فصلنامه مطالعات اندازه‌گیری و ارزشیابی آموزشی
سال دهم، شماره ۳۲، زمستان ۱۳۹۹، صص ۲۳٤-۲۰۵

Journal of Educational Measurement & Evaluation Studies
Vol.10, No. 32, Winter2021

کاربرد مدل‌های تشخیصی شناختی (CDM) در مطالعه کارکرد افتراقی سؤال‌های ریاضیات پایه هشتم تیمز ۲۰۰۷ در بین دختران و پسران

اصغر مینائی*
جلیل یونسی**
بهروز کاظمی دانا***

تاریخ دریافت مقاله: ۱۳۹۸/۱۰/۲۳

تاریخ پذیرش مقاله: ۱۴۰۰/۰۳/۱۸

چکیده

هدف از اجرای پژوهش حاضر این بود که کارکرد افتراقی سؤال‌های ریاضی پایه هشتم تیمز ۲۰۰۷ در بین دختران و پسران با استفاده از مدل DINA مورد آزمون قرار گیرد. نخست برازش داده با مدل و برآورد پارامترها انجام گرفت. سپس با استفاده از آزمون والد کارکرد افتراقی سؤال‌ها مورد آزمون قرار گرفت. نتایج نشان داد که از ٤٤ سؤال مورد مطالعه، ٢٦ سؤال لنگر و ۱۸ سؤال (٤٠٪) سؤال دارایDIF است که ۱۲ سؤال دارایDIF یکنواخت و ٦ سؤال غیریکنواخت است. از ۱۸ سؤال دارایDIF ، ۷ سؤال به نفع دانش‌آموزان دختر مسلط و ۱۰ سؤال به نفع دانش‌آموزان پسر مسلط دارای کارکرد افتراقی است و یک سؤال به نفع دانش‌آموزانی که مسلط نبودن دارای کارکرد افتراقی بود. همین طور از ۱۸ سؤال دارای DIF ، ۹ سؤال به نفع دانش‌آموزان دختر و ۷ سؤال به نفع دانش‌آموزان پسری که مسلط نبودند، دارای کارکرد افتراقی بودند و ۲ سؤال نیز به نفع دانش‌آموزان مسلط دارای کارکرد افتراقی بودند. می‌توان نتیجه گرفت که آزمون والد در چارچوب مدل‌های (CDM) می‌تواند علاوه بر آزمون کارکرد افتراقی یکنواخت و غیریکنواخت در شناسایی کارکرد افتراقی در افرادی مسلط و غیر مسلط در صفات به‌کارگیری شود.

واژگان کلیدی: مدل‌های تشخیصی شناختی، کارکرد افتراقی سؤال، آزمون تیمز

---

* دانشیار گروه سنجش و اندازه گیری (روان‌سنجی)، دانشگاه علامه طباطبائی، تهران، ایران
** دانشیار گروه سنجش و اندازه گیری (روان‌سنجی)، دانشگاه علامه طباطبائی، تهران، ایران
*** دانشجوی دکتری سنجش و اندازه گیری (روان‌سنجی)، دانشگاه تهران (نویسنده مسئول؛ dana_behrooz@yahoo.com)

مقدمه

ارزشیابی برای بقا و تداوم مطلوب هر نظام آموزشی همچون هر سیستم دیگر لازم است. ارزشیابی پیشرفت تحصیلی یا به‌عبارتی ارزشیابی از فعالیت‌های یاددهی-یادگیری آموزش‌وپرورش از جایگاه ویژه‌ای برخوردار است. این ارزشیابی زمانی خواهد توانست نقش اساسی خود را ایفا کند که مبتنی بر اصول و قواعد و تکنیک‌های حوزه سنجش و اندازه‌گیری پیشرفت تحصیلی باشد. آنچه در استفاده از هر آزمون آموزشی و روان‌شناسی باید در نظر گرفته شود عادلانه بودن آزمون برای تمامی متقاضیان و سودار نبودن آن علیه دسته‌ای از جامعه شرکت‌کنندگان است (زومبو[1]، ۱۹۹۹). ابزار اندازه‌گیری نباید افرادی را که ازنظر ویژگی مورد اندازه‌گیری در سطح یکسان قرار دارند متفاوت نشان دهد. در این غیر این صورت گفته می‌شود که آزمون دارای سوگیری است (هورن و مک آردل[2]، ۱۹۹۲). یکی از مهم‌ترین تهدیدها برای روایی[3] آزمون، سوگیری سؤال یا فقدان تغییرناپذیری اندازه‌گیری[4] است که از آن با عنوان «کارکرد افتراقی سؤال[5] (DIF)» یاد می‌شود (گرامی‌پور و فلسفی‌نژاد، ۱۳۹۳).

تأمین فرصت‌های برابر و فراهم آوردن شرایطی برای آزمودنی‌ها تا بتوانند دانش و توانایی خود را در آزمون به نمایش بگذارند ضرورتی انکارناپذیر است. تعیین اینکه تفاوت‌های مشاهده شده بین عملکرد گروه‌ها در آزمون‌های پیشرفت تحصیلی ناشی از تفاوت‌های واقعی آنان در توانایی مورد اندازه‌گیری است یا ناشی از سوگیری ابزار است و واقعی نیست، مهم و اساسی است (ملاهادی، ۱۳۸۸). فهم تفاوت‌های گروه‌ها در ریاضیات برای پژوهشگران، معلمان و پرورش‌دهندگان آزمون‌ها در جهت دستیابی به هدف‌ها و غایت‌های آموزش‌وپرورش از اهمیت اساسی برخوردار است. از این‌رو، شناسایی سوگیری یا DIF سؤال‌های ریاضی به پرورش‌دهندگان آزمون‌ها در ساخت سنجش‌های عادلانه و همچنین در استنتاج‌های معتبر و روا از نمره‌های آزمون‌ها کمک می‌کند (فلاحی‌سرشت، ۱۳۹٤).

---

1. Zumbo
2. Horn & MacArdle
3. validity
4. measurement invariance
5. differential item functioning

آزمون تیمز[1] از مهم‌ترین سنجش‌های بین‌المللی است و نمره‌های حاصل از آن، هسته اصلی مقایسه عملکردهای کشورها، تصمیم‌گیری‌های مربوط به سیاست‌های آموزشی و پژوهشی در زمینهٔ نتایج ارزیابی است (مینائی، ۱۳۹۱). مطالعات زیادی در خصوص کارکرد افتراقی این آزمون صورت گرفته است. ازجمله این مطالعات می‌توان به ارسیکان و کاه[2] (۲۰۰۵)، ارسیکان (۱۹۹۸)، بیتون[3] (۱۹۹۸)، وو و همکاران[4] (۲۰۰۷) اشاره کرد. به‌عنوان مثال، ارسیکان (۱۹۹۸) دریافت که تعداد ۱۸ سؤال (۲۶درصد) از ۷۰ سؤال مورد مطالعه در درس علوم TIMSS سال ۱۹۹۵ ۱۸ سؤال دارای DIF بودند. از این ۱۸ سؤال، ۱۰ سؤال به نفع گروه مرجع[5] که دانش‌آموزان انگلیسی‌زبان بودند و تعداد ۸ سؤال به نفع گروه هدف[6] که دانش‌آموزان فرانسوی‌زبان کشور کانادا بودند، عمل می‌کردند. پژوهش ارسیکان و کاه (۲۰۰۵) نیز نتایج مشابهی به دست داد. آنها دریافتند که سؤال‌های هر دو بخش ریاضی و علوم علیه دانش‌آموزان فرانسوی‌زبان عمل می‌کنند. ارسیکان و مک‌کریس[7] (۲۰۰۲) نیز دریافتند که در درس ریاضی تیمز ۱۹۹۵، از ۱۵۶ سؤال مورد مطالعه، ۲۲ سؤال در بین دانش‌آموزان فرانسوی و انگلیسی‌زبان کشور کانادا ۶۱ سؤال در بین دانش‌آموزان کشور فرانسه و انگلستان و از بین ۱۵۴ سؤال ۹۱ سؤال در بین دانش‌آموزان کشور فرانسه و آمریکا دارای (DIF) هستند (نقل از مینائی، ۱۳۹۲).

در مطالعات مربوط به کارکرد افتراقی سؤال‌های تیمز برای انجام تحلیل‌ها به روش‌های مختلفی عمل شده است. در برخی از مطالعات مانند ترسی و همکاران[8] (۲۰۰۷) صرفاً از روش مبتنی بر نظریهٔ سؤال- پاسخ [9] (IRT)، در برخی دیگر مانند ارسیکان (۱۹۹۸؛ به نقل از مینائی، ۱۳۹۲) از روش χ^2 منتل هنزل[10] (MH)، و در برخی دیگر از تحلیل عاملی تأییدی و IRT (مانند مینائی، ۱۳۹۱؛ مینائی و غفاری، ۱۳۹۴) استفاده شده است.

