Biquarterly Journal of Cognitive Strategies in Learning *Vol 10, Num 19, Autumn & Winter 2022-2023*

https://10.22084/j.psychogy.2022.23446.2269

20.1001.1.24237906.1401.10.19.1.6

Research Articl **Page 1-23**

Challenges of modeling in Cognitive diagnostic assessment and solving them in TIMSS data

Masoud Kabiri[1*], Mahmood Ghazi-tabatabaei[2]

1. Assistant Professor, National Center for TIMSS and PIRLS, Research Institute for Education (RIE), Tehran, Iran
2. Professor, Sociology Department, University of Tehran, Tehran, Iran

Submit Date: 13 November 2021 **Revise Date:** 14 March 2022
Accept Date: 26 April 2022 **Publication Date:** 31 December 2022

Abstract

Objective: Cognitive diagnostic assessment has been introduced as a new issue in educational measurement. In this approach, more information was examined about how people learn and master cognitive attributes in school. There are several data modeling issues in cognitive diagnostic assessment due to differences with another statistical modeling.

Methods: In the present study, science data of grade eight in TIMSS was analyzed by cognitive diagnostic assessment, as an empirical example, and the problems were entitled as modeling challenges. Each challenge has been explained in order to highlight differences from the usual statistical modeling.

Results: The challenges included; unidimensionality versus multidimensionality, number of attributes, correlation between attributes, number of items in each attribute, operationalization of attribute, reliability of attribute, validity, item parameters, fit of the model, identification and specification, convergence, and complex sampling.

Conclusion: Each topic was discussed in the context of modeling TIMSS data in a science course and the experience of solving these challenges was shared.

Keywords: Evaluation, Concept map-Based Tests, Validity, Reliability.

Citation: Kabiri, M., Ghazi-tabatabaei, M., (2022). Challenges of modeling in Cognitive diagnostic assessment and solving them in TIMSS data. *Biquarterly Journal of Cognitive Strategies in Learning*, 10(19), 1-23.

***Corresponding Author:** Masoud Kabiri
E-mail: maskabiri@yahoo.com

https:// 10.22084/j.psychogy.2022.23446.2269

Extended Abstract

1. Introduction

Cognitive Diagnostic Assessment (CDA) as a new assessment approach attempts to provide rich information about the learning of individuals as well as mastery or non-mastery over cognitive competencies. The assumption of CDA is to succeed in an item, a set of unobserved sub-skills should be mastered, therefore, it is tried to model the response probability of an item as a function of latent attributes of an examinee. A CDA is designed to measure an examinee's knowledge and skills. By measuring these competencies, the examinees' cognitive strengths and weaknesses can be identified, and, thus, diagnostic inferences about their competencies is determined in response to items (Gierl, Alves & Majeau, 2010).

The objective of the diagnostic model is to draw attribute profiles (Finkelman, Kim, Roussos & Verschoor, 2010; von Davier, 2007). The simplest model is the unidimensional model that is used more commonly in Item Response Theory (IRT) which reports the performance of a student in the ability dimension. In addition, multidimensional models are used when the dimensions are calibrated simultaneously and the associations of latent variables are estimated directly. The cognitive diagnostic models are to be considered multidimensional models.
Considering objectives, perspectives, and specific characteristics of modeling of CDA, the common procedure of modeling cannot be used and we encounter specific challenges and difficulties. In the present article, we explain the major challenges of modeling in CDA

2. Results

Many of the problems in common modeling are true for the modeling CDA. However, there are more challenges in CDA due to complexities of structure, factor loadings, the interaction of attributes, shortage of proper software, and using time-consuming estimation methods.

Unidimensiality versus multidimensionality: The main problem of cognitive diagnostic modeling into TIMSS data is, like many scaling, it is designed basically upon a unidimensional approach, but the objective of CDA is to extract multiple dimensions. If the purpose is mainly to rank examinees along a single competency scale, a unidimensional continuous model may be the best solution. The challenge of CDA is whether extraction of multiple dimensions can enable it to state richer information about the performance of the examinee. It is an important issue because it is possible that a unidimensional model considers an acceptable fit like the k-dimension model.

In order to solve this problem, in the first step, the unidimensional model was modeled and then more complex models were obtained by adding more dimensions. This process continues when the more complex model is not adequate rather than the simpler one based on fit indices.

https:// 10.22084/j.psychogy.2022.23446.2269

The number of attributes: Whereas the CDA hope to extract more dimension, the question is how many dimension is adequate? In the case of identifying many attributes, they highly correlate to each other and they are not as separate as to discriminate and also a danger of non-identifiability. This leads to non-accuracy estimations (DiBello, Roussos & Stout, 2007). Despite recommending models with low dimensions, these models cause the attributes to losing their utility and interpretability. Therefore, it is essential to make a balance between the number of attributes in the case that not so many which not estimated statistically and not so little to lose the diagnostic interpretation.

In the TIMSS cognitive diagnostic modeling, the first 19 attributes were recognized by literature review and adaptation with TIMSS items, but 7 attributes were confirmed as required skills. In addition, some attributes were integrated and added to the model.

Correlation between attributes: If there are many attributes that are perfectly correlated, then reporting these attributes separately provides no real information not already provided by a single composite score (Haberman and von Davier, 2007). Also, a large degree of correlation between attributes is likely due to the large number of them (Rupp, Templin, and Henson, 2010). In our modeling of TIMSS data, the correlations between attributes were considerably low, partially because the first dimension (basic knowledge) explained much of the variances.

The number of items in each attribute: The number of items issue is related to reliability, but it should be examined more because the reliability of the subscale is lower than the composite score. There are a number of recommendations that each attribute had to be included in at least three or four items (Haberman and von Davier, 2007; Chen, et al., 2008b). However, in an international large-scale assessment like TIMSS, the number of items should be more because of matrix sampling used in the study which leads to the unbalanced number of items in the attributes. In our cognitive modeling 8 items were considered in each skill as a rule of thumb.

Grain size: Grain size refers to both the depth and breadth of knowledge and skills measured and consists of fine and coarse grain size. Cognitive models are recommended that contain attributes that are specified at the fine grain size (Leighton and Gierl, 2011), however, their appropriateness depends on the objective of the diagnostic assessment and diagnostic complexities (Rupp, Templin, and Henson, 2010). In our diagnostic model the coarse grain size attributes were used because of poor literature in science education.

Reliability of attributes: Attribute reliability refers to the precision of score decisions about examinees' attribute mastery. However, standard reliability coefficients as estimated for assessments modeled with a continuous unidimensional latent trait do not translate directly to discrete latent space modeled CDA, because there are a few items in attributes and also, for items that measure more than one attribute, each attribute only contributes to a part of the total item-level variance. In our diagnostic model, we used the output of General

https:// 10.22084/j.psychogy.2022.23446.2269

Diagnostic Model's software (von Davier, 2008) which provides the reliability of attributes.

Items Parameters: The estimates for the item parameters should be evaluated for internal consistency, reasonability, and concurrence with substantive expectations in order to reach convergence in the parameter estimation. If an attribute turned out much harder or easier than expected, the Q matrix should be revisited. In our diagnostic model the Q matrix has been changed somewhat the item parameters are estimated reasonably.

Model fit: Although parsimony is important in CDA, the aim is not to achieve the simplest model that fits the data, but rather the simplest model that accomplishes the diagnostic purpose and fits the data reasonably well. Thus, one could choose a more complex model, even if some traditional fit measures were lower (DiBello, Roussos, and Stout, 2007). In our model, we used Akaike's Information Criterion (AIC) and Bayesian Information Criterion (BIC) as well as Q-matrix revising to get the proper fit.

3. Discussion

Analysis CDA needs modeling, however, the modeling process in CDA is quite different from other assessment methods that relate to the nature of CDA especially categorical variables, multidimensional scaling, and so on. In this paper, we discuss some of the challenges of modeling in CDA.

4. Ethical Considerations

Compliance with ethical guidelines: All ethical principles are considered in this article. The participants were informed about the purpose of the research and its implementation stages. They were also assured about the confidentiality of their information and were free to leave the study whenever they wished, and if desired, the research results would be available to them.

Funding: This research did not receive any grant from funding agencies in the public, commercial, or non-profit sectors.

Authors' contributions: All authors have participated in the design, implementation and writing of all sections of the present study.

Conflicts of interest: The authors declared no conflict of interest.

دوفصلنامه راهبردهای شناختی در یادگیری
سال دهم، شماره نوزدهم، پاییز و زمستان ۱۴۰۱

Biquarterly Journal of Cognitive Strategies in Learning
Vol 10, Num 19, Autumn & Winter 2022-2023

(از ص ۵ تا ۲۳)

DOI: 10.22084/j.psychogy.2022.23446.2269
DOR: 20.1001.1.24237906.1401.10.19.1.6

مقاله پژوهشی

چالش‌های مدل‌سازیِ سنجش شناختی-تشخیصی و چگونگی رفع آن‌ها در داده‌های مطالعه تیمز

Challenges of modeling in cognitive diagnostic assessment and solving them in TIMSS data

مسعود کبیری[1*]، محمود قاضی طباطبایی[2]

دریافت مقاله: ۱۴۰۰/۰۸/۲۲ **بازنگری مقاله:** ۱۴۰۰/۱۲/۲۳
پذیرش مقاله: ۱۴۰۱/۰۲/۰۶ **انتشار مقاله:** ۱۴۰۱/۱۰/۱۰

چکیده

هدف: سنجش شناختی- تشخیصی به‌عنوان یکی از مباحث جدید در سنجش آموزشی مطرح شده است. در این روش، اطلاعات وسیع‌تری در مورد چگونگی یادگیری افراد و نحوهٔ تسلط بر مهارت‌های شناختیِ لازم برای بهبود فرایند یادگیری بررسی می‌شود. به علت تفاوت‌های سنجش شناختی-تشخیصی با مدل‌های دیگر سنجش، چالش‌های خاصی در مدل‌سازی این روش وجود دارد.

