen, A. (2013). Adapting the CEFR for the classroom assessment of young learners' writing. *The Canadian Modern Language Review*, *69*, 415–435. https://doi.org/10.3138/cmlr.1705.415

Henson, R., DiBello, L., & Stout, B. (2018). A Generalized Approach to Defining Item Discrimination for DCMs. *Measurement: Interdisciplinary Research and Perspectives*, *16*(1), 18–29. https://doi.org/10.1080/15366367.2018.1436855

Huang, H. Y. (2017). Multilevel cognitive diagnosis models for assessing changes in latent attributes. *Journal of Educational Measurement*, *54*(4), 440–480. https://doi.org/10.1111/jedm.12156

Jang, E. (2009). Cognitive diagnostic assessment of L2 reading comprehension ability: Validity arguments for Fusion Model application to *LanguEdge* assessment. *Language Testing*, *26*, 31–73. https://doi.org/10.1177%2F0265532208097336

Jurich, D. P., & Bradshaw, L. P. (2014). An illustration of diagnostic classification modeling in student learning outcomes assessment. *International Journal of Testing*, *14*(1), 49–72. https://doi.org/10.1080/15305058.2013.835728

Kaya, Y., & Leite, W. L. (2017). Assessing change in latent skills across time with longitudinal cognitive diagnosis modeling: An evaluation of model performance. *Educational and psychological measurement*, *77*(3), 369–388. https://doi.org/10.1177%2F0013164416659314

Köhn, H. F., & Chiu, C. Y. (2018). How to Build a Complete Q-Matrix for a Cognitively Diagnostic Test. *Journal of Classification*, *35*(2), 273–299. https://doi.org/10.1007/s00357-018-9255-0

Kunina-Habenicht, O., Rupp, A. A., & Wilhelm, O. (2009). A practical illustration of multidimensional diagnostic skills profiling: Comparing results from confirmatory factor analysis and diagnostic classification models. *Studies in Educational Evaluation*, *35*(2-3), 64–70. https://doi.org/10.1016/j.stueduc.2009.10.003

Kusseling, F., & Lonsdale, D. (2013). A corpus-based assessment of French CEFR lexical content. *The Canadian Modern Language Review*, *69*, 436–461. https://doi.org/10.3138/cmlr.1726.436

Little, D. (2007). The common European framework of reference for languages: Perspectives on the making of supranational language education policy. *The Modern Language Journal*, *91*, 645–655. https://doi.org/10.1111/j.1540-4781.2007.00627_2.x

Liu, R., Huggins-Manley, A. C., & Bradshaw, L. (2017). The impact of Q-matrix designs on diagnostic classification accuracy in the presence of attribute hierarchies. *Educational and psychological measurement*, *77*(2), 220–240. https://doi.org/10.1177%2F0013164416645636

Liu, R., Huggins-Manley, A. C., & Bulut, O. (2018). Retrofitting diagnostic classification models to responses from IRT-based assessment forms. *Educational and psychological measurement*, *78*(3), 357–383. https://doi.org/10.1177%2F0013164416685599

Madison, M. J., & Bradshaw, L. P. (2015). The effects of Q-matrix design on classification accuracy in the log-linear cognitive diagnosis model. *Educational and Psychological Measurement*, *75*(3), 491–511. https://doi.org/10.1177%2F0013164414539162

McGill, R. J., Styck, K. M., Palomares, R. S., & Hass, M. R. (2016). Critical issues in specific learning disability identification: What we need to know about the PSW model. *Learning Disability Quarterly*, *39*(3), 159–170. https://doi.org/10.1177%2F0731948715618504

Rupp, A. A., & Templin, J. L. (2008). Unique characteristics of diagnostic classification models: A comprehensive review of the current state-of-the-art. *Measurement*, *6*(4), 219–262. https://doi.org/10.1080/15366360802490866

Sessoms, J., & Henson, R. A. (2018). Applications of Diagnostic Classification Models: A Literature Review and Critical Commentary. *Measurement: Interdisciplinary Research and Perspectives*, *16*(1), 1–17. https://doi.org/10.1080/15366367.2018.1435104

Templin, J., & Bradshaw, L. (2013). Measuring the reliability of diagnostic classification model examinee estimates. *Journal of Classification*, *30*(2), 251–275. https://doi.org/10.1007/s00357-013-9129-4

Templin, J., & Hoffman, L. (2013). Obtaining diagnostic classification model estimates using Mplus. *Educational Measurement: Issues and Practice*, *32*(2), 37–50. https://doi.org/10.1111/emip.12010

Tu, D., Gao, X., Wang, D., & Cai, Y. (2017). A new measurement of internet addiction using diagnostic classification models. *Frontiers in psychology*, *8*, 1768. https://doi.org/10.3389%2Ffpsyg.2017.01768

Walker, G. M., Hickok, G., & Fridriksson, J. (2018). A cognitive psychometric model for assessment of picture naming abilities in aphasia. *Psychological assessment*, *30*(6), 809–826. https://doi.org/10.1037%2Fpas0000529

Wang, C. (2013). Mutual information item selection method in cognitive diagnostic computerized adaptive testing with short test length. *Educational and Psychological Measurement*, *73*(6), 1017–1035. https://doi.org/10.1177%2F0013164413498256