---

1. Trends in International MathematicsScience Study
2. Ercikan & Koh
3. Beaton
4. Wu et al.
5. Reference group
6. Focal group
7. Ercikan & McCreith
8. Teresi et al.
9. Item response theory
10. Mantel-Haenszel

به‌طور کلی، رویکردهای روان‌سنجی و آماری بررسی کارکرد افتراقی سؤال (DIF) را می‌توان به سه دسته یا طبقه تقسیم‌بندی کرد که شامل: ۱- روش‌های مبتنی بر نظریه کلاسیک (CTT)؛ ۲- روش‌های مبتنی بر نظریه سؤال پاسخ (IRT)؛ ۳- روش‌های مبتنی بر نظریه تشخیصی شناختی[1] (CDM) است.

از جمله روش‌های شناسایی DIF بر اساس نظریه کلاسیک می‌توان به روش‌های منتل هنزل و منتل هنزل تعمیم داده شده[2] (GMH) (پنفیل و آلجینا[3]، ۲۰۰۳؛ سو و وانگ[4]، ۲۰۰۵)، روش اختلاف میانگین استانداردشده (دورانز و هالند[5]، ۱۹۹۳) و روش رگرسیون لجستیک (سوامیناتان و راجرز، ۱۹۹۰؛ زومبو، ۱۹۹۹) و روش مجذور خی که شومان[6] در ۱۹۷۹ ارائه کرد. این روش‌ها عموماً بر اساس نظریه کلاسیک آزمون هستند، زیرا نمره‌های مشاهده‌شده آزمون به‌عنوان متغیر همتا برای پیش‌بینی نمره‌های واقعی در نظر گرفته می‌شوند (بیرنباوم و همکاران[7]، ۱۹٦۸)؛ اما به دلیل ضعف‌های مطرح شده در نظریه کلاسیک مانند تغییرپذیری اندازه‌گیری در استفاده از روش‌های کلاسیک، پژوهشگران به مدل‌سازی متغیرهای مکنون یا نظریه سؤال - پاسخ روی آورده‌اند. روش‌های مختلفی برای بررسی DIF در رویکرد IRT مطرح شده است که می‌توان به روش‌هایی چون؛ مقایسه پارامترهای برآورد شده با الگوی یک پارامتری، مقایسه منحنی‌های ویژگی‌های سؤال برای گروه‌های مختلف و روش‌های مبتنی بر آزمون برازش الگو اشاره کرد (فتوحی، ۱۳۸۷). از محدودیت‌های مهم روش‌های DIF مبتنی بر IRT می‌توان به‌ضرورت وجود مدل‌های تک‌بعدی، شکل توزیع جامعه، مقدار DIF، نوع DIF (یکنواخت و غیریکنواخت بودن) و طول آزمون، انسجام توزیع‌های توانایی، یکسانی منحنی ویژگی سؤال و نیاز به حجم بالای نمونه اشاره کرد. در این بین، برای بررسی DIF در مدل‌های تشخیصی شناختی (CDMs)، لازم است به دلیل پرداختن به مشخصات تسلط ویژگی‌های مکنون به جای رتبه در سطح توانایی مکنون، به شکل جدیدی تعریف شود. مدل‌های تشخیصی شناختی، علاوه بر برآورد توانایی کلی فرد در یک سازه، نیمرخی را

---

1. cognitive diagnostic mode
2. Generalized Mantel-Haenszel
3. Penfield &Algina
4. Su & Wang
5. Dorans & Holland
6. Shoeman
7. Birnbaum et al.

برای هر آزمودنی ارائه می‌کنند که نشانگر میزان تسلط وی به مجموعه‌ای از صفات و مهارت‌ها اساسی از پیش تعیین‌شده برای موفقیت در آزمون است و می‌توان بر اساس نتایج آنها به بررسی تفاوت در مهارت‌ها و صفات در گروه‌های مختلف به‌منظور دستیابی به اطلاعات تشخیصی و افتراقی معتبر و همراه با جزئیات اقدام کرد (افضلی و همکاران، ۱۳۹۳). در طی چند دهۀ اخیر، پژوهش‌های گسترده‌ای دربارۀ مبحث یادگیری شناختی تشخیصی دانش‌آموزان صورت‌گرفته (دلاتوره[۱]، ۲۰۱۱؛ دلاتوره و داگلاس، ۲۰۰٤؛ امبرستون[۲]، ۱۹۹۷؛ هارتز، ۲۰۰۲؛ جانکر و سیجتسما، ۲۰۰۱؛ تاتسوکو، ۱۹۸۵؛ تمپلین[۳]، ۲۰۰٤). تعداد زیادی مدل‌های تشخیصی شناختی (CDMs) برای فراهم آوردن اطلاعات جزئی درباره نقاط قوت و ضعف دانش‌آموزان در یادگیری، ارائه شده است، شامل: مدل سلسله‌مراتبی صفت[٤] (AHM) (لینگتون و همکاران[٥]، ۲۰۰٤)، مدل تشخیصی شناختی لگ خطی (LCDM) (هنسون و همکاران[٦]، ۲۰۰۹)، مدل مرتبه بالاتر دینا (HO-DINA[۷]) (دلاتوره و داگلاس، ۲۰۰٤) و مدل تعمیم‌یافته DINA (G-DINA[۸]) (دلاتوره، ۲۰۱۱)، مدل (DINA) (دلاتوره و داگلاس، ۲۰۰٤؛ هارتل[۹]، ۱۹۸۹؛ جانکر و سیجتسما[۱۰]، ۲۰۰۱)، مدل (DINO) (تمپلین و هنسون، ۲۰۰٦) و مدل یکپارچۀ کاهش‌یافته باز برآورد شده (RUM[۱۱]) (هارتز[۱۲]، ۲۰۰۲). مدل تشخیصی عمومی (GDM) (داوییر[۱۳]، ۲۰۰۵)، این تلاش‌ها به محبوبیت فزاینده این مدل‌های تشخیصی شناختی منجر شده و نیز موجب تحولی در بین پژوهشگران شده تا مدل‌های CDM را برای پژوهشگران کاربردی ساده‌تر و امکان‌پذیرتر شود (دلاتوره و لی[۱٤]، ۲۰۱۳).

---

1. De La Torre
2. Embretson et al.
3. Templin
4. Attribute Hierarchy Model
5. Leighton et al.
6. Henson et al.
7. Higher-order Deterministic Inputs, Noisy (AND) gate
8. Generalized DINA
9. Haertel
10. Junker & Sijtsma,
11. reparameterized unified model
12. Hartz
13. Davier
14. De la Torre & Lee

همان‌طور که بیان شد انواع متعددی از مدل‌های تشخیصی شناختی وجود دارد که می‌تواند جنبه جبرانی و غیر جبرانی صفت‌های مورد نیاز به پاسخگویی به سؤال‌ها را مدنظر قرار دهند. به نظر می‌رسد موضوعی مانند آزمون ریاضی تیمز که صفات غیر جبرانی برای حل مسئله نیاز باشد و به تبعیت از پژوهش‌های افضلی و همکاران (۱۳۹۳)؛ محسن‌پور و همکاران (۱۳۹٤)؛ مینائی و همکاران (۱۳۹۳)؛ تاتساکو و همکاران (۲۰۰٤)؛ [1] چوی و همکاران[2] (۲۰۱۵) که در پژوهش‌های خود صفات مورد نیاز برای حل مسائل ریاضی را غیر جبرانی در نظر گرفته‌اند. همچنین، به‌کارگیری مدل‌های غیر جبرانی[3] می‌تواند در به دست آوردن اطلاعات تشخیصی و افتراقی در علوم ریاضی قابل اعتمادتر باشد.

مدل DINA[4] از ساده‌ترین مدل‌های غیر جبرانی تشخیصی شناختی است که فقط به دو پارامتر تغییرپذیر نیاز دارد. این مدل اساس سایر مدل‌های استفاده شده در آزمون‌های تشخیصی شناختی است. مدل DINA یک مدل غیر جبرانی و عطفی است و فرض می‌کند که یک آزمودنی باید تمام صفت‌های مورد نیاز را داشته باشد تا بتواند به یک سؤال به‌درستی پاسخ دهد (هنسون و همکاران، ۲۰۰۹). برای هر سؤال، پاسخ‌های آزمودنی‌ها در دوطبقه مکنون نمره‌گذاری می‌شوند: یکی طبقه‌ای که نشان‌دهنده پاسخ‌دهی صحیح است (نمره ۱)، شامل آزمودنی‌هایی است که تمامی صفات مورد نیاز برای پاسخ‌دهی صحیح به سؤال را دارا هستند؛ طبقه دیگر که نشان‌دهنده پاسخ‌گویی غلط به سؤال است (نمره صفر)، شامل آزمودنی‌هایی است که حداقل یکی از صفات مورد نیاز برای پاسخگویی صحیح به سؤال را ندارند (دلاتوره، ۲۰۱۱). تحلیل پاسخ‌های آزمون دهندگان با استفاده از مدل مدل‌های تشخیصی شناختی (CDM ها) مزیت ارائه اطلاعات تشخیصی را داراست؛ اما برای اطمینان از اعتبار نتایج حاصل از این مدل‌ها، کارکرد افتراقی سؤال در CDM باید بیشتر مورد بررسی قرار گیرد (دلاتوره، ۲۰۱٤).