روش: به‌عنوان نمونه‌ای عملی از مدل‌سازی، داده‌های علوم مطالعهٔ تیمز با رویکرد شناختی-تشخیصی تحلیل شده و مشکلات فرایند مدل‌سازی آن مستند گردید. هر یک از چالش‌ها مورد تشریح قرار گرفته تا تفاوت‌های آن با رویه‌های مرسوم مدل‌سازی آشکار گردد.

یافته‌ها: چالش‌های مورد بحث شامل تک بُعدی بودن در مقابل چند بُعدی بودن، تعداد خصیصه‌ها، همبستگی بین خصیصه‌ها، تعداد مناسب سؤال در هر خصیصه، درجهٔ دقت خصیصه‌ها، اعتبار خصیصه‌ها، روایی سنجش شناختی-تشخیصی، پارامترهای سؤال، برازش مدل، شناسایی و تعیّن مدل، هم‌گرایی و نمونه‌گیری‌های پیچیده بودند.

نتیجه‌گیری: برای نشان دادن چگونگی حل این چالش‌ها در نمونه‌ای عملی، تجربهٔ مدل‌سازی شناختی-تشخیصی داده‌های مطالعهٔ تیمز به بحث گذاشته شد.

کلید واژه‌ها: سنجش شناختی تشخیصی، مدل‌سازی، تیمز، علوم تجربی.

۱. استادیار، پژوهشگاه مطالعات آموزش و پرورش، تهران، ایران
۲. استاد، گروه جمعیت‌شناسی دانشگاه تهران، تهران، ایران
* نویسنده مسئول
Email: maskabiri@yahoo.com

۱. مقدمه

رویه‌های مرسوم در سنجش آموزشی، همانند نظریهٔ کلاسیک و نظریه سؤال پاسخ[1] اندازه‌گیری، به دلیل توجه کم به نظریه‌های شناختی مطرح در علوم یادگیری و عدم‌تلفیق با نظریه‌های شناختی در معرض انتقاد قرار گرفته‌اند. نظریه‌های شناختی و یادگیری درباره فرایندها، راهبردها و ساختار دانشِ زیربناییِ حل سؤال اطلاعاتی فراهم می‌کنند، ولی این اطلاعات به ندرت در سنجش آموزشی استفاده می‌شوند. از این‌رو، لیتون، گیرل و هونکا[2] (۲۰۰۴) تلفیق نظریات شناختی با فرایند روان‌سنجی را به علت فقدان چارچوب مرتبط، کُند توصیف کرده‌اند. در همین راستا، به علت آنکه در سنجش‌های سنتی تنها نمرهٔ کل برای هر آزمودنی به‌دست می‌آید[3]، اطلاعات حاصل به‌عنوان برآوردهای بیرونی[4] شناخته می‌شود (چن، گورین، تامپسون و تاتسوکا[5]، ۲۰۰۸ب). در بحث هدف آزمون نیز این ایراد مطرح شده است که آزمون‌های سنتی بیشتر به‌منظور انتخاب، جای‌گذاری[6] و گواهی‌دهیِ[7] دانش و توانایی آزمودنی استفاده می‌شوند و برای مقاصد دیگرِ آزمون، هم‌چون تشخیص نقاط قوت و ضعف آزمودنی برای ارائه بازخورد جهت بهبود، مناسب نیست. لذا به‌عنوان روشی برای حل مشکلات اشاره شده در فرایند سنتی آموزشی، سنجش شناختی تشخیصی[8] پیشنهاد شده است.

سنجش شناختی-تشخیصی به‌عنوان رویکرد جدیدِ سنجش آموزشی شناخته شده است. در این رویکرد، سعی می‌شود با برطرف کردن نقاط مغفولِ نظریات رایجِ سنجش آموزشی، اطلاعات وسیع‌تری در مورد چگونگی یادگیری افراد و میزان تسلط یا عدم تسلط بر مهارت‌های شناختیِ لازم برای بهبود فرایند یادگیری–یاددهی قرار گیرد. در این شیوهٔ سنجش از طریق مدل‌سازی روابط پاسخ‌های سؤال و شایستگی‌های دانش‌آموزان در فرایندهای متفاوت شناختی، ساختار شناختی زیربناییِ نمرات آزمون آشکارشده (چن و همکاران، ۲۰۰۸الف) و مهارت‌ها یا خصیصه‌های مکنونِ لازم برای پاسخ‌گویی به سؤالات آزمون کشف شود (دی‌کارلو[9]، ۲۰۱۱). در این روش فرض می‌شود برای موفقیت در سؤال لازم است مجموعه‌ای از زیرتکالیفِ غیرقابل‌مشاهده، کسب شده باشند. بنابراین، سعی می‌شود احتمال پاسخ به سؤال به‌عنوان تابعی از خصیصه‌های مکنون یک آزمودنی مدل‌سازی شود.

ازجمله اهداف اصلی این سنجش‌ها فراهم کردن اطلاعات بیشتر در مورد آزمودنی‌هاست که قابل استفادهٔ طیف وسیعی از ذی‌نفعان باشد تا هم نحوهٔ یادگیری دانش‌آموزان آشکارشده و هم نقاط قوت و ضعف دانش‌آموزان در هر خصیصه، و نه عملکرد کلی، مشخص شود. به‌عبارت‌دیگر، سنجش

1. Item Response Theory (IRT)
2. Leighton, Gierl & Hunka
۳. به‌طور مثال در نظریه کلاسیک نمره خام و در نظریه سؤال پاسخ برآورد تتا برای هر آزمودنی محاسبه می‌شود.
4. external estimate
5. Chen, Gorin, Thompson & Tatsuoka
6. Placement
7. Certification
8. Cognitive diagnostic assessment
9. DeCarlo

شناختی-تشخیصی برای درک چگونگی عملکرد افراد در آزمون بر مبنای فرایندهای ذهنی و مهارت‌های فرایندی آنان طراحی شده است تا اطلاعاتی در مورد قوت و ضعف‌های شناختی آنان فراهم نماید. بنابراین، همه فعالیت‌ها با هدف نشان دادن عملکرد آزمودنی در سؤال براساس مهارت‌های لازم برای پاسخ‌گویی به آن صورت می‌پذیرد (روسوس، تمپلین و هنسون[1]، ۲۰۰۷). به‌طورکلی، سنجش شناختی-تشخیصی بر این فرض استوار است که در پاسخ آزمودنی‌ها به سؤالات آزمون، اطلاعات بیشتری به‌جز برآوردِ تک‌بُعدی نمرهٔ توانایی وجود دارد. لذا اطلاعاتی که از این تحلیل حاصل می‌شود در زمینهٔ سنجش میزان تسلط دانش‌آموزان به هر یک از خصایص زیربنایی برای حل مسئله یا پاسخ‌گویی به سؤال است. سنجش شناختی-تشخیصی به‌طور بالقوه برای بهبود تشخیص و استنباط در زمینهٔ نمرهٔ آزمون به‌کار می‌آید و در نتیجه به‌دلیل مرتبط کردن قوت‌ها و ضعف‌های آزمودنی‌ها در پاسخ به سؤال‌ها، ارتباط آموزش با سنجش را بهتر برقرار می‌کند (گیرل، آلوز و ماجران[2]، ۲۰۱۰).

بررسی منابع و اسناد موجود در سنجش شناختی-تشخیصی دو کارکرد متفاوت آن را نشان می‌دهد. اولین و رایج‌ترین کارکرد آن است که سنجش شناختی-تشخیصی ، فضای توانایی مکنونی متشکل از دانش، مهارت‌ها یا خصیصه‌ها را در نظر گرفته، عملکرد دانش‌آموز را در فضای چندبُعدی قرار داده و احتمال رسیدن به تسلط در هر مهارت را به‌صورت نیمرخ‌های مهارتی نشان می‌دهد (تاتسوکا، کورتر[3] و تاتسوکا، ۲۰۰۴؛ کیم[4]، ۲۰۱۱؛ ون‌داویر[5]، ۲۰۰۹). به عبارت دیگر، از طریق تجزیهٔ نمرهٔ کل فرد به مجموعه‌ای از نمرات مهارت‌ها یا خصیصه‌ها (یا به عبارت بهتر احتمال پاسخ‌گویی فرد به مجموعه‌ای از ابعاد) نمرات متعددی تولید شده و اطلاعات تشخیصی بیشتری فراهم می‌گردد. نمونه چنین دیدگاهی در پژوهش چن و همکارانش (۲۰۰۸، ب) وجود دارد که نشان داده شد دانش‌آموزان تایوانی از ۲۳ خصیصه ریاضی تنها در ۵ خصیصه به حد تسلط نرسیده بودند. کارکرد دوم به توصیف و مدل‌سازی عملکرد فرد در برخورد با مسئله مربوط می‌شود، بدین‌صورت که با توصیف ساده شده‌ای از حل مسئله در افراد، بررسی می‌کند که دانش و مهارت‌های آنان چگونه به حل مسئله منجر شده و با انجام مدل‌سازی عملکرد افراد را تبیین و پیش‌بینی می‌کند (گیرل، سویی و ژو[6]، ۲۰۰۹).