Xia, Y., & Zheng, Y. (2018). Asymptotically Normally Distributed Person Fit Indices for Detecting Spuriously High Scores on Difficult Items. *Applied psychological measurement*, *42*(5), 343–358. https://doi.org/10.1177%2F0146621617730391

Xie, Q. (2017). Diagnosing university students' academic writing in English: Is cognitive diagnostic modeling the way forward? *Educational Psychology*, *37*(1), 26–47. https://doi.org/10.1080/01443410.2016.1202900

Appendix A

Mplus syntax file for DCM model of Figure 4. Model constraint statement includes only the first two items for illustration purposes

```
TITLE: EPT RC
DATA: FILE IS nRC1.dat;
VARIABLE:
     NAMES = x1-x14 id;
     USEVARIABLE = x1-x8;
     CATEGORICAL = x1-x8;
     CLASSES = c(4);
     idvariable = id;
ANALYSIS:
     TYPE=MIXTURE;
     STARTS=0;
MODEL:
%OVERALL%
[C#1] (m1); !latent variable mean for attribute pattern [0,0];
[C#2] (m2); !latent variable mean for attribute pattern [0,1];
[C#3] (m3); !latent variable mean for attribute pattern [1,0];

%c#1% !for attribute pattern [0,0];
[x1$1] (t1_1); !threshold for item 1 LCDM kernel 1
[x2$1] (t2_1); !threshold for item 2 LCDM kernel 1
[x3$1] (t3_1); !threshold for item 3 LCDM kernel 1
[x4$1] (t4_1); !threshold for item 4 LCDM kernel 1
[x5$1] (t5_1); !threshold for item 5 LCDM kernel 1
[x6$1] (t6_1); !threshold for item 6 LCDM kernel 1
[x7$1] (t7_1); !threshold for item 7 LCDM kernel 1
[x8$1] (t8_1); !threshold for item 7 LCDM kernel 1

%c#2% !for attribute pattern [0,1];
[x1$1] (t1_2); !threshold for item 1 LCDM kernel 2
[x2$1] (t2_2); !threshold for item 2 LCDM kernel 2
[x3$1] (t3_2); !threshold for item 3 LCDM kernel 2
[x4$1] (t4_2); !threshold for item 4 LCDM kernel 2
[x5$1] (t5_2); !threshold for item 5 LCDM kernel 2
[x6$1] (t6_2); !threshold for item 6 LCDM kernel 2
[x7$1] (t7_2); !threshold for item 7 LCDM kernel 2
[x8$1] (t8_2); !threshold for item 7 LCDM kernel 2

%c#3% !for attribute pattern [1,0];
[x1$1] (t1_3); !threshold for item 1 LCDM kernel 3
[x2$1] (t2_3); !threshold for item 2 LCDM kernel 3
[x3$1] (t3_3); !threshold for item 3 LCDM kernel 3
[x4$1] (t4_3); !threshold for item 4 LCDM kernel 3
[x5$1] (t5_1); !threshold for item 5 LCDM kernel 1
[x6$1] (t6_1); !threshold for item 6 LCDM kernel 1
[x7$1] (t7_1); !threshold for item 7 LCDM kernel 1
[x8$1] (t8_1); !threshold for item 7 LCDM kernel 1

%c#4% !for attribute pattern [1,1];
[x1$1] (t1_4); !threshold for item 1 LCDM kernel 4
[x2$1] (t2_4); !threshold for item 2 LCDM kernel 4
[x3$1] (t3_4); !threshold for item 3 LCDM kernel 4
[x4$1] (t4_4); !threshold for item 4 LCDM kernel 4
[x5$1] (t5_2); !threshold for item 5 LCDM kernel 2
[x6$1] (t6_2); !threshold for item 6 LCDM kernel 2
[x7$1] (t7_2); !threshold for item 7 LCDM kernel 2
[x8$1] (t8_2); !threshold for item 7 LCDM kernel 2
```

```
MODEL CONSTRAINT: !used to define LCDM parameters and constraints
!Mplus uses P(X=0) rather than P(X=1) so terms must be multiplied by -1

!ITEM 1:
!Q-matrix Entry [1 0]
!One attibute measured: 1 intercept; 1 main effect
NEW(l1_0 l1_11 l1_12 l1_212);    !define LCDM parameters present for item 1
t1_1=-(l1_0);
t1_2=-(l1_0+l1_12);
t1_3=-(l1_0+l1_11);
t1_4=-(l1_0+l1_12+l1_11+l1_212);
l1_12>0;            !the order constraints necessary for the main effect
l1_11>0;            !the order constraints necessary for the main effect
l1_212>-l1_11;      !the order constraints necessary for the interaction
l1_212>-l1_12;      !the order constraints necessary for the interaction

!ITEM 2:
!Q-matrix Entry [1 1]
!2 attibutes measured: 1 intercept; 2 main effects; 1 two-way interaction
NEW(l2_0 l2_11 l2_12 l2_212);    !define LCDM parameters present for item 2
t2_1=-(l2_0);
t2_2=-(l2_0+l2_12);
t2_3=-(l2_0+l2_11);
t2_4=-(l2_0+l2_12+l2_11+l2_212);
l2_12>0;            !the order constraints necessary for the main effect
l2_11>0;            !the order constraints necessary for the main effect
l2_212>-l2_11;      !the order constraints necessary for the interaction
l2_212>-l2_12;      !the order constraints necessary for the interaction
```