تعداد محدودی پژوهش وجود دارند که DIF را با استفاده از قالب CDM می‌سنجند (هو و همکاران[5]، ۲۰۱٤؛ لی[6]، ۲۰۰۸؛ ژانگ، ۲۰۰٦). ژانگ (۲۰۰٦) با بررسی DIF با

---

1. Tatsuoka et al.
2. Choi et al.
3. Compensatory
4. Deterministic Input Noisy and Gate Model
5. Hou et al.
6. Li

رویکرد مدل‌های CDM و با به‌کارگیری روش‌های MH[1] و SIBTEST و مدل (DINA) نتایج را مورد مقایسه قرارداد. مدل DINA در نمایه صفات ملاک مرجع بهتر از روش‌های MH و SIBTEST بهتر عمل کرد، این دو روش قدرت کمی در شناسایی DIF غیریکنواخت دارند. البته این رویکرد نقص‌های متعدی دارد. اول اینکه به‌منظور کسب برآورد بی‌طرف و ناریب پژوهش باید بدون DIF باشد در غیر این صورت برآوردها برای پروفایل صفت‌ها ناعادلانه بوده و متغیر برآورد شده نامعتبر خواهد بود. درنتیجه باعث بی‌اعتباری ارزیابی DIF می‌شود. آزمون‌های واقعی بدون نقص نیستند و معمولاً شامل سؤال‌های DIF دار خواهند بود. دستیابی به یک آزمون بی‌نقص بسیار دشوار است از سوی دیگر، اگر این فرضیه معتبر باشد، ارزیابی DIF الزامی نیست. دوم، صفت‌های ارزیابی‌شده به‌عنوان برآوردهای واقعی و بدون خطا در نظر گرفته شده است اما همان‌طور که می‌دانیم در عمل ممکن است برآورد پارامترها دارای میزان قابل‌توجهی عدم قطعیت باشند. سوم، همان‌طور که ژانگ (۲۰۰۶) مشاهده کرد این روش قدرت خود را در شناسایی DIF های غیریکنواخت از دست می‌دهد.

هو و همکاران (۲۰۱٤) آزمون والد[2] را برای شناسایی DIF در مدل DINA به کار گرفتند. به‌طوری که مدل DINA برای داده‌های گروه‌های مرجع و هدف به طور جداگانه، بررسی شده و اختلاف در برآورد پارامترهای سؤال به روش آماری والد بررسی شد. شبیه‌سازی آنها نشان داد که روش آزمون والد برای ارزیابی DIF کارآمد است و تحت تأثیر درصد سؤال‌های دارای DIF موجود در آزمون قرار نمی‌گیرد. هو و همکارانش این مطالعهٔ آزمایشی و شبیه‌سازی‌شده را طراحی کردند تا امکان‌پذیری آزمون والد (Wald-CDM) را برای بررسی DIF سؤال‌ها مورد آزمایش قرار دهند؛ نتایج پژوهش آزمایشی آنا نشان داد که آماره والد (Wald –CDM) روشی مناسب برای شناسایی DIF که در داده‌های تشخیصی شناختی است. در بیشتر شرایط شبیه‌سازی‌شده، روش والد در مقایسه با روش‌های SIBTEST و MH بهتر عمل کرد در مواردی نیز از آنها سبقت گرفت (جورج و رابیتج[3]، ۲۰۱٤). با توجه به لزوم مقایسه پذیری سؤال‌ها و نتایج آزمون، در پژوهش حاضر تلاش شد تا کارکرد افتراقی سؤال‌های ریاضی پایه هشتم تیمز ۲۰۰۷ در دانش‌آموزان دختر و پسر ایران با استفاده از مدل DINA که یکی از مدل‌های غیر جبرانی

---

1. Mantel-Hanzel
2. Wald statistic
3. George & Robitzsch

است و در مقایسه با سایر مدل‌های تشخیص شناختی، نسبت به حجم نمونه، حساس نیست، بررسی شود. از این‌رو، پرسش پژوهش حاضر این بود که آیا در چارچوب مدل DINA سؤال‌های ریاضی پایه هشتم تیمز ۲۰۰۷ در دانش‌آموزان دختر و پسر دارای کارکرد افتراقی (DIF) هستند؟

روش

از آنجا که در پژوهش حاضر به آزمون کارکرد افتراقی سؤال‌های ریاضی پایه هشتم تیمز که در سال ۲۰۰۷ برگزار شد، پرداخته شد می‌توان آن را جزو مطالعات تحلیل ثانویه به‌حساب آورد. جامعه آماری پژوهش همان جامعه آماری پایه هشتم در آزمون ریاضی تیمز ۲۰۰۷ است که در سال تحصیلی ۱۳۸۵–۱۳۸٦ به اجرا درآمد. در سال تحصیلی مورد نظر تعداد ۱/٤۷۵/۳٦۸ دانش‌آموز سوم راهنمایی در ۲۹/۹۵٦ کلاس، مشغول به تحصیل بودند (السن و همکاران[1]، ۲۰۰۸) است. نمونه پژوهش حاضر بخشی از دانش‌آموزان ایرانی سال سوم راهنمایی شرکت‌کننده در تیمز ۲۰۰۷ است که به دفترچه‌های شماره ۱، ۲، ۳ و ۱٤ آزمون تیمز ۲۰۰۷ پاسخ داده‌اند و تعداد آنها برابر با ۱۱۳۷ است.

در خصوص نحوه انتخاب این بلوک‌ها باید متذکر شد که آزمون ریاضی تیمز دارای ۱٤ دفترچه است که از ۱ تا ۱٤ شماره‌گذاری شده‌اند. هر دفترچه با دفترچه بعد و قبل خود دارای سؤال‌های مشترک است و هر دفترچه در قالب دو بلوک تقسیم‌بندی می‌شود. سؤال‌های یکی از بلوک‌ها با دفترچه ماقبل و بلوک دیگر با دفترچه بعد، مشترک است. به‌عنوان مثال، دفترچه یک شامل بلوک ۱ و ۲ و دفترچه دو شامل بلوک ۳ و ٤ که سؤال‌های بلوک ۳ دفترچه دو همان سؤال‌های بلوک دو دفترچه ۱ است. به دلیل اینکه در مدل تشخیصی شناختی برای برآورد مناسب به نمونه‌ای با حجم بالا مورد نیاز است، داده‌های بلوک‌های مشابه را با یکدیگر ادغام کردیم تا حجم نمونه افزایش یابد. ابزار مورد استفاده در این پژوهش، آزمون ریاضی پایه هشتم تیمز ۲۰۰۷ بود که دارای ۲۱۵ سؤال است و در ۱٤ دفترچه که می‌توان گفت در فرم‌های موازی ارائه شده‌اند (مینائی، ۱۳۹۲). هرکدام از دفترچه‌ها بین ۳۲ –۳٦ سؤال دارند. آزمون ریاضی پایه هشتم، چهار حیطه محتوایی و سه حیطه شناختی را پوشش می‌دهد.

---

[1]. Olson et al.

سؤال‌های آزمون تیمز به دو صورت پاسخ‌گزین[۱] (دو یا چندگزینه‌ای) و پاسخ‌ساز[۲] است. در پژوهش حاضر، به‌منظور سهولت در تحلیل داده‌ها و به پیروی از مینائی (۱۳۹۱) و لی و همکاران[۳] (۲۰۱۱) سؤال‌های چندارزشی به سؤال‌های دوارزشی تبدیل شدند. اگر فردی در یک سؤال چندارزشی بالاترین نمره را به دست آورده بود، نمره ۱ و چنانچه نمره کامل نگرفته بود، نمره صفر تعلق می‌گرفت.

به دلیل وجود تعداد بالای سؤال‌های آزمون تیمز نمونه‌ای از بلوک‌ها انتخاب شد، بلوک ۱ ،۲ ، ۳ و ١٤ به‌عنوان بلوک‌های مورد بررسی، انتخاب شدند. تحلیل‌ها بر اساس ٤٤ سؤال انجام گرفت. برای انجام تحلیل‌ها از SPSS، پکیج NPCD (ژنگ و همکاران[٤]، ٢٠١٦)، پکیج CDM (روبتزوچ و همکاران[٥]، ٢٠١٦)، پکیج DIFT (کرونتس و کرونتس[٦]، ٢٠١٥) و پکیج GDINA (ونچو و دلاتوره[٧]، ٢٠١٦) که در محیط R اجرا می‌شوند، استفاده شد.