برای آگاهی از مدل‌های شناختی-تشخیصی باید در نظر داشت که به‌طور کلی، مدل‌های آماری سعی در بیان ریاضی رابطهٔ بین داده‌های گردآوری شده دارند. در ارتباط با سنجش شناختی تشخیصی، هدف مدل‌های تشخیصی نشان دادن عملکرد آزمودنی در سؤال بر مبنای خصیصه‌های لازم برای پاسخ‌گویی به سؤال و شایستگی آن آزمودنی در این خصیصه‌هاست (دی‌بلو، روسوس و استات[7]، ۲۰۰۷؛

1. Roussos, Templin & Henson
2. Gierl, Alves & Majeau
3. Corter
4. Kim
5. Von Davier
6. Cui & Zhou
7. DiBello, Roussos & Stout

روسوس، تمپلین و هنسون، ۲۰۰۷). لیتون و گیرل (۲۰۰۷: ۶) نشان دادند که مدل شناختی تشخیصی «توصیف ساده شده‌ای از حل مسئله افراد در تکالیف آموزشیِ استاندارد است. این مدل دانش و مهارت‌های کسب شدهٔ دانش‌آموزان در سطوح مختلف یادگیری را بازنمایی کرده و توصیف و پیش‌بینی عملکرد دانش‌آموزان را تسهیل می‌کند». با توجه به اینکه فرایند تفکر و حل مسئله به‌طور مستقیم مشاهده نمی‌شود، از این مدل‌ها زمانی استفاده می‌شود که پژوهشگران بخواهند به درک بهتری از دانش و مهارت‌های مورد استفادهٔ آزمودنی‌ها در حل‌مسئله دست یابند. از نگاهی دیگر می‌توان هدف مدل‌های تشخیصی را تعیین نیمرخ‌های خصیصه‌ای دانست، به‌طوری‌که طبقه‌بندی‌های چندگانه براساس الگوهای پاسخیِ مشاهده‌شده با توجه به خصیصه‌های لازم برای پاسخ‌گویی انجام می‌شود (فینکلمن، کیم، روسوس و ورشور[1]، ۲۰۱۰؛ ون‌داویر، ۲۰۰۸). بنابراین، ایدهٔ کلی مدل‌ها، تبدیل نظریه‌های شناختی به مدل‌های احتمالی است (سینهارا و المونڈ[2]، ۲۰۰۷). استفاده از مدل‌های شناختی در آزمون‌های تشخیصیِ پیشرفت تحصیلی زمانی توجیه می‌شود که این مدل‌ها بتوانند درک آزمودنی‌ها را تفسیر کنند.

ساده‌ترین مدل‌های تشخیصی، مدل‌های مرتبط با یک متغیر مکنونِ تکی هستند و به‌عنوان شاخصی برای اندازه‌گیری سازهٔ غالب به‌کار می‌آیند. این مدل‌ها، مدل‌های تک‌بعدی[3] نامیده شده و در چارچوب نظریهٔ سؤال پاسخ به‌وفور استفاده می‌شوند. به‌صورتی‌که عملکرد دانش‌آموزان در بُعد شایستگی به‌طور جداگانه گزارش داده می‌شود. علاوه‌براین، هنگامی که ابعاد با هم‌دیگر به‌طور توأمان برآورد می‌شوند، مدل‌های چندبُعدی[4] به‌کار می‌آیند. این مدل‌ها همبستگی‌های متغیرهای مکنون را به‌طور مستقیم برآورد کرده و درجهٔ وابستگی سازه‌ها به هم‌دیگر و اطلاعاتی در مورد ابعاد چندگانه در هنگام برآورد پارامترهای سؤال یا نیمرخ‌های خصیصه را مشخص می‌کنند. به‌دلیل وابستگی مدل‌های شناختی-تشخیصی در طبقه‌بندی آزمون‌ها به متغیرهای مکنونِ چندگانه، این مدل‌ها جزو مدل‌های چندبُعدی محسوب می‌شوند.

برای آگاهی از چگونگی عمل مدل‌های شناختی تشخیصی، اطلاع از مفهوم ماتریس Q لازم است. به منظور برآورد تابع‌های پاسخ سؤال باید فهرستی از مهارت‌های اندازه‌گیری‌شده توسط ابزار و توصیفی از چگونگی اندازه‌گیری این مهارت‌ها توسط هر یک از سؤالات وجود داشته باشد (روسوس، تلمپین و هنسون، ۲۰۰۷). نتیجه این فعالیت تشکیل ماتریس وقوع[5] یا ماتریس Q است. در حقیقت ماتریس Q ماتریس سؤال × خصیصه است که تعیین می‌نماید برای پاسخ‌گویی به هر سؤال چه خصیصه‌هایی نیاز

1. Finkelman, Kim, Roussos & Verschoor
2. Sinharay & Almond
3. unidimensional models
4. multidimensional models
5. incidence

است. معمولاً درایه‌های ماتریس به‌صورت صفر و یک بیان می‌شوند که اعداد یک به منزلهٔ لازم بودن خصیصه در پاسخ‌گویی به آن سؤال است.

با توجه به اهداف و چشم‌اندازهای سنجش شناختی-تشخیصی و هم‌چنین ویژگی‌های خاص مدل‌سازی در آن، رویه‌های مرسوم در مدل‌سازی نمی‌توانند پاسخ‌گوی نیازها در سنجش شناختی-تشخیصی باشند. به‌عبارت‌دیگر، مدل‌سازی در سنجش شناختی- تشخیصی با دشواری‌ها و چالش‌های خاصی روبه‌رو است. در این مقاله سعی شده است مهم‌ترین چالش‌های مدل‌سازی شناختی-تشخیصی تشریح گردد تا پژوهشگران دیدگاه عمیق‌تری نسبت به مدل‌سازی شناختی-تشخیصی و دشواری‌های آن پیدا کنند. این چالش‌ها شامل تک‌بُعدی بودن در مقابل چندبُعدی بودن، تعداد خصیصه‌ها، همبستگی بین خصیصه‌ها، تعداد مناسب سؤال در هر خصیصه، درجهٔ دقت خصیصه‌ها، اعتبار خصیصه‌ها، روایی سنجش شناختی تشخیصی، پارامترهای سؤال، برازش مدل، شناسایی و تعیّن مدل، هم‌گرایی، و نمونه‌گیری‌های پیچیده بودند. بدین منظور، تجربهٔ به‌کارگیری داده‌های مطالعهٔ تیمز در مدل‌سازی شناختی-تشخیصی مطرح شده و چالش‌های مدل‌سازی بحث می‌شود. در این مباحث چگونگی رفع چالش‌ها نیز به بحث گذاشته شده است. آگاهی از این مباحث باعث می‌شود تا پژوهشگرانی که در حوزه سنجش شناختی-تشخیصی مشغول به تحقیق هستند نسبت به حساسیت و مشکلات موجود در سنجش شناختی-تشخیصی وقوف بیشتری پیدا کرده و در نهایت مدل‌های بهتری را ارائه نمایند.

۲. روش پژوهش

روش به‌کار گرفته‌شده در این پژوهش مطالعه موردی[1] در زمینهٔ مشکلات در هنگام مدل‌سازی داده‌های علوم مطالعهٔ تیمز با رویکرد سنجش شناختی-تشخیصی است. از آن‌جاکه این تجربه می‌تواند مورد استفادهٔ سایر پژوهشگران شناختی-تشخیصی قرار گیرد، آگاهی از آن بسیار سودمند خواهد بود.

۳. یافته‌های پژوهش

بسیاری از مشکلات موجود در فرایند مدل‌سازی‌های رایج، در سنجش شناختی-تشخیصی نیز وجود دارند. باوجوداین، به علت پیچیدگی ساختار، بارهای عاملی و تعامل بین خصیصه‌ها چالش‌های بیشتری رویاروی مدل‌سازی شناختی-تشخیصی قرار دارد. به موارد فوق کمبود نرم‌افزارهای تحلیلی و استفاده از روش‌های برآوردِ زمان‌بر نیز اضافه می‌شود که دشواری‌های موردبحث را دوچندان می‌کند. در این بخش مهم‌ترین چالش‌های مدل‌سازی شناختی-تشخیصی اشاره می‌شوند.

1. Case Study

تک‌بُعدی بودن در مقابل چندبُعدی بودن: چالش بسیار مهم در مدل‌سازی شناختی-تشخیصی داده‌های مطالعهٔ تیمز آن است که اساساً این مطالعه براساس منطق تک‌بُعدی بودن طراحی شده است. حال آنکه هدف سنجش شناختی-تشخیصی، استخراج چندین بُعد برای بیان توصیفات تشخیصی است. مطالعهٔ تیمز همانند عموم مقیاس‌سازی‌ها، تک‌بُعدی است. بدین‌صورت که محوری در نظر گرفته شده و افراد در سازهٔ تعریف‌شده جایی روی این محور، در بین ابتدا تا انتهای آن، پیدا می‌کنند. ساده‌ترین مثال، نمرهٔ کلاسی است که دانش‌آموزان در متغیری، مثل عملکرد در درس فیزیک، نمره‌ای در طیف صفر تا بیست به‌دست می‌آورند. باوجوداین، جهت‌گیری اصلیِ سنجش شناختی تشخیصی، پیدا کردن چندین بُعد (k خصیصه) برای ارائهٔ عملکرد آزمودنی‌ها در هر یک از ابعاد است. این موضوع، عمده‌ترین مزیت کاربرد سنجش‌های شناختی-تشخیصی نسبت به سایر مدل‌ها به‌شمار می‌رود. بنابراین، چالش سنجش شناختی-تشخیصی آن است که آیا امکان استخراج چندین بُعد وجود دارد که بتواند شناخت عمیق‌تری در مورد عملکرد آزمودنی‌ها را حاصل کند؟ این چالش با رویکرد بازبرازش[1]، که مدل‌سازی داده‌های مطالعهٔ تیمز نیز جزو آن است، نشان داده می‌شود (یاماگوشی و اوکادا[2]، ۲۰۱۸). آزمون طراحی‌شده و برازش‌یافته با مدل تک‌بُعدی با روش‌های پیشرفته‌ای تحلیل می‌گردند که در طراحیِ اولیه مورد توجه نبودند. در چالش بُعدیت این امکان وجود دارد که مدل تک‌بُعدی با مدل‌های k بُعدی نیز برازش قابل‌قبولی داشته باشد. این امر فرایند مدل‌سازی را با پیچیدگی‌های زیادی روبه‌رو می‌کند. برخی از مسائل در قسمت برازش مدل توضیح داده می‌شوند.

انتخاب بین مدل تک‌بُعدی یا چندبُعدی به موارد متعددی ربط پیدا می‌کند که هدف سنجش از جملهٔ آن‌هاست. اگر هدف سنجش مرتب کردن آزمودنی‌ها در یک مقیاس شایستگی تک‌بُعدی است، مدل‌های تک‌بُعدی مناسب هستند. ولی اگر هدف، طبقه‌بندی گسستهٔ افراد در فضاهای مهارتیِ چندگانه باشد، کاربرد مدل‌های چندبُعدیِ سنجش شناختی مرجح است (دی‌بلو، روسوس و استات، ۲۰۰۷).

به‌منظور حل چالش بُعدیت، در مدل‌سازی شناختی-تشخیصی داده‌های تیمز، براساس توصیه‌ها ابتدا مدل تک‌بُعدی تحلیل شد و سپس با افزودن ابعاد بیشتر مدل پیچیده‌تری به‌دست آمد. این وضعیت تا جایی ادامه پیدا کرد که به استناد آماره‌های برازش، مدل با ابعاد بیشتر نسبت به مدل قبلی مناسب‌تر شناخته نشد. در این حالت براساس اصل ایجازِ[3] مدل‌سازی، مدل با ابعاد کم‌تر با آماره‌های برازش مشابه نسبت به مدل پیچیده‌تر ترجیح داده می‌شود (دلاتوره و مینشن[4]، ۲۰۱۹).

1. Retrofitting
2. Yamaguchi & Okada
3. Parsimony
4. de la Torre & Minchen

تعداد خصیصه‌ها: موضوع دیگر در مدل‌سازی شناختی-تشخیصی داده‌های مطالعهٔ تیمز، انتخاب تعداد خصیصه‌های لازم است. هر چند بیان شد که در این مدل‌سازی نسبت به استخراج ابعاد بیشتر امیدواری وجود دارد، ولی سؤال اینجاست که تعداد مناسب خصیصه‌ها چه تعداد است؟ اگر تعداد زیادی از خصیصه‌ها را برای مدل تشخیصی در نظر بگیریم، این احتمال وجود دارد که خصیصه‌ها همبستگی بسیار بالایی با هم‌دیگر پیدا کنند. در نتیجه خصیصه‌ها به اندازه‌ای از هم متمایز نباشند که بتوان آن‌ها را از هم‌دیگر جدا پنداشت. از این‌رو، به‌کارگیری تعداد زیاد خصیصه‌ها چالشی ناظر به جداسازی تجربی خصیصه‌ها به‌همراه دارد (راپ[1] و تمپلین، ۲۰۰۸). علاوه‌براین، مدل‌ها با تعداد زیاد خصیصه در معرض خطر عدم شناسا بودن[2] قرار دارند. در این حالت امکان برآورد دقیق برای مدل‌هایی که تعداد زیادی از مهارت‌ها را در هر سؤال مشارکت می‌دهند، وجود ندارد (دی‌بلو، روسوس و استات، ۲۰۰۷). علاوه‌براین، حجم نمونه نیز با تعداد خصیصه‌ها در ارتباط است؛ به‌طوری‌که توصیه شده است در صورت دسترسی به تعداد نمونهٔ کم، خصیصه‌های محدودی برای مدل‌سازی در نظر گرفته شوند (اسکیجز، ویلکینز و هین[3]، ۲۰۱۶)

با توجه به نکات فوق، بسیاری از پژوهشگران انتخاب مدل‌ها با تعداد خصیصه‌های کم را پیشنهاد داده‌اند. زیرا هم مدل‌سازی را ساده‌تر می‌کند (چیو و سئو[4]، ۲۰۰۹) و هم پارامترهای کمی برآورد می‌گردند که باعث برآوردهای بهتری می‌شود (روسوس، تمپلین و هنسون، ۲۰۰۷). در کل با توجه به اصل ایجازِ مدل‌سازی (به معنای افزایش پیچیدگی مدل بنا به ضرورت)، مدل با تعداد کمی از خصیصه‌ها پیشنهاد شده است. باوجوداین، مدل‌های با تعداد کم خصیصه‌ها نیز مشکلاتی را به‌بار می‌آورند. کاهش تعداد خصیصه‌ها ممکن است باعث افزایش سؤالات در هر خصیصه شده و در نتیجه افزایش اعتبار را به همراه داشته باشد، ولی احتمال دارد این کار در نظریهٔ زیربناییِ مدل قابل‌تأیید نباشد. به این صورت که خصیصه‌ها کاربرد[5] یا تفسیرپذیری[6] خود را از دست بدهند (دی‌بلو، روسوس و استات، ۲۰۰۷). بنابراین، ایجاد توازن بین تعداد خصیصه‌ها، نه خیلی زیاد که از نظر آماری قابل برآورد نباشند و نه خیلی کم که قابلیت تفسیر تشخیصی خود را از دست بدهند، ضروری است. به‌عنوان راهنمای عملی، استفادهٔ ۴ تا ۸ خصیصه برای مدل‌سازی شناختی-تشخیصی توصیه شده است (راپ، تمپلین و هنسون، ۲۰۱۰). اسکیجز، ویلکینز و هین (۲۰۱۶) در مطالعهٔ خود نتیجه گرفتند که وجود بیش از ۸ خصیصه، یا پارامترهای سؤال و خصیصه را بی‌ثبات می‌کند یا مانع از رسیدن به هم‌گرایی در مدل می‌گردد.

1. Rupp
2. Nonidentifiability
3. Skaggs, Wilkins, & Hein
4. Chiu & Seo
5. Utility
6. Interpretability

به‌منظور رسیدن به تعداد منطقی خصیصه‌ها، جانگ (۲۰۰۵، به نقل از دی‌بلو، روسوس و استات، ۲۰۰۷) از روش ترکیب خصیصه‌های قابل تلفیق و یا حذف برخی خصیصه‌ها به‌منظور کاهش تعداد آن‌ها استفاده کرد. در پژوهش فوق، ابتدا ۳۲ فرایند یا خصوصیت قابل استفاده به‌عنوان مهارت یا خصیصه تعیین شد. سپس به علت عدم‌امکان برآورد، این تعداد خصیصه‌ها به ۱۶ و در نهایت به ۹ مورد کاهش پیدا کرد. کاهش خصیصه‌ها تا جایی انجام شد که او به نتیجه رسید که مهارت‌ها آن چنان از هم متمایز شده بودند که قابل‌ترکیب نبودند و اگر برخی از خصیصه‌ها حذف می‌شدند، تعدادی از سؤالات بدون خصیصه باقی می‌ماندند. روش دیگری برای رسیدن به تعداد منطقی خصیصه‌ها، استفاده از رویکردی است که مدل‌سازی را از مدل‌های با ابعاد کم (از تک‌بُعدی به ابعاد بیشتر) شروع کنیم تا جایی که بر اساس ملاک‌های مربوط به برازش مدل، مقادیر پارامترها، منطقی بودن مدل و ... مشخص شود که افزودن ابعاد بیشتر به مدل پیچیدگیِ غیرضروری را به آن تحمیل می‌کند (ون‌داویر، ۲۰۰۸). این کار در فرایند مدل‌سازی داده‌های مطالعهٔ تیمز در پیش گرفته شد؛ به‌طوری‌که مدل ۸ خصیصه‌ای به‌دلیل برتری نسبت به مدل‌های با تعداد خصیصه‌های کم‌تر انتخاب شد (کبیری و همکاران، ۲۰۱۷).

در تجربه مدل‌سازی شناختی داده‌های مطالعهٔ تیمز ابتدا براساس مرور منظمِ منابع آموزش علوم، ۲۲ شایستگی آموزش علوم در پایه هشتم تعیین شد که ۱۹ خصیصهٔ آن قابل بررسی در مطالعهٔ تیمز بود. در جریان مدل‌سازی تنها هفت خصیصه به‌عنوان مهارت‌های لازم در مدل قرار گرفت. با بررسی خصیصه‌های باقیمانده، چندین خصیصه شامل طراحی تحقیق، فرضیه‌سازی، دانش کاربرد ابزار علمی، ارزشیابی شواهد و نتیجه‌گیری با هم‌دیگر تلفیق‌شده و با عنوان کاوشگری علمی به مدل افزوده شد. از بین بقیهٔ خصیصه‌ها نیز بازشناسیِ نمادهای علمی و دانش اصطلاحات علمی با دانش پایه و مقایسه و طبقه‌بندی با استدلال علمی ترکیب شدند. بدین‌ترتیب، عموم خصیصه‌ها به‌نحوی در جریان مدل‌سازی شناختی-تشخیصی قرار گرفتند.

همبستگی بین خصیصه‌ها: موضوع دیگرِ مرتبط با بحث بالا همبستگی بین خصیصه‌هاست. اگر خصیصه‌ها همبستگی کاملی داشته باشند، ارائه نتایج در مورد خصیصه‌ها اطلاعات واقعی در مورد نمرهٔ هر یک از خصیصه‌ها را فراهم نمی‌کند و نمی‌توان بین نیمرخ خصیصه‌ها تمایز معتبری قائل شد (هابرمن[1] و ون‌داویر، ۲۰۰۷). برای درک این موضوع در نظر بگیرید که اگر همبستگی بین خصیصه‌ها کامل و یا بسیار بالا باشند، بیشتر الگوهای خصیصه‌ای در صورت عدم تسلط بر خصیصه‌ها به شکل (۰،۰،...،۰) و یا در صورت تسلط بر خصیصه‌ها به شکل (۱،۱،...،۱) می‌شود و سایر الگوهای خصیصه‌ای (ترکیبی از صفرها و یک‌ها) کم‌تر دیده خواهند شد (ون‌داویر، ۲۰۰۷). بنابراین، تشخیص بین خصیصه‌ها به مفهومی که هدف سنجش شناختی-تشخیصی باشد روی نخواهد داد. در این حالت، ویژگی چندبُعدی بودن سنجش شناختی زیر سؤال خواهد رفت. در نتیجه، وجود همبستگی بالا بین

1. Haberman

خصیصه‌ها در سنجش شناختی-تشخیصی مطلوب نیست بلکه بایستی بین خصیصه‌ها تمایز مناسبی وجود داشته باشد. تعداد زیاد خصیصه‌ها از عواملی است که احتمالاً درجه همبستگی بین آن‌ها را بالا می‌برد (راپ، تمپلین و هنسون، ۲۰۱۰). هم‌چنین برای مدل‌سازی همبستگی بین خصیصه‌ها استفاده از مدل‌هایی که رابطه بین خصیصه‌ها را به‌صورت سلسله‌مراتبی و روابط سطح بالاتر در نظر می‌گیرند، پیشنهاد شده است (دلاتوره و داگلاس[1]، ۲۰۰۴).