گام اول- تهیه ماتریس Q: در ماتریس Q رابطه میان سؤال‌ها و صفات نمایش داده می‌شود. در پژوهش حاضر از ماتریس Q استفاده شد که توسط مینائی (۱۳۹۱) تهیه شده بود و با استفاده از پکیج NPCD (ژنگ و همکاران، ٢٠١٦) و با به‌کارگیری روش (RSS[٨]) اصلاحات جزئی روی ماتریس اولیه به منظور برازش بهتر مدل با داده‌ها انجام گرفت. در مدل‌هایی که برازش مدل پس از طراحی و اجرای آزمون انجام می‌گیرد و ماتریس کیو زیر نظر متخصصان ساخته می‌شود همواره دارای خطا و انحراف است؛ از این‌رو، پروفایل صفات آزمودنی با استفاده از به حداقل رساندن مجموع مجذورات باقی‌مانده بین پاسخ واقعی[۹] و پاسخ مورد انتظار[۱۰] (RSS) با استفاده از روش ناپارامتری[۱۱]

---

1. Multiple- choice
2. Constructed response
3. Lee et al.
4. Zheng et al.
5. Robitzsch et al.
6. Cervantes & Cervantes
7. Wenchao
8. residual sum of square
9. real responses
10. ideal responses
11. nonparametric

تخمین زده می‌شود (چیو[1]، ۲۰۱۳). به‌طوری که مهارت‌ها و صفاتی که برای سؤال j برای فرد i دارای کمترین میزان برای RSS که طبق فرمول زیر برآورد شده و در صورت نیاز ماتریس جهت نیاز یا عدم نیاز به صفت مورد تحلیل و اصلاح قرار می‌گیرد.

$$RSS_j = \sum_{i=1}^{N} (Yij - \acute{\eta}\ ij\ )^2$$

$\acute{\eta}ij$ مقدار مورد انتظار پاسخ به سؤال j ام توسط فرد i ام

Yij مقدار مورد واقعی پاسخ به سؤال j ام توسط فرد i ام

ماتریس نهایی در جدول (۱) ارائه شده است. در اینجا علامت * نشان‌دهنده اصلاحات جزئی روی ماتریس اولیه با استفاده از روش (RSS) است. یادآور می‌شود در پژوهش حاضر از ماتریس Q استفاده شد که توسط مینائی (۱۳۹۱) و کورتر و تاتسکو (۲۰۰۲) ارائه شده بود. این ماتریس شامل ۲۰ صفت بود که برای تمام سؤال‌های آزمون ریاضی تیمز که متشکل از ١٤ دفترچه است، تهیه شده بود. با توجه به اینکه در پژوهش حاضر صرفاً از سؤال‌های بلوک ۱، ۲، ۳ و ١٤ که مربوط به دفترچه‌های شماره ۱، ۲ و ١٤ هستند، استفاده شد، عملاً تعدادی از صفات در سؤال‌ها شرکت ندارند و از ماتریس نهایی این پژوهش کنار گذاشته شدند.

[1]. Chiu

جدول (۱) ماتریس Q نهایی

	مهارت / سؤالات	C2	C3	C4	C5	P1	P2	P3	P5	P6	P7	P9	S3	S5	S7	S8	S9	S10	S11
بلوک ۱	M01_01	۱	–	۰	–	۰	۰	۱	۰	–	۰	–	۱	۱	–	–	۱	۰	۰
	M01_02	۱	–	۰	–	۱	❋۰	۰	۰	–	۰	–	۰	۰	–	–		۱	۱
	M01_03	۰	–	۱	–	۰	۰	۱	۰	–	۱	–	۱	❋۰	–	–	۱	۰	۰
	M01_04	۰	–	۰	–	۱	۱	۰	۱	–	۰	–	۰	۱	–	–	۰	۱	۰
	M01_05	۰	–	۱	–	۱	۱	۰	۱	–	۰	–	۰	۰	–	–	۰	۱	۱
	M01_06	۱	–	۰	–	۱	۱	۱	۱	–	۰	–	۰	۱	–	–	۰	۰	۱
	M01_07	۰	–	۰	–	۱	۱	۱	۰	–	۰	–	۰	۱	–	–	۰	۰	۱
	M01_08	۰	–	❋۰	–	۰	۱	۱	۱	–	۱	–	۱	۱	–	–	۱	۰	۰
	M01_09	۱	–	۰	–	۰	۱	❋۱	۰	–	۰	–	۰	۱	–	–	❋۱	۰	۰
	M01_10	❋۰	–	۰	–	❋۰	❋۰	۰	۱	–	۱	–	۱	۰	–	–	۰	۱	۱
	M01_11A	۱	–	۱	–	۱	۱	۰	۱	–	۱	–	۱	۰	–	–	۱	۱	۱
	M01_11B	۱	–	۱	–	۱	۱	۰	۰	–	۱	–	۱	۰	–	–	۱	۱	۰
	M01_12	۰	–	۱	–	۱	۱	۱	۱	–	۱	–	۱	۰	–	–	۱	۱	۰
بلوک ۲	M02_01	–	۰	۱	–	۰	۰	۰	۰	۰	۱	۰	۰	❋۱	–	۰	–	–	۰
	M02_02	–	۰	۱	–	۰	۱	۰	۰	۰	۰	۱	۰	۰	–	۰	–	–	۰
	M02_03	–	۰	۰	–	۰	۱	❋۱	۱	۱	۱	۱	۰	۰	–	۰	–	–	۰
	M02_04	–	۰	۱	–	۰	۱	۰	۱	۱	۰	۱	❋۰	۰	–	۱	–	–	۱
	M02_05	–	۰	۱	–	۰	۱	۰	۰	۱	۱	۱	۱	۰	–	۱	–	–	۱
	M02_06	–	۰	۱	–	۰	❋۰	۰	۰	۰	۰	۱	۰	۰	–	۱	–	–	۰
	M02_07A	–	۱	۰	–	۱	۱	❋۰	۰	❋۰	۱	۰	۱	۱	–	۱	–	–	۱
	M02_07B	–	۱	۰	–	۱	۱	۱	۱	۱	۱	۰	❋۰	۱	–	۱	–	–	۰
	M02_07C	–	۱	۰	–	۱	۱	۰	۱	۱	۱	۱	۱	۱	–	۰	–	–	۰
	M02_08	–	۱	۰	–	۰	۱	۰	۱	۱	۱	۱	۱	۰	–	۰	–	–	۱
	M02_09	–	۱	۱	–	۰	۱	۱	۰	۰	۱	۰	۰	۱	–	۰	–	–	۰
	M02_10	–	۰	۱	–	۱	۱	۱	۰	۱	۱	۱	۱	۱	–	۱	–	–	۰
	M02_11	–	۰	۱	–	۱	۱❋	۱	۰	۰	۱	۱	۰	۱	–	۱	–	–	۰
	M02_12	–	۰	۱	–	۱	۱	۱	۰	۱	۱	۱	۰	۰	–	۰	–	–	۱
	M02_13	–	۱	۱	–	۱	۱	۰❋	۱	۱	۱	۰	۱	۰	–	۰	–	–	۱
	M02_14	–	۱	❋۱	–	۱	۱	۱	۱	۱	۱	۱	۱	۰	–	۰	–	–	۱

	مهارت / سؤالات	C2	C3	C4	C5	P1	P2	P3	P5	P6	P7	P9	S3	S5	S7	S8	S9	S10	S11
بلوک ۳	M03_01	–	–	–	۰	۰	۱	۰	۰	۰	۱	۰	۰	۱	۰	–	–	۰	–
	M03_02	–	–	–	۱	۱	۱	۱	۱	۰	۱	۱	۰	۱	۱	–	–	۰	–
	M03_03	–	–	–	۰	۱	۰	۱	۰	۱	۱	۱	۱	۰	۱	–	–	۰	–
	M03_04	–	–	–	۰	۱	۱	۱	۱	۰	۰	۱	۱*	۰	۰	–	–	۰	–
	M03_05	–	–	–	۰	۰	۰	۱	۰	۰	۱	۰	۱	۱	۰	–	–	۱	–
	M03_06	–	–	–	۱	۱	۱	۱	۱	۰	۱*	۰	۱	۱	۰	–	–	۰	–
	M03_07	–	–	–	۰	۰	۰	۱	۰	۰	۰	۱	۱	۰	۰	–	–	۱	–
	M03_08	–	–	–	۰	۱	۱	۱	۱	۰	۰	۱	۰	۰	۱	–	–	۰	–
	M03_09	–	–	–	۰	۰	۰	۰*	۰	۰	۱	۱	۱	۱	۰	–	–	۱	–
	M03_10	–	–	–	۰	۱	۰	۱	۰	۰	۰*	۰	۱	۱	۰	–	–	۰	–
	M03_11	–	–	–	۱	۱	۰	۱	۰	۱	۱	۱	۱	۱	۰	–	–	۰	–
	M03_12	–	–	–	۱	۰	۰	۰*	۰	۱	۱	۰*	۱	۱	۰	–	–	۰	–
	M03_13	–	–	–	۰	۱	۰	۰*	۰	۰	۰	۰	۱	۱	۰	–	–	۰	–
	M03_14	–	–	–	۰	۱	۱	۱	۱	۰	۱	۱	۱	۱	۰	–	–	۰	–
	M03_15	–	–	–	۰	۰	۱	۱	۱	۱	۱	۱	۱	۰	۰	–	–	۱	–

گام دوم – برازش مدل، برآورد پارامترهای مدل: انواع متعددی از مدل‌های تشخیصی شناختی وجود دارد. در پژوهش حاضر، با توجه به غیر جبرانی بودن صفات در ریاضی، از مدل DINA که یکی از مدل‌های غیر جبرانی است، استفاده شد. برازش مدل با روش بیشینه درست‌نمایی (ML[1]) انجام گرفت. برازش مدل با داده‌ها با استفاده از ملاک اطلاعات آکائیک (AIC[2])، ملاک اطلاعات بیزی (BIC)، (SRMSR) و (RMSEA)[3] مورد ارزشیابی قرار گرفت.