در تحلیل داده‌های مطالعهٔ تیمز مشخص شد که همبستگی بین خصیصه‌ها به‌طور قابل توجهی کم بود. این مسئله تا اندازه‌ای از این موضوع نشأت می‌گرفت که بُعد اول (دانش پایه) مقدار بسیار زیادی از واریانس سؤالات را به خود اختصاص می‌داد. از طرف دیگر، کم بودن همبستگی بین خصیصه‌ها نشان داد که خصیصه‌ها به‌خوبی از یکدیگر متمایز شده‌اند و لذا وابستگی بسیار کمی به یکدیگر دارند.

تعداد مناسب سؤال در هر خصیصه: از جمله پرسش‌ها در مدل‌سازی داده‌های مطالعهٔ تیمز این موضوع بود که هر خصیصه باید از چند سؤال تشکیل شود. این موضوع تا اندازه‌ای به اعتبار نیز مربوط می‌شود. باوجوداین، به لحاظ این‌که اعتبار زیرمقیاس‌های سنجش شناختی-تشخیصی قاعدتاً از اعتبار نمرهٔ کل کم‌تر است، باید به‌طور خاص بررسی شود. هابرمن و ون‌داویر (۲۰۰۷) براساس کاربردهای سنجش شناختی-تشخیصی توصیه کردند که حداقل سه یا چهار سؤال باید در هر خصیصه درنظر گرفته شوند. چن و همکاران (۲۰۰۸ب) در مطالعه خود خصیصه‌ها را به‌صورتی عملیاتی کردند که حداقل سه سؤال را در برگرفته باشند. بااین‌حال، یکی از محدودیت‌های مدل‌سازی داده‌های مطالعات کلان مقیاس هم‌چون مطالعهٔ تیمز، این است که تعداد سؤالات در خصیصه‌ها نامتوازن است (اسکیجز، ویلکینز و هین، ۲۰۱۶).

با وجود آنکه ملاک‌های ساده‌تری در مورد تعداد سؤالات در هر خصیصه وجود دارد ولی به لحاظ استفاده از نمونه‌گیری ماتریسی[2] در مطالعهٔ تیمز و هم‌چنین بیشتر شدن اعتبار خصیصه‌ها، تعداد سؤالات بیشتری در هر خصیصه در نظر گرفته شد. علاوه‌براین، به‌عنوان قانون سرانگشتی وجود هشت سؤال برای هر خصیصه لازم شمرده شده است. بنابراین، ملاک هشت سؤال در مدل‌سازی داده‌های مطالعه تیمز مورد توجه قرار گرفت.

درجهٔ دقت خصیصه‌ها: در مدل‌سازی شناختی-تشخیصی تأکیدات روی استفاده از خصیصه‌ها در جزئی‌ترین سطح مهارت است. باوجوداین، مدل‌سازی داده‌های مطالعهٔ تیمز در درس علوم به علت عدم وجود سابقهٔ پژوهشی در سطح جزئی خصیصه‌ها قابل انجام نبود. علاوه‌براین، اگر مهارت‌ها در سطح بسیار جزئی استفاده می‌شد، تعداد خصیصه‌ها به اندازه‌ای زیاد می‌شد که سؤالات کمی در هر

1. Douglas
2. Matrix Sampling

خصیصه قرار می‌گرفتند. بنابراین، برای تعریف خصیصه‌ها آگاهی از مفهوم درجهٔ دقت[1] خصیصه‌ها لازم است. درجهٔ دقت، عمق و گسترهٔ دانش و مهارت‌های اندازه‌گیری را نشان می‌دهد (لیتون و گیرل، ۲۰۱۱) و به دو حالتِ درشت[2] و ریز[3] تعریف می‌شوند. در درجهٔ دقت درشت، خصیصه‌ها در سطح گسترده‌تری تعریف می‌شوند. در مقابل، درجهٔ دقت ریز به خصیصه‌های تعریف‌شده در سطح جزئی و بسیار محدود اشاره دارد که مطلوب پژوهشگران سنجش شناختی-تشخیصی برای ارائه تفاسیر شناختی ویژه است. هر چند توصیه شده است که در سنجش تشخیصی، خصیصه‌ها به سمت درجهٔ دقت‌های ریز تمایل پیدا کنند (اسکیجز، ویلکینز و هین، ۲۰۱۶؛ لیتون و گیرل، ۲۰۱۱)، ولی انتخاب درجهٔ دقت خصیصه‌ها تابعی از هدف سنجش، پیچیدگی شناختی و جامعهٔ مورد بررسی است (راپ، تمپلین و هنسون، ۲۰۱۰).

در مدل‌سازی شناختی-تشخیصی درس علوم تیمز به علت عدم وجود سابقهٔ تحلیل شناختی-تشخیصی در آموزش علوم و عدم وجود خصیصه‌هایی که با درجهٔ دقت ریز تعریف شده باشند، خصیصه‌های با درجهٔ دقت درشت به‌کار رفت. نمونه چنین خصیصه‌هایی شامل کاوشگری علمی و پیش‌بینی بودند.

اعتبار[4] خصیصه‌ها: انتظار می‌رود که خصیصه‌های مورد بررسی دارای اعتبار قابل قبولی باشند. بااین‌حال، مفاهیم استاندارد اعتبار به‌طور مستقیم قابل‌انتقال به سنجش شناختی-تشخیصی نیست. زیرا خصیصه‌ها بر اساس تعداد کمی از سؤالات قرار دارند، در نتیجه اعتبار کم‌تری را نسبت به نمرهٔ کل نشان می‌دهد. علاوه‌براین، به علت آنکه مقیاس‌های در سنجش شناختی-تشخیصیِ طبقه‌ای هستند، بررسی اعتبار با دشواری‌های بیشتری دست به گریبان است. برای درک مشکل، اگر اعتبار را به صورت نسبت واریانس نمرهٔ واقعی به واریانس نمرهٔ مشاهده‌شده در نظر بگیریم، در سؤالاتی که در بیش از یک خصیصه مشارکت دارند، هر خصیصه تنها در بخشی از واریانس سطح سؤال مشارکت خواهد داشت (گیرل، سویی و ژائو، ۲۰۰۹). بنابراین، رویکردهای مطرح در نظریهٔ کلاسیک یا نظریهٔ سؤال پاسخ در سنجش شناختی-تشخیصی قابل استفاده نخواهند بود. برای بهبود بخشیدن اعتبار زیرمقیاس‌ها سه روش تجربی بیزی، رویکرد رگرسیون محور و نمرات وابسته به مدل‌های چندبُعدی پیشنهاد شده‌اند که همگی حول این ایده می‌گردند که تلفیق داده‌های متوازن در پاسخ آزمودنی‌ها خطای نمرات زیرمقیاس‌ها را کاهش می‌دهد (استون، یی، ژو و لین[5]، ۲۰۱۰).

برای رفع مشکل فوق راه حل‌هایی پیشنهاد شده است. گیرل، سویی و ژائو (۲۰۰۹) با تعریف دو احتمال تسلط بر خصیصه (آزمودنی که بر خصیصه تسلط دارد می‌تواند به سؤال پاسخ درست دهد) و

1. Grain Size
2. Coarse
3. Fine
4. Reliability
5. Stone, Ye, Zhu & Lane

احتمال عدم تسلط بر خصیصه (آزمودنی که بر خصیصه تسلط ندارد می‌تواند به سؤال پاسخ درست دهد) فرمول‌های آلفای کرونباخ متناظر با سنجش شناختی-تشخیصی ارائه کردند. البته فرمول‌های آنان قابل کاربرد در مدل سلسله‌مراتبی خصیصه[1] بود. با وجود این‌گونه تلاش‌ها، به‌نظر می‌رسد که مطالعات بیشتر در مورد تلفیق اعتبار با سنجش شناختی-تشخیصی نیاز است.

با توجه به اینکه در تحلیل داده‌های مطالعهٔ تیمز از مدل تشخیصی کلی[2] استفاده شد و در خروجی نرم‌افزار این مدل اعتبار خصیصه‌هاگزارش می‌گردد لذا مشکلی در زمینهٔ محاسبه وجود نداشت. باوجوداین، گزارش اعتبار خصیصه‌ها همچنان به‌عنوان یکی از مشکلات سایر مدل‌ها وجود دارد.