روش برآورد پارامترها در بیشتر پکیج نرم‌افزار R برای مدل‌های CDM به‌طور پیش‌فرض روش احتمال وزنی حداقل مربعات، فراوانی‌های مورد انتظار برای طبقات

[1]. Maximum likelihood
[2]. Akaike information criterion
[3]. Bayesian Information Criterion , standardized root mean square root of squared residuals, Root Mean Square Error

مهارت‌ها (WLS[۱]) است که قابل تغییر به روش‌های احتمال غیر وزنی حداقل مربعات فراوانی مورد انتظار (ULS[۲]) و بیشینه درست‌نمایی (ML[۳]) است.

روش مبتنی بر الگوریتم بیشینه درست‌نمایی (ML) در نرم‌افزار R دارای سرعت کندتری نسبت به روش‌های دیگر است؛ اما «مقاوم» [٤] و «ثبات» [٥] بیشتری در برآورد پارامترها، نسبت به روش‌های دیگر دارد (روبیتزوچ و همکاران[٦]، ۲۰۱٦). ویژگی «ثبات»، الگوریتم ML به این معنی است که با افزایش تعداد مشاهدات دقت برآورد افزایش می‌یابد (امبرستون و رایس، ۲۰۰۰). ویژگی «مقاوم» الگوریتم ML به این معنا است که برآوردها تحت تأثیر مقادیر نامتعارف بسیار بزرگ یا بسیار کوچک قرار نمی‌گیرد. به‌ویژه برای داده‌هایی که دارای توزیع نرمال نیستند. روش‌های آماری مقاوم برای حل بسیاری از مسائل رایج مانند تخمین پارامتر مکان، پارامتر مقیاس و رگرسیون خطی توسعه یافته‌اند به‌طوری که بی‌جهت تحت تأثیر داده‌های پرت قرار نمی‌گیرند (هابر[۷]، ۲۰۱۱). بر همین اساس، الگوریتم برآورد در پژوهش حاضر بر اساس بیشینه درست‌نمایی (ML) انتخاب شد.

گام سوم- شناسایی سؤال‌های دارای کارکرد افتراقی: برای پاسخگویی به سؤال‌های پژوهش از آزمون والد استفاده شد. برای تعیین یکنواخت و غیریکنواخت بودن و تعیین جهت DIF بر اساس تسلط و عدم تسلط افراد به صفات در گروه‌های مورد مقایسه، اختلاف پارامتر لغزش در بین گروه کانونی و گروه مرجع، (Δsj) و اختلاف پارامتر حدس برای گروه کانونی و مرجع، (Δgj)، بررسی شد. نحوه تفسیر Δgj و Δsj به‌طور خلاصه در جدول (۲) ارائه شده است.

تعیین گروه مرجع و هدف اختیاری است و در محاسبات DIF تأثیری ندارد (تی و همکاران[۸]، ۲۰۱۵، به نقل از اکبری ورمزیار، ۱۳۹۳) و معمولاً گروه مرجع، گروه اکثریت

---

۱. weights probabilitiesleast squares

۲. unweighted least squares

۳. maximization of the marginal likelihood

٤. stable

٥. consistency

6. Robitzsch et al.

۷. Huber

۸. Tay et al.

و گروه هدف، گروه اقلیت یا محروم در نظر گرفته می‌شود (سوامیناتان و راجرز، ۱۹۹۰)، پسران به‌عنوان گروه مرجع و دختران به‌عنوان گروه هدف انتخاب شدند.

جدول (۲) خلاصه تفسیر Δgj و Δsj

جهت DIF	نوع DIF	(Δ_{gj})	(Δ_{sj})
به نفع افراد مسلط در گروه کانونی و به نفع افراد غیر مسلط در گروه کانونی	یکنواخت	+	+
به نفع افراد مسلط در گروه مرجع و به نفع افراد غیر مسلط در گروه مرجع		−	−
به نفع افراد مسلط درگروه مرجع و به نفع افراد غیر مسلط در گروه کانونی	غیریکنواخت	+	−
به نفع افراد مسلط در گروه کانونی و به نفع افراد غیر مسلط در گروه مرجع		−	+
بدون DIF برای افراد مسلط و به نفع افراد غیر مسلط در گروه مرجع		−	۰
به نفع افراد مسلط در گروه مرجع و بدون DIF برای افراد غیر مسلط		۰	−
بدون DIF برای افراد مسلط و به نفع افراد غیر مسلط در گروه کانونی		+	۰
به نفع افراد مسلط در گروه کانونی و بدون DIF برای افراد مسلط		۰	+

نتایج

در جدول (۳) اطلاعات توصیفی نمونه آزمودنی‌ها و سؤال‌های تیمز مورد استفاده در پژوهش حاضر ارائه شده است.

جدول (۳) اطلاعات توصیفی نمونه آزمودنی‌ها و سؤال‌ها

	جنسیت	فراوانی	تعداد سؤال	دفترچه	حیطه محتوایی	آلفای کرونباخ
تیمز ۱	پسر	۳۱۶	۱۳	۱، ۱٤	اعداد (٤)، هندسه (۳)، جبر (٤)، احتمال (۲)	۰/۷۳
	دختر	۲۵٦				
تیمز ۲	پسر	۳۱۰	۱٦	۱، ۲	اعداد (۱۱)، هندسه (۲) جبر (۱)، احتمال (۲)	۰/۸٦
	دختر	۲٦۳				
تیمز ۴	پسر	۳۰٦	۱۵	۲، ۳	اعداد (۳)، هندسه (۲) جبر (۳)، احتمال (۷)	۰/۷٦
	دختر	۲۵۹				

بر اساس شاخص‌های برازش جدول (٤) می‌توان گفت که مدل از برازش مطلوب با داده‌ها برخوردار است.

جدول (٤) شاخص‌های برازش مدل DINA

	P	χ^2	SRMSR	RMSEA	BIC	AIC
بلوک ۱	۰/۱۹۴۳	۹/۱۴۶۸	۰/۰۳۵۶	۰/۰۳	۳۳۴۹۷	۱۵۵۷۵
بلوک ۲	/۰۰۰۴	۴۳/۷۱۷	۰/۰۴۳۱۶	۰/۰۳۲	۶۱۷۷۸	۲۶۰۰۱
بلوک ۳	۰/۲۲۹۹	۹/۳۸۳۴	۰/۰۳۹۳۳	۰/۰۴۳	۳۵۶۵۰	۱۷۷۶۱

به‌منظور تعیین برازش سؤال با مدل از شاخص RMSEA و شاخص تشخیص سؤال[1] (IDI) (لی و همکاران، ۲۰۱۲) استفاده شد. شاخص تشخیص از رابطه معکوس با شاخص‌های حدس و لغزش برخوردار است و هرچه این شاخص بزرگ‌تر باشد، بهتر است و بیانگر بالا بودن قدرت تشخیص سؤال است، اما ملاک دقیق و روشنی برای تعیین نقطه برش برای شاخص تشخیص وجود ندارد (دلاتوره، ۲۰۱۲) در مدل DINA، شاخص تشخیص سؤال از طریق محاسبه اختلاف بین احتمال پاسخ درست به سؤال توسط افرادی که به همه مهارت‌ها تسلط یافته‌اند و احتمال پاسخ درست به سؤال توسط افرادی که در همه مهارت‌ها به تسلط نرسیده‌اند، محاسبه می‌شود (راپ و همکاران[2]، ۲۰۱۰). رابطه پارامتر تشخیص سؤال با پارامترهای لغزش و حدس به شکل زیر است (دلاتوره، ۲۰۰۸):

$$IDIj = 1 - sj - gj$$

شاخص RMSEA نیز میزان برازندگی سؤال با مدل شناختی را نشان می‌دهد، این شاخص بر اساس توزیع مجذور کای است، هر چه میزان این شاخص کوچک‌تر باشد، نشان‌دهنده برازش بالای سؤال با مدل است. به‌صورت معمول سؤال‌ها با شاخص RMSEA کوچک‌تر از ۰/۰۵ را سؤال دارای بهترین برازندگی با مدل و سؤال‌های با RMSEA بزرگ‌تر از ۰/۱ را فاقد برازش با مدل معرفی می‌کنند (افضلی، ۱۳۹۳). با توجه به جدول (۵) می‌توان نتیجه گرفت بیشتر سؤال‌ها از برازش مطلوبی با مدل برخوردارند.

---

1. Item discrimination index
2. Rupp et al.

جدول (۵) شاخص‌های برازش سؤال‌ها با مدل DINA

	سؤال‌ها	پسران			دختران		
		تشخیص (IDI)	تمیز	RMSEA	تشخیص (IDI)	تمیز	RMSEA
بلوک ۱	M01_01	۰/۶۶۷	۰/۴۳۰	۰/۰۲۰	۰/۶۳۸	۰/۳۹۸	۰/۰۱۳
	M01_02	۰/۱۹۳	۰/۱۰۹	۰/۰۷۵	۰/۲۶۳	۰/۲۴۸	۰/۰۵۵
	M01_03	۰/۳۷۳	۰/۳۰۳	۰/۰۵۰	۰/۳۲۶	۰/۲۶۶	۰/۰۳۸
	M01_04	۰/۸۰۶	۰/۶۲۱	۰/۰۱۹	۰/۷۳۲	۰/۵۹۳	۰/۰۱۹
	M01_05	۰/۲۹۰	۰/۲۰۳	۰/۰۶۱	۰/۲۲۲	۰/۱۳۷	۰/۰۵۹
	M01_06	۰/۳۸۸	۰/۲۳۴	۰/۰۶۱	۰/۴۸۳	۰/۳۵۹	۰/۰۳۶
	M01_07	۰/۶۴۶	۰/۴۰۳	۰/۰۲۸	۰/۴۹۴	۰/۳۵۱	۰/۱۰۴
	M01_08	۰/۷۶۵	۰/۵۳۷	۰/۰۸۵	۰/۶۱۷	۰/۴۱۵	۰/۰۲۳
	M01_09	۰/۱۸۱	۰/۳۸۹	۰/۰۴۹	۰/۱۰۵	۰/۳۴۰	۰/۰۶۶
	M01_10	۰/۳۹۲	۰/۲۷۳	۰/۰۴۱	۰/۳۰۱	۰/۲۳۵	۰/۰۶۱
	M01_11A	۰/۶۴۸	۰/۵۷۸	۰/۰۲۶	۰/۴۸۲	۰/۴۵۶	۰/۰۲۷
	M01_11B	۰/۲۶۲	۰/۳۹۳	۰/۰۲۰	۰/۳۰۸	۰/۴۴۸	۰/۰۳۶
	M01_12	۰/۷۰۵	۰/۵۴۴	۰/۰۴۳	۰/۵۰۱	۰/۴۲۵	۰/۰۳۳
	میانگین	۰/۴۸۵	۰/۳۸۵	۰/۰۴۴	۰/۴۲۰	۰/۳۵۹	۰/۰۴۳
بلوک ۲	M02_01	۰/۵۵۲	۰/۳۳۳	۰/۰۶۷	۰/۴۶۰	۰/۲۲۴	۰/۰۶۰
	M02_02	۰/۳۸۳	۰/۲۶۲	۰/۰۷۷	۰/۵۷۵	۰/۳۱۴	۰/۰۵۵
	M02_03	۰/۴۲۷	۰/۳۴۰	۰/۰۶۶	۰/۳۲۶	۰/۲۲۸	۰/۰۷۵
	M02_04	۰/۴۷۰	۰/۳۱۶	۰/۰۴۹	۰/۵۴۴	۰/۳۱۳	۰/۰۶۰
	M02_05	۰/۴۴۳	۰/۳۳۶	۰/۰۴۰	۰/۳۸۲	۰/۲۹۹	۰/۰۴۲
	M02_06	۰/۶۰۸	۰/۴۲۶	۰/۰۳۳	۰/۳۱۹	۰/۳۰۵	۰/۰۴۸
	M02_07A	۰/۵۰۶	۰/۳۳۲	۰/۰۶۸	۰/۵۱۴	۰/۱۸۶	۰/۰۸۱
	M02_07B	۰/۸۶۷	۰/۵۲۴	۰/۰۴۶	۰/۸۳۹	۰/۴۵۴	۰/۰۶۸
	M02_07C	۰/۶۷۲	۰/۵۵۳	۰/۰۰۳	۰/۸۳۸	۰/۵۵۲	۰/۰۰۸
	M02_08	۰/۱۳۴	۰/۳۰۹	۰/۰۳۲	۰/۲۰۲	۰/۲۴۲	۰/۰۳۹
	M02_09	۰/۴۶۵	۰/۳۲۳	۰/۰۵۷	۰/۳۱۷	۰/۲۱۹	۰/۰۸۳
	M02_10	۰/۱۹۸	۰/۱۸۳	۰/۰۷۶	۰/۲۲۵	۰/۲۳۴	۰/۰۹۷
	M02_11	۰/۲۲۱	۰/۲۵۹	۰/۱۱۳	۰/۲۵۶	۰/۳۵۰	۰/۱۱۵
	M02_12	۰/۳۸۶	۰/۳۵۹	۰/۰۷۲	۰/۳۲۵	۰/۲۳۶	۰/۰۸۴
	M02_13	۰/۳۶۵	۰/۳۴۵	۰/۱۲۷	۰/۳۸۲	۰/۳۵۹	۰/۰۸۱
	M02_14	۰/۴۶۴	۰/۴۲۵	۰/۰۵۰	۰/۵۶۱	۰/۴۸۵	۰/۰۲۰
	میانگین	۰/۴۴۷		۰/۰۶۱	۰/۴۴۱		۰/۰۶۳

	سؤال‌ها	پسران			دختران		
		تشخیص (IDI)	تمیز	RMSEA	تشخیص (IDI)	تمیز	RMSEA
بلوک ۳	M03_01	۰/٦۸۳	۰/٤۳٤	۰/۰۳۰	۰/٤۸٥	۰/٤۳٤	۰/۰٥٥
	M03_02	۰/٤٦۹	۰/۳٥٦	۰/۰۸۷	۰/٤٦۷	۰/۳٥٦	۰/۰۹۱
	M03_03	۰/٦٦٦	۰/٤٥۹	۰/۰٦٦	۰/٥۹٤	۰/٤٥۹	۰/۰۲۸
	M03_04	۰/۱٥۹	۰/۱۹۱	۰/۱۱٦	۰/۳٤۸	۰/۱۹۱	۰/۰۸۰
	M03_05	۰/۳٥٤	۰/۳٦۷	۰/۰٥۸	۰/۳۹۳	۰/۳٦۷	۰/۰۸۱
	M03_06	۰/۲۷۸	۰/۲٦٤	۰/۰۷٥	۰/۳۰۲	۰/۲٦٤	۰/۰۷٥
	M03_07	۰/٦۲٦	۰/٤٥۲	۰/۰۷٥	۰/٥۹٥	۰/٤٥۲	۰/۰۳۸
	M03_08	۰/٦٥۲	۰/٤۲۳	۰/۰٤۰	۰/٤۲۳	۰/٤۲۳	۰/۰٤٤
	M03_09	۰/۲٥۰	۰/۳۸٥	۰/۰٤٦	۰/۲۲٦	۰/۳۸٥	۰/۰٥٦
	M03_10	۰/۱۹۱	۰/۲۱۰	۰/۰۸٤	۰/۲٦۰	۰/۲۱۰	۰/۱۱۸
	M03_11	۰/۳۳۱	۰/۳۱٦	۰/۰۹۱	۰/۳۰۱	۰/۳۱٦	۰/۰۷۲
	M03_12	۰/٥۲۸	۰/۳۸٦	۰/۰۸۷	۰/٥۹۳	۰/۳۸٦	۰/۰۷٥
	M03_13	۰/۷۰۲	۰/٤٤۱	۰/۰٤٤	۰/۸۳۷	۰/٤٤۱	۰/۰۱٦
	M03_14	۰/٥۰۰	۰/٤۳۲	۰/۰۷۹	۰/٤۰۸	۰/٤۳۲	۰/۱۰۹
	M03_15	۰/٥۹۸	۰/٤۷۸	۰/۰۹۳	۰/٤۱۰	۰/٤۷۸	۰/۰۸۰
	میانگین	۰/٤٦٥		۰/۰۷۱	۰/٤٤۲		۰/۰٦۷

برای شناسایی سؤال‌های دارای DIF از آزمون والد استفاده شد که نتایج آن در جدول (٦) ارائه شده است. این آزمون که دارای توزیع خی‌دو است، به آزمون معنی‌داری تفاوت بین پارامترهای حدس و لغزش در بین گروه مرجع و هدف می‌پردازد که اگر معنادار باشد بیانگر کارکرد افتراقی سؤال است؛ برای تعیین اندازه یا بزرگی DIF از شاخص اندازه اثر UA[1] استفاده شد که مقادیر آن در جدول (٦) ارائه شده است. پژوهشگران پیشنهاد می‌کنند که مقادیر کمتر از ۰/۰٥۹ به عنوان DIF ناچیز، مقادیر بین ۰/۰٦ تا ۰/۰۸۷ به عنوان DIF متوسط و مقادیر بزرگ‌تر از ۰/۰۸۸ به عنوان DIF بزرگ تفسیر

[1]. Unsigned area

شود (راوند و روبیتز[1]، ۲۰۱۵). جدول (٦) نشان می‌دهد که از ٤٤ سؤال، ٢٦ سؤال لنگر و ۱۸ سؤال (٤٠٪) دارای DIF است که ۸ سؤال از DIF ناچیز و ۱۰ سؤال از DIF بزرگ برخوردارند.