روایی سنجش شناختی تشخیصی: به‌طورکلی هر گاه صحبت از اعتبار به میان می‌آید، موضوع روایی نیز به ذهن متبادر می‌گردد. سنجش شناختی-تشخیصی نیز این قاعده مستثنی نیست. سنجش شناختی-تشخیصی ابزار قدرتمندی برای بررسی‌های زیربنایی دربارهٔ روایی سازه است. به‌طوری که می‌تواند برای تائید کردن ارتباط بین پاسخ‌های سؤال و سازهٔ فرضی استفاده شود (چن و همکاران، ۲۰۰۸، الف). برای کمّی کردن روایی درونی دو آماره *IMStats* و *EMStats* پیشنهاد شده است. برای محاسبهٔ این آماره‌ها میانگین نمرهٔ افراد به حد تسلط رسیده در سؤال با نمرهٔ افراد به حد تسلط نرسیده مقایسه می‌گردد. اگر دو نسبت با هم نزدیک باشند، طبقه‌بندی خصیصه مناسب نبوده و باید مورد تجدیدنظر قرار گیرد (دی‌بلو، روسوس و استات، ۲۰۰۷؛ روسوس، تمپلین و هنسون، ۲۰۰۷).

با این‌حال، آماره‌های ذکرشده در بالا مربوط به برخی از مدل‌های شناختی-تشخیصی همچون مدل یک‌شکل بازپارامتری شده[3] است. نظیر این آماره‌ها در مدل‌های تشخیصی کلی وجود ندارد، لذا گزارشی در این زمینه در مورد روایی مدل ارائه نشده است. هر چند که به‌نظر متخصصان از دیگر آماره‌ها و پارامترهای مدل می‌توان در مورد روایی مدل استنباط کرد، ولی به‌نظر می‌رسد که توسعهٔ روش‌شناسی مدل تشخیصی کلی برای شامل کردن آماره‌های فوق و یا پیشنهاد آماره‌های مشابه لازم باشد.

پارامترهای سؤال: مناسب بودن پارامترهای سؤال از جمله مواردی بود که در مدل‌سازی شناختی-تشخیصی داده‌های مطالعهٔ تیمز توجه شد. این چالش به قابل قبول بودن پارامترهای سؤال اعم از پارامترهای شیب، موقعیت، همبستگی سؤال با خصیصه و خطاهای مرتبط اشاره دارد. قابل قبول بودن برآوردهای پارامترهای سؤال باید از نظر تجانس درونی، منطقی بودن[4] و توافق با انتظارات بعدی ارزیابی شوند (دی‌بلو، روسوس و استات، ۲۰۰۷). به‌عنوان مثال، باید بررسی گردد که تقسیم‌بندی آزمودنی‌ها در هر خصیصه به‌صورت مسلط‌ها یا غیر مسلط‌ها تا چه اندازه با انتظارات نظری همگونی دارد. اگر خصیصه‌ای نسبت به انتظار خیلی ساده یا خیلی دشوار باشد، ماتریس Q باید مورد تجدیدنظر

1. Attribute hierarchy method (AHM)
2. General Diagnostic Model (GDM)
3. Reparameterized unified model (Full- UM)
4. Reasonability

قرار گیرد (روسوس، تمپلین و هنسون، ۲۰۰۷). برخی از کاربرد سنجش شناختی-تشخیصی نشان داده‌اند که حدود ۹ درصد از درایه‌های ماتریس Q به علت عدم وجود قدرت تمیز مناسب در برخی سؤالات حذف شدند (جانگ[1]، ۲۰۰۶).

با وجود اهمیتِ بررسی پارامترهای سؤال، برخی از پژوهشگران به نامناسب بودن آماره‌های نظریهٔ کلاسیک و نظریهٔ سؤال پاسخ اشاره کرده و استفاده از آماره‌های خاصی همچون شاخص شناختی-تشخیصی را پیشنهاد می‌نمایند. بااین‌حال، ایراداتی به این شاخص از جمله عدم مشخص شدن توان تشخیص سؤال برای خصیصه‌ای خاصی وارد است (هنسون، روسوس، داگلاس و هه[2]، ۲۰۰۸). جدای از شاخص شناختی-تشخیصی توصیه‌هایی در مورد میزان قابل قبولِ خطای برآورد مطرح شده است که مقدار آن نباید از ۰/۰۵ بیشتر باشد (روسوس، تمپلین و هنسون، ۲۰۰۷).

در مدل‌سازی شناختی-تشخیصی داده‌های مطالعهٔ تیمز، در سؤالاتی که مقادیر نامناسبی از پارامترها داشتند ماتریس Q تجدیدنظر شد. این تغییرات تا جایی انجام شد که پارامترهای سؤالات، مقادیر قابل قبولی پیدا کردند. در مورد خطای برآورد نیز به علت استفاده مدل تشخیصی کلی از روش خطای استاندارد جک‌نایف[3] انتظار میزان بیشتری از خطای استاندارد نسبت به مقدار بیان شده توسط روسوس، تمپلین و هنسون (۲۰۰۷) وجود داشت.

برازش مدل: چنان‌چه در بحث چندبُعدی بودن مدل گفتیم، در مدل‌سازی شناختی-تشخیصی مدل‌های بسیار پیچیده موردنظر نیستند. اما باید توجه داشت که انتخاب ساده‌ترین مدل نیز مطلوب سنجش تشخیصی نیست بلکه به ساده بودن مدل در کنار اهداف تشخیصی توجه می‌شود. در این‌جا، موضوع برازش مدل مطرح می‌شود. به‌طورکلی در مدل‌سازی شناختی ساده‌ترین مدلی مطلوب است که اهداف تشخیصی را برآورده کرده و به‌طور منطقی و به‌خوبی برازش پیدا کند. با این تفاسیر امکان دارد که مدل ساده‌تری با مقادیر برازش نامناسب‌تر به علت منطق زیربناییِ تأییدکننده انتخاب شود (دی‌بلو، روسوس و استات، ۲۰۰۷).

برای محاسبه برازش همان منطق مدل‌های نظریهٔ سؤال پاسخ به‌کار گرفته می‌شود. در نظریهٔ سؤال پاسخ نموداری ترسیم می‌شود که نسبت نمرات مشاهده‌شدهٔ صحیح با نمرات پیش‌بینی‌شده مقایسه شده و میزان تناسب در قالب آماره χ^2 به کمّیت درمی‌آید. در صورتی‌که سؤالات زیادی برازش نداشته باشند، نتیجه گرفته می‌شود مشکلی در برازش مدل با داده‌ها وجود دارد. همین ایده قابل گسترش به مدل‌های شناختی-تشخیصی است. در این مدل‌ها نیز دو نوع شاخص برازش کلی مدل و برازش سؤال محاسبه می‌شود (سینهارا و الموند، ۲۰۰۷). در برازش مدل علاوه بر χ^2 دو آمارهٔ ملاک اطلاعاتی

1. Jang
2. He
3. Jacknife Standard Error (JRR)

آکایک[1] و ملاک اطلاعات بیزی[2] نیز استفاده می‌شود. برای انتخاب بهترین مدل، مقادیر کمِ آماره‌های فوق به‌عنوان مدل مطلوب در نظر گرفته می‌شود. برازش سؤال معمولاً با همان آماره χ^2 بررسی می‌شود (راپ و تمپلین، ۲۰۰۸). علاوه‌براین، تلاش‌هایی در زمینه تعریف آمارۀ برازش فرد انجام شده است که از آن جمله می‌توان به شاخص تجانس سلسله‌مراتبی[3] برای مدل سلسله‌مراتبی خصیصه (لیتون، سویی و کور[4]، ۲۰۰۹) و آمارۀ برازش فرد برای مدل بازپارامتری‌شدۀ یک‌شکلِ کامل (هارتز، ۲۰۰۵، به نقل از راپ و تمپلین، ۲۰۰۸) اشاره کرد.

عدم برازش به دلایل متعددی رخ می‌دهد. عدم برازش می‌تواند ناشی از نامناسب بودن مدل انتخاب شده (استفاده از مدل جبرانی به جای غیرجبرانی یا برعکس)، اشکال در ساختار بار در ماتریس Q (حذف شدن برخی از خصیصه‌های لازم از ماتریس Q یا عدم‌انطباق داده‌ها با ماتریس Q)، وجود محدودیت‌های پارامتر غیرضروری (وجود محدودیت‌هایی مثل مساوی در نظر گرفتن برخی پارامترها) و یا جامعه‌هایی ناهمگون از آزمودنی‌ها (وجود ترکیبی از دو یا چند زیرجامعه با پارامترهای سؤال متفاوت) باشد (راپ و تمپلین، ۲۰۰۸). لذا در صورت عدم‌برازش مدل توجه به دلایل احتمالی فوق می‌تواند در جهت برازش مؤثر واقع شود.

در مدل‌سازی شناختی داده‌های مطالعۀ تیمز علاوه‌بر حصول برازش مناسب مدل، برازش مناسبی در هر یک از سؤالات نیز حاصل شد. به‌عبارت‌دیگر، برازش سؤالات به‌عنوان یکی از ملاک‌های تصمیم گیری برای کیفیت ماتریس Q در مورد هر سؤال به‌کار رفت. در مورد سؤالاتی که برازش نامناسبی داشتند، خصیصه‌های لازم برای حل سؤال تجدیدنظر شد تا برازش قابل قبول سؤال حاصل گردد.

شناسایی و تعیین مدل: در فرایند مدل‌سازی داده‌های مطالعۀ تیمز، به‌خصوص در مراحل اول مدل‌سازی، در مواقعی تحلیل بدون رسیدن به راه حلی به پایان می‌رسید. پس از بررسی اشکالات نتیجه گرفته شد که عدم‌تعیّن ماتریس Q ازجمله عوامل بسیار مهم در رخ دادن این وضعیت است. استفاده از مجموعه داده‌های بزرگ با تعداد زیادی از سؤالات، کار پیدا کردن مصادیق عدم تعیّن را دشوار می‌کند. لذا دو موضوع شناسایی[5] و تعیّن[6] مدل از جمله موارد تکنیکی و مهم مدل‌سازی به‌شمار می‌روند. شناسایی به معنای برآورد منحصربه‌فردِ مقادیر برای هر پارامتر است و ارتباط آن با مدل‌سازی در بخش تعداد خصیصه‌ها بیان شد. به‌طور خلاصه باید مدل به‌گونه‌ای طراحی شود که پارامترهای آن شناساشده باشند. تعیّن مدل به معنای این است که خصیصه‌هایی را که برای پاسخ‌گویی به سؤال تعریف کرده‌ایم با شرایط نظری و واقعی مطابقت داشته باشد. در مواقعی امکان دارد که فرد خصیصه‌های کم‌تری را نسبت به آن‌چه در ماتریس Q مشخص شده است به‌کار بگیرد یا از طرف

1. Akaike's Information Criterion (AIC)
2. Bayesian Information Criterion (BIC)
3. Hierarchy Consistency Index (HCI)
4. Cor
5. identifiability
6. specification

دیگر، خصیصه‌ای برای پاسخ‌گویی به سؤال نیاز باشد که در ماتریس Q وجود نداشته باشد. در این حالت مدل تعیّن نشده[1] می‌شود. بنابراین، کیفیت ماتریس Q برای مدل‌سازی شناختی اهمیت زیادی دارد (کیم، ۲۰۱۱). برآوردهای نامناسب، عدم برازش، نرسیدن به هم‌گرایی و ... پیامدهای مدل تعیّن نشده هستند. هم‌چنین گزارش داده شده است که وجود ۵ تا ۱۰ درصد درایه‌های دارای شرایط عدم تعیّن در ماتریس Q برآوردهای پارامتر خصیصه و سؤال را به‌طور جدی دچار مشکل می‌کنند (ایم و کورتر، ۲۰۱۱).

در تحلیل داده‌های مطالعهٔ تیمز نیز مصادیق بسیاری در زمینه عدم‌تعیّن ماتریس Q مشاهده شد. این مصادیق به عدم‌مشارکت خصیصه‌های لازم برای پاسخگویی به سؤال و یا تعیین مشارکت غیر لازم خصیصه‌ها برای پاسخگویی مربوط می‌شد. در صورت بروز چنین مواردی پارامترهای نامناسبی به‌دست می‌آمد. جهت شناسایی و رفع مصادیق عدم‌تعیّن، تغییرات ماتریس Q در هر یک از سؤالات به‌طور جداگانه انجام و سپس تحلیل مجدداً اجرا می‌شد. اجرای تحلیل مدل پس از تغییرات هر سؤال و به‌طور جداگانه باعث می‌شود که وجود عدم‌تعیّن در هر یک از سطور ماتریس Q تشخیص داده شود.

هم‌گرایی[2]: در جریان مدل‌سازی، مقادیر پارامترها بایستی از طریق یک فرآیند تکراری به نتیجه برسند. تکرار فرایند تعیین پارامترها باید به نقطه‌ای برسند که تغییرات پارامترها نسبت به پارامترهای دور قبل از سطح مشخصی کم‌تر باشد. به این معنا که تکرار محاسبات تغییرات زیادی در مقادیر پارامترها به وجود نیاورد. رسیدن به این نقطه مشخص هم‌گرایی نام دارد که مطلوب مدل‌سازی است. ازجمله عوامل تأثیرگذار بر رسیدن به هم‌گرایی، مدل است که انتخاب شده است، به‌طوری‌که در مدل‌های ساده‌ترِ تشخیصی شناختی (مثل *DINA*، *NIDA* یا حالت راش مدل تشخیصی کلی) با تعداد کمی از آزمودنی‌ها و تعداد متوسطی از سؤال و خصیصه (چند آزمودنی، ۲۰ تا ۴۰ سؤال و ۴ تا ۶ خصیصه) رسیدن به هم‌گرایی به سادگی امکان‌پذیر است. ولی در مدل‌های پیچیده‌تر (مثل شبکه استنباط بیزی[3] یا مدل یک شکل بازپارامتری شدهٔ غیرجبرانی[4]) که حجم نمونه بسیار بزرگ و تعداد زیادی سؤالات و خصیصه‌ها را می‌طلبند، رسیدن به هم‌گرایی دشوارتر است (راپ و تمپلین، ۲۰۰۸).

سرعت هم‌گرایی با الگوریتم مورد استفاده نیز در ارتباط است. در مدل DINA با نرم‌افزار Ox یا R و یا در مدل تشخیصی کلی با نرم‌افزار mdltm از الگوریتم انتظار-بیشینه‌سازی[5] استفاده می‌شود و نتایج در چند ثانیه یا چند دقیقه به‌دست می‌آید. در مقابل مدل‌های یک شکل بازپارامتری شدهٔ غیرجبرانی کامل با نرم‌افزار Arpeggio Suite یا مدل شبکه استنباط بیزی با استفاده از الگوریتم زنجیره مونت‌کارلو مارکوف[6] برآورد می‌شوند که رسیدن به هم‌گرایی چندین ساعت یا حتی چندین

1. Misspecitied
2. Convergence
3. Bayesian Inference network (BIN)
4. NC-RUM
5. Expectation- Maximization algorithm (EM)
6. Markov Chain Monte Carlo *(*MCMC*)*

روز نیز طول می‌کشد (راپ و تمپلین، ۲۰۰۸؛ روسوس، تمپلین و هنسون، ۲۰۰۷). به مشکل فوق تفسیر نتایج را نیز اضافه کنید که به دانش پیشرفته‌ای از برآورد بیز و مواردی از این دست نیاز دارد. در صورت عدم هم‌گراییِ تکرارها می‌توان گام‌های تشکیل دهندهٔ مدل، ماتریس Q و پارامترها را تجدیدنظر کرد. به‌طور مثال، اگر مشاهده شود خصیصه‌ای به سؤالی اختصاص دارد که گستره وسیعی از دشواری‌ها را پوشش می‌دهد، احتمال عدم هم‌گرایی قوت می‌یابد (دی‌بلو، روسوس و استات، ۲۰۰۷).

در مدل‌سازی داده‌های درس علوم مطالعهٔ تیمز، عدم رسیدن به هم‌گرایی در برخی موارد ملاحظه شد. همان گونه که در قبل بیان شد، عدم تعیّن ماتریس Q یکی از اصلی‌ترین عوامل عدم رسیدن به هم‌گرایی بود. علاوه‌براین، به علت استفاده از حجم نمونه بزرگ و تعداد سؤالات زیاد، رسیدن به هم‌گرایی با فواصل بسیار کم دشوار بود. بنابراین، برای رسیدن به هم‌گرایی هم ماتریس Q تجدیدنظر شد و هم پارامترهای هم‌گرایی تغییر پیدا کرد. در تغییر پارامترهای مربوط به هم‌گرایی در مواردی که هم‌گرایی حاصل نمی‌شد با بیشتر کردن فاصلهٔ قابل قبول بین تکرارها امکان وقوع هم‌گرایی تسهیل گردید.

نمونه‌گیری‌های پیچیده: نمونه‌گیری مطالعهٔ تیمز به لحاظ بهره‌گیری از روش نمونه‌گیری خاص، که در آن مدارس به صورت خوشه‌ای و دو مرحله‌ای انتخاب می‌شوند، ویژگی‌های خاصی دارد که پیچیدگی‌های خاصی را به تحلیل داده‌ها تحمیل می‌کند. نمونه‌گیری‌های پیچیده به نمونه‌گیری‌های ماتریسی اطلاق می‌شود که در آن افراد تنها به بخشی از سؤالات پاسخ می‌دهند. در این نمونه‌ها معمولاً افراد با طی چندین مرحله به‌صورت خوشه‌ای از کلاس‌ها برگزیده می‌شوند. در نتیجه واریانس خطا در این نمونه‌ها بیشتر از نمونه‌گیری‌های تصادفی ساده بوده و افراد نمونه با احتمالات نامتناسبی انتخاب می‌شوند. هم‌چنین، نمونه‌های پیچیده خطاهای همبسته‌ای را به‌وجود می‌آورند که باید در مدل‌سازی توجه شود. علاوه‌براین، به لحاظ شیوهٔ خاص نمونه‌گیری، به‌کارگیری وزن‌های نمونه‌گیری در تحلیل‌های مطالعهٔ تیمز لازم است. برای برخورد مناسب با هر یک از این مشکلات راه‌حل‌هایی پیشنهاد شده است. برای اطلاع بیشتر به منابع روتوسکی، گنزانس، یانکاس و ون‌داویر[1] (۲۰۱۰) و راپ، تمپلین و هنسون (۲۰۱۰) مراجعه کنید. چگونگی برخورد مناسب با این مشکلات در مدل‌سازی شناختی-تشخیصی نیز یکی از چالش‌های مدل‌سازی است. تاکنون تنها مدل تشخیصی کلی برخورد مناسب با این مشکلات را در فرایند مدل‌سازی مورد توجه قرار داده است.

با توجه به اینکه نمونه مطالعهٔ تیمز جزو نمونه‌های پیچیده به‌شمار می‌رود، استفاده از روش مدل‌سازی که قابلیت تحلیل نمونه‌های پیچیده را داشته باشد ضروری است. بدین منظور از مدل تشخیصی کلی استفاده شد. در تحلیل، هم‌چنین وزن‌های نمونه‌گیری و هم‌چنین خطای استاندارد جک‌نایف نیز به‌کار گرفته شد.

1. Rutkowski, Gonzalez, Joncas & von Davier

بحث و نتیجه‌گیری

سنجش شناختی-تشخیصی به‌منظور ارائه اطلاعات اصلاحی برای تشخیص مناسب‌تر در مورد شناخت و یادگیری افراد طرح‌ریزی شده است. در این روش سعی می‌شود که نقطه ضعف سایر روش‌های سنجش شامل ارتباط کم‌تر با نظریات شناختی و عدم توجه به چگونگی یادگیری برطرف شود. با اجرای سنجش شناختی تشخیصی، اطلاعات دقیق‌تری در مورد نحوهٔ یادگیری افراد به‌دست می‌آید که می‌تواند در زمینهٔ برنامه‌ریزی برای رفع کاستی‌ها و در نتیجه بهبود یادگیری به‌کار آید.