جدول (٦) آماره آزمون والد و اندازه اثر

	سؤال	آماره والد	df	p	اندازه اثر UA	لنگر و DIF	اندازه DIF
بلوک ۱	M01_01	۱/۳۷٦	۲	۰/۵۰۲	۰/۰۲۰	لنگر	–
	M01_02	۲/٤۲٤	۲	۰/۲۹۷	۰/۰٤۰	لنگر	–
	M01_03	۱/۸۵۷	۲	۰/۳۹۵	۰/۰۳۳	لنگر	–
	M01_04	۸۷/۹۸٤	۲	۰/۰۰۰	۰/۰۲۷	DIF	کوچک
	M01_05	۳/۳۳۵	۲	۰/۱۸۸	۰/۰٤۱	لنگر	–
	M01_06	۳/٤٤۵	۲	۰/۱۷۸	۰/۰٦۷	لنگر	–
	M01_07	۳/۲۸۲۰	۲	۰/۱۹۳	۰/۰۲۸	لنگر	–
	M01_08	۳/۳۷۷	۲	۰/۱۸٤	۰/۱۸۳	لنگر	–
	M01_09	۸۰۱/۳٦۵	۲	۰/۰۰۰	۰/۰۰۹	DIF	کوچک
	M01_10	۵/۵۹۹۸	۲	۰/۰٦۰	۰/۰۸۷	لنگر	–
	M01_11A	۳۲٦/۳۳۸	۲	۰/۰۰۰	۰/۱۲۱	DIF	بزرگ
	M01_11B	۵۳۵٦/۰۳	۲	۰/۰۰۰	۰/۰۲٦	DIF	کوچک
	M01_12	۱۳/۸۱۰	۲	۰/۰۰۱	۰/۲۷٦	DIF	بزرگ
بلوک ۲	M02_01	۱۳/٦۰٦	۲	۰/۰۰۱	۰/۰۱٤	DIF	کوچک
	M02_02	۰/۹٦۱۵	۲	۰/٦۱۸	۰/۰٤٦	لنگر	–
	M02_03	۲۷/٤٤۸	۲	۰/۰۰۰	۰/۰٤٦	DIF	کوچک
	M02_04	۲/۲٤۱۷	۲	۰/۳۲٦	۰/٦٦٤	لنگر	–
	M02_05	۱/۱۰۱۵	۲	۰/۵۷٦	۰/۰۱۲	لنگر	–
	M02_06	۲۸/۱۳۳	۲	۰/۰۰۰	۰/٦۸۰	DIF	کوچک
	M02_07A	۲/۷۲٤	۲	۰/۲۵٦	۰/۱۹٦	لنگر	–
	M02_07B	۱۷/۵۹۵	۲	۰/۰۰۰	۰/۲۱۲	DIF	بزرگ
	M02_07C	۱۷۱/۲۳	۲	۰/۰۰۰	۰/۳٤۷	DIF	بزرگ
	M02_08	۳۳/٤۵۰	۲	۰/۰۰۰	۰/۰۱۷	DIF	کوچک
	M02_09	۲/۷٤۹	۲	۰/۰۲۰	۰/۲۱۲	DIF	بزرگ
	M02_10	۰/٤۷۸	۲	۰/۷۸۷	۰/٤٤۱	لنگر	–
	M02_11	۱/۳۲۱	۲	۰/۵۱٦	۱/۵۲	لنگر	–
	M02_12	۱۲/۳٦۰	۲	۰/۰۰۲	٤/۳۱۵	DIF	بزرگ

[1]. Ravand & Robitzsch

	سؤال	آماره والد	df	p	اندازه اثر UA	لنگر و DIF	اندازه DIF
	M02_13	۱/۱۰۲	۲	۰/۹۵۰	۰/۹۷۸	لنگر	–
	M02_14	۱۵/٦٤۹	۲	۰/۰۰۰	۰/۲۷۷	DIF	بزرگ
بلوک ۳	M03_01	۱۱۷/۵۸۵	۲	۰/۰۰۰	۰/۰٤۹	DIF	کوچک
	M03_02	۳/۹۸۰	۲	۰/۱۳٦	۲/۳۱	لنگر	–
	M03_03	۱/٦۳۳	۲	۰/٤٤۱	۰/۸۷۹	لنگر	–
	M03_04	۱/٤۷۳	۲	۰/۲۹۰	۰/٤۱۷	لنگر	–
	M03_05	٤/۸٤۳	۲	۰/۰۸۸	۰/۰۲۰	لنگر	–
	M03_06	۱/٦۲۵	۲	۰/٤٤۳	۰/۰۱٦	لنگر	–
	M03_07	۰/۹۰۲	۲	۰/٦۳٦	۰/۰۲۱	لنگر	–
	M03_08	۹/۸۸	۲	۰/۰۰۷	۰/۲۸۵	DIF	بزرگ
	M03_09	۰/٦۷٦	۲	۰/۷۱۳	۰/۰۰۰	لنگر	–
	M03_10	۲٤/۱۵٤	۲	۰/۰۰۰	۰/۰۲۷	DIF	بزرگ
	M03_11	۱/۵٤۵	۲	۰/۷٦۱	۸/۲٤۲	لنگر	–
	M03_12	۰/۸٤۸	۲	۰/٦۵٤	۰/۳۲۱	لنگر	–
	M03_13	٤/٤۵٤	۲	۰/۱۰۷	۰/۰۱۷	لنگر	–
	M03_14	۰/۵۷۸	۲	۰/۷٤۸	۰/٦۷۳	لنگر	–
	M03_15	۱۰/۲۵۰	۲	۰/۰۰۵	۲/۲٤۱	DIF	بزرگ

در جدول (۷) تفاوت بین مقادیر پارامتر لغزش در گروه مرجع و هدف (Δ_{sj}) و همچنین تفاوت بین پارامتر حدس در گروه مرجع و هدف (Δ_{gj})، برای تعیین یکنواخت یا غیریکنواخت بودن و جهت DIF بر اساس تسلط یا عدم تسلط در صفات ارائه شده است. شاخص CDIF نیز که برای تعیین جهت تأثیر DIF بر گروه‌های مرجع و هدف در چارچوب رویکرد DIFT[1] به کار می‌رود با استفاده از پکیج DIFT (کرونتس و کرونتس[2]، ۲۰۱۵) که در محیط R اجرا می‌شود، ارائه شده است. اگر علامت CDIF[3] مثبت باشد بیانگر این است که سؤال به نفع گروه مرجع (پسران) و اگر علامت آن منفی باشد یعنی سؤال به نفع گروه هدف (دختران) دارای کارکرد افتراقی است (مینائی، ۱۳۹۲). همان‌طور که جدول (۷) نشان می‌دهد، ۱۲ سؤال دارای Δ_{gj} و Δ_{sj} هم‌علامت هستند که نشان‌دهنده DIF یکنواخت این سؤال‌ها است. همچنین، ٦ سؤال دارای Δ_{gj} و Δ_{sj} با

[1]. Differerital Functioning of Item and Test
[2]. Cervantes& Cervantes
[3]. Compensatory Differerntial Item Functtioning

علامت نایکسان هستند که بیانگر DIF غیریکنواخت این سؤال‌ها است. بر اساس علامت شاخص CDIF، ٤ سؤال به نفع گروه هدف (دختران) و ١٤ سؤال به نفع گروه مرجع (پسران) دارای کارکرد افتراقی است.

جدول (۷) شاخص CDIF و Δ_{gj} و Δ_{sj}

سؤال	Δ_{gj}	Δ_{sj}	CDIF	نوع DIF	جهت DIF	جهت DIF برحسب تسلط در صفات
M01_01	۰/۰۰۵	۰/۰۲٤	۰/۳۸۸	–	–	–
M01_02	−۰/۰۷۱	−۰/۰۰۲	۰/۱۰٦	–	–	–
M01_03	۰/۰۳۷	−۰/۰۱۱	۰/۳۳٦	–	–	–
M01_04	−۰/۰۳۳	−۰/۱۰۷	۰/۹۱٤	یکنواخت	به نفع گروه مرجع	به نفع افراد مسلط درگروه مرجع(R) و به نفع افراد غیر مسلط در گروه مرجع(R)
M01_05	−۰/۰٦٤	−۰/۱۳۱	۰/۰۰٦	–	–	–
M01_06	−۰/۰۸٤	۰/۰۱۲	۰/۲٦۹	–	–	–
M01_07	۰/۰۰۲	−۰/۱۵۰	۰/۳۸۲	–	–	–
M01_08	۰/۰۸۹	−۰/۰۵۹	۰/۵۹۳	–	–	–
M01_09	−۰/۰۰۱	−۰/۰۷٦	۰/۰۵۸	یکنواخت	به نفع گروه مرجع	به نفع افراد مسلط در (R) و به نفع افراد غیرمسلط در (R)
M01_10	−۰/۰۹۸	−۰/۱۸۹	۰/۰۳۷	–	–	–
M01_11A	−۰/۰۲٦	−۰/۱۹۳	۰/۲۵۲	یکنواخت	به نفع گروه مرجع	به نفع افراد مسلط در (R) و به نفع افراد غیر مسلط در (R)
M01_11B	۰/۰۰۰	−۰/۰٤٦	۰/۲۰٦	یکنواخت	به نفع گروه مرجع	به نفع افراد مسلط در (R) و بدون DIF برای افراد غیر مسلط
M01_12	۰/۰۳۹	−۰/۱٦٤	۰/٤۸۱	غیریکنواخت	به نفع گروه مرجع	به نفع افراد مسلط در (R) و به نفع غیر مسلط در گروه (F)