تحلیل در زمینهٔ سنجش شناختی-تشخیصی نیازمند مدل‌سازی است. با این وجود، فرایند مدل‌سازی در سنجش شناختی-تشخیصی با مدل‌سازی در سایر روش‌های سنجش آموزشی متفاوت است. این تفاوت‌ها ناشی از ماهیت سنجش شناختی-تشخیصی به‌خصوص مقوله‌ای فرض کردنِ سطح متغیرها، چندبُعدی بودن مقیاس‌سازی و مواردی از این دست است. در نوشته حاضر، چالش‌هایی که در هنگام تحلیل داده‌های سنجش شناختی-تشخیصی به‌وجود می‌آید و باعث تغییراتی در رویه‌های مدل‌سازی می‌شود، بحث شد. موارد بحث‌شده شامل تک‌بُعدی بودن در مقابل چندبُعدی بودن، تعداد خصیصه‌ها، همبستگی بین خصیصه‌ها، تعداد مناسب سؤال در هر خصیصه، درجهٔ دقت خصیصه‌ها، اعتبار خصیصه‌ها، روایی سنجش شناختی تشخیصی، پارامترهای سؤال، برازش مدل، شناسایی و تعیّن مدل، هم‌گرایی، و نمونه‌گیری‌های پیچیده بودند که هر یک به‌طور جداگانه بحث شدند. آگاهی از این چالش‌ها پژوهشگران سنجش شناختی-تشخیصی را نسبت به دشواری‌های مدل‌سازی شناختی-تشخیصی آگاه‌تر می‌سازد.

اصلی‌ترین پیچیدگی در مدل‌سازی شناختی-تشخیصیِ داده‌های سنجش کلان‌مقیاسی همانند مطالعهٔ تیمز، توجه هم‌زمان به چندین چالش است که به‌طور معمول در مدل‌سازی‌های تک‌بُعدی به‌وجود نمی‌آیند. به‌دلیل آنکه هیچ اطلاعاتی دربارهٔ تعداد ابعاد زیربناییِ سنجش وجود نداشت، فرایند مدل‌سازی از ابعاد کم شروع به‌سمت مدل‌های با ابعاد بیشتر ادامه پیدا کرد ولی تعداد خصیصه‌ها آن چنان زیاد درنظر گرفته نشد که با کاهش تعداد سؤالات در خصیصه‌ها، اعتبار آنها کاهش یافته و همبستگی بین آنها به‌طور تصنعی بالا رود. این راهبرد اگرچه منجر به افزایش تا حداقل ۸ سؤال در هر خصیصه می‌شد که از توصیه‌های هوبرمن و وان‌داویر (۲۰۰۷) بالاتر بود و درجهٔ دقت درشتی را ایجاد می‌کرد که بر خلاف توصیهٔ لیتون و گیرل (۲۰۱۱) و اسکیجز، ویکینز و هیل (۲۰۱۶) در استفاده از درجهٔ دقت ریز بود، ولی از لحاظ عملیاتی بهترین روش را برای مدل‌سازی در داده‌هایی با حجم نمونهٔ زیاد و سؤالات گسترده فراهم می‌کرد. علاوه‌براین، اصلاح مداوم ماتریس Q براساس توصیه‌های کیم (۲۰۱۱) توانست به‌طور هم‌زمان چالش‌هایی از قبیل پارامترهای سؤال، برازش مدل، هم‌گرایی و شناسایی و تعّین مدل را برطرف کند. پیروی از این راهبردها می‌تواند راهنمای مناسبی برای پژوهشگران جهت چگونگی برخورد و رفع مشکلات فراهم آورد.

منابع

Chen, Y. H., Gorin, J. S., Thompson, M. S., Tatsuoka, K. K. (2008a). An alternative examination of Chinese Taipei mathematics achievement: Application of the rule-space method to TIMSS 1999 data. In M. v. Davier & D. Hastedt (Eds.), *Issues and Methodologies in Large-Scale Assessments* (Vol. 1, pp. 23-49). Hamburg: IEA-ETS Research Institute.

Chen, Y. H., Gorin, J. S., Thompson, M. S., Tatsuoka, K. K. (2008b). Cross-cultural validity of the TIMSS-1999 mathematics test: verification of a cognitive model. *International Journal of Testing*, 8(3), 251-271.

Chiu, C. Y., Seo, M. (2009). Cluster analysis for cognitive diagnosis: An application to the 2001 PIRLS reading assessment. In M. v. Davier & D. Hastedt (Eds.), *Issues and Methodologies in Large-Scale Assessments* (Vol. 2, pp. 137-159). Hamburg: IEA-ETS Research Institute.

De la Torre, J., Douglas, J. A. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika*, 69(3), 333-353.

De la Torre, J., Minchen, N. D. (2019). *The G-DINA Model Framework. In M. von Davier & Y. Lee (Eds.), Handbook of diagnostic classification models: Models and model extensions, applications, software packages (pp. 155-170).* Switzerland, Cham: Springer.

DeCarlo, L. T. (2011). On the analysis of fraction subtraction data: the DINA model, classification, latent class sizes, and the Q-matrix. *Applied Psychological Measurement*, 35(1), 8-26.

DiBello, L. V., Roussos, L. A., Stout, W. (2007). Review of cognitively diagnostic assessment and a summary of psychometric models. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics psychometrics* (Vol. 26, pp. 979-1030). Amesterdom: Elsevier Science Publishers.

Finkelman, M. D., Kim, W., Roussos, L., Verschoor, A. (2010). A binary programming approach to automated test assembly for cognitive diagnosis models. *Applied Psychological Measurement*, 34(5), 310-326.

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.

Gierl, M. J., Cui, Y., Zhou, J. (2009). Reliability and attribute based scoring in cognitive diagnostic assessment. *Journal of Educational Measurement*, 46(3), 293-313.

Haberman, S. J., von Davier, M. (2007). Some notes on models for cognitively based skills diagnosis. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics* (Vol. 26, pp. 1031-1038). Amesterdom: Elsevier Science Publishers.

Henson, R., Roussos, L., Douglas, J., He, X. (2008). Cognitive diagnostic attribute-level discrimination indices. *Applied Psychological Measurement*, 32(4), 275-288.

Im, S., Corter, J. E. (2011). Statistical Consequences of Attribute Misspecification in the Rule Space Method. *Educational and Psychological Measurement*, 71(4), 712-731.

Jang, E. E. (2006). *Pedagogical implications of cognitive skills diagnostic assessment for teaching and learning.* Paper presented at the the annual meeting of the American Educational Research Association, San Francisco, California.

Kabiri, M., Ghazi-Tabatabaei, M., Bazargan, A., Shokoohi-Yekta, M., Kharrazi, K. (2017) Diagnosing Competency Mastery in Science: An Application of GDM to TIMSS 2011 Data, *Applied Measurement in Education,* 30(1), 27-38,

Kim, Y. H. (2011). Diagnosing EAP writing ability using the reduced reparameterized unified model. *Language Testing,* 28(4), 509-541.

Leighton, J., Gierl, M. (2007). Why cognitive diagnostic assessment? In J. Leighton & M. Gierl (Eds.), *Cognitive diagnostic assessment for education: Theory and applications* (pp. 3-18). Cambridge: Cambridge University Press.

Leighton, J. P., Cui, Y., Cor, M. K. (2009). Testing expert-based and student-based cognitive models: An application of the attribute hierarchy method and hierarchy consistency index. *Applied Measurement in Education,* 22(3), 229-254.

Leighton, J. P., Gierl, M. J. (2011). *The learning sciences in educational assessment: The role of cognitive models*. New York: Cambridge University Press.

Leighton, J. P., Gierl, M. J., Hunka, S. M. (2004). The attribute hierarchy method for cognitive assessment: A variation on Tatsuoka's rule-space approach. *Journal of Educational Measurement,* 41(3), 205-237.

Roussos, L. A., Templin, J. L., Henson, R. A. (2007). Skills diagnosis using IRT based latent class models. *Journal of Educational Measurement, 44*(4), 293-311.

Rupp, A. A., Templin, J., Henson, R. A. (2010). *Diagnostic measurement: Theory, methods, and applications*. New York: The Guilford Press.

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.

Rutkowski, L., Gonzalez, E., Joncas, M., von Davier, M. (2010). International large-scale assessment data. *Educational Researcher,* 39(2), 142-151.

Sinharay, S., Almond, R. G. (2007). Assessing fit of cognitive diagnostic models: A case study. *Educational and Psychological Measurement,* 67(2), 239-257.

Stone, C. A., Ye, F., Zhu, X., Lane, S. (2010). Providing subscale scores for diagnostic information: A case study when the test is essentially unidimensional. *Applied Measurement in Education,* 23(1), 63-86.

Skaggs, G., Wilkins, J. L. M., Hein, S. F. (2016). Grain Size and Parameter Recovery with TIMSS and the General Diagnostic Model, *International Journal of Testing*, 16(4), 310-330. DOI: 10.1080/15305058.2016.1145683

Tatsuoka, K. K., Corter, J. E., & Tatsuoka, C. (2004). Patterns of diagnosed mathematical content and process skills in TIMSS-R across a sample of 20 countries. *American Educational Research Journal,* 41(4), 901-926.

Von Davier, M. (2007). *Mixture distribution diagnostic models* (No. RR-07-32). Princeston, NJ: ETS.

Von Davier, M. (2008). A general diagnostic model applied to language testing data. *British Journal of Mathematical and Statistical Psychology*, 61(2), 287-307.

Von Davier, M. (2009). *Using the general diagnostic model to measure learning and change in a longitudinal large-scale assessment* (No. RR-09-28). Princeston, NJ: ETS.

Yamaguchi, K., Okada, K. (2018). Comparison among cognitive diagnostic models for the TIMSS 2007 fourth grade mathematics assessment. *PLoS ONE*, 13(2), e0188691.