سؤال	Δ_{gj}	Δ_{sj}	CDIF	نوع DIF	جهت DIF	جهت DIF برحسب تسلط در صفات
M02_01	۰/۱۶۳	۰/۰۷٤	-۰/۱۰۷	یکنواخت	به نفع گروه هدف	به نفع افراد مسلط در (F)و به نفع افراد غیر مسلط در (F)
M02_02	-۰/۰۵۹	۰/۱۳۳	۰/۲۰۹	–	–	–
M02_03	-۰/۰۲۸	-۰/۱۲۸	۰/۰۱۲	یکنواخت	به نفع گروه مرجع	به نفع افراد مسلط در (R)و به نفع افراد غیر مسلط در (R)
M02_04	-۰/۰٦۸	۰/۰۰٦	۰/۰۳٤	–	–	–
M02_05	۰/۰۳٦	-۰/۰۲۵	۰/۱۳٦	–	–	–
M02_06	۰/۳۰٤	۰/۰۱۵	۰/۱۹۹	یکنواخت	به نفع گروه مرجع	به نفع افراد مسلط در (F) و به نفع افراد غیر مسلط در (F)
M02_07A	۰/۰۸۰	۰/۰۸۹	۰/۱۲۹	–	–	–
M02_07B	۰/۰۳۵	۰/۰۰۷	۰/۱۹۱	یکنواخت	به نفع گروه مرجع	به نفع افراد مسلط در (R)و به نفع افراد غیر مسلط در (R)
M02_07C	۰/۰۰۰	۰/۱٦٦	۰/۱۷۹	یکنواخت	به نفع گروه مرجع	به نفع افراد مسلط در (F)و بدون DIF برای افراد غیر مسلط
M02_08	-۰/۰۱۱	۰/۰۵۷	۰/۰۱٦	غیریکنواخت	به نفع گروه مرجع	به نفع افراد مسلط در (F)و به نفع افراد غیر مسلط در (R)
M02_09	۰/۱۳۲	-۰/۰۱٦	۰/۱۳۱	غیریکنواخت	به نفع گروه مرجع	به نفع افراد مسلط در (R)و به نفع افراد غیر مسلط در (F)
M02_10	۰/۰۰۸	۰/۰۳٦	۰/۱۲۷	–	–	–
M02_11	۰/۰۳۷	۰/۰۷٤	۰/۱۳۱	–	–	–

جهت DIF برحسب تسلط در صفات	جهت DIF	نوع DIF	CDIF	Δ_{sj}	Δ_{gj}	سؤال
به نفع افراد مسلط در (F)و به نفع افراد غیر مسلط در (F)	نفع گروه مرجع	یکنواخت	۰/۱۶۳	۰/۱۲۳	۰/۱۸٤	M02_12
-	-	-	۰/۱۳٦	۰/۰٥٤	۰/۰۳٦	M02_13
به نفع افراد مسلط در (F)و به نفع افراد غیر مسلط در (R)	به نفع گروه مرجع	غیریکنواخت	۰/٤۰۳	۰/۰٤٤	-۰/۰٥۳	M02_14
به نفع افراد مسلط در (R)و به نفع افراد غیر مسلط در (F)	به نفع گروه هدف	غیریکنواخت	-۰/۰٤۹	-۰/۱۷٦	۰/۰۲۲	M03_01
-	-	-	۰/۰۰۹	۰/۰۷۷	۰/۰۷۹	M03_02
-	-	-	۰/۰۲۱	-۰/۰۷۲	۰/۰۰۰	M03_03
-	-	-	-۰/۰۰۲	۰/۰۹۸	-۰/۰۹۲	M03_04
-	-	-	۰/۰۰۹	۰/۰٤۱	۰/۰۰۲	M03_05
-	-	-	۰/۰۱۳	۰/۰۱٥	-۰/۰۰۹	M03_06
-	-	-	۰/۰۱٤	۰/۰۱٤	۰/۰٤٦	M03_07
بدونDIF برای افراد مسلط و غیر مسلط در (F)	به نفع گروه هدف	یکنواخت	-۰/۰۰۹	۰/۰۰۰	۰/۲۳۲	M03_08
-		-	۰/۰۰۹	-۰/۰۱۲	۰/۰۱۲	M03_09
به نفع افراد مسلط در (F) و به نفع افراد غیر مسلط در (F)	به نفع گروه هدف	یکنواخت	-۰/۰۰۳	۰/۰۸۳	۰/۰۱٤	M03_10
-	-	-	۰/۰۲٤	۰/۰۰۰	۰/۰۳۰	M03_11
-	-	-	۰/۰۱۷	۰/۰۳۹	-۰/۰۲۷	M03_12
-	-	-	-۰/۰۲٦	۰/۱۷۰	۰/۰۳٥	M03_13
-	-	-	۰/۰۰۹	۰/۰۰۰	۰/۰۹۲	M03_14
به نفع افراد مسلط در (R) و به نفع افراد غیر مسلط در (F)	به نفع گروه مرجع	غیریکنواخت	۰/۰۱۱	-۰/۰٤۱	۰/۱٤۸	M03_15

بحث و نتیجه‌گیری

در جهان امروز، بسیاری از تصمیم‌ها بر اساس نتایج آزمون‌ها صورت می‌گیرد. یکی از آزمون‌های بسیار مهم که در سطح بین‌المللی اجرا می‌شود آزمون بین‌المللی تیمز است. ازآنجاکه این آزمون بین گروه‌های مختلف دختر و پسر و قومیت‌های مختلف اجرا می‌شود امکان وجود سؤال‌های سودار به نفع یا ضرر گروه خاصی به دور از انتظار نیست؛ بنابراین، در پژوهش حاضر کارکرد افتراقی سؤال‌های آزمون ریاضی پایه هشتم تیمز ۲۰۰۷ در بین دانش‌آموزان دختر و پسر در چارچوب رویکرد سنجش تشخیصی شناختی (CDMs) با استفاده از مدل DINA و بر اساس صفات و مهارت‌های شناختی ارائه شده توسط مینائی (۱۳۹۱) مورد مطالعه قرار گرفت.

یافته‌ها نشان می‌دهد که از ۱۸ سؤال دارای کارکرد افتراقی، ۱۲ سؤال دارای DIF یکنواخت و ٦ سؤال دارای DIF غیریکنواخت است، ۷ سؤال به نفع دانش‌آموزان دختر که در صفات مورد نیاز سؤال‌ها به تسلط رسیده‌اند و ۹ سؤال نیز به نفع دانش‌آموزان دختر که در صفات مورد نیاز سؤال‌ها به تسلط نرسیده‌اند، دارای کارکرد افتراقی هستند. تعداد ۱۰ سؤال به نفع دانش‌آموزان پسر که در صفات مورد نیاز سؤال‌ها به تسلط رسیده‌اند و ۷ سؤال نیز به نفع دانش‌آموزان پسر که در صفات مورد نیاز سؤال‌ها به تسلط نرسیده‌اند دارای کارکرد افتراقی هستند. بر اساس شاخص CDIF ، ۳ سؤال به نفع گروه هدف (دختران) و ۱۵ سؤال به نفع گروه مرجع (پسران) دارای DIF هستند. این یافته به میزان زیادی با پژوهش نتایج مینائی و غفاری (۱۳۹٤)، صادقی (۱۳۹۲)، هاشمی (۱۳۸۹) و فابر (۲۰۰۹) که در خصوص آزمون ریاضی تیمز در دوره‌های مختلف با روش‌های مختلف در بافت نظریه کلاسیک و سؤال – پاسخ انجام گرفته‌اند، همخوانی دارد. فابر (۲۰۰۹) کارکرد افتراقی سؤال‌های پایه هشتم تیمز ۲۰۰۳ را با استفاده از رویکرد اندازه‌گیری چندوجهی راش در بین دختران و پسران بررسی کرد. نتایج پژوهش او نشان داد که از پنج حوزه محتوایی، سؤال‌های دو حوزه اندازه‌گیری و جبر از DIF برخوردارند. پژوهش صادقی (۱۳۹۲) نیز نشان داد که بر اساس روش کلاسیک ٦ سؤال و در رویکرد IRT یک سؤال دارای DIF است. پژوهش مینائی و غفاری (۱۳۹٤) نیز نشان داد از ۲۱۹ سؤال آزمون ریاضی ، ۷۵ سؤال دارای DIF است که ٥٦ سؤال دارای DIF یکنواخت و ۱۹ سؤال دارای DIF غیریکنواخت بود.

فهم تفاوت‌های گروهی و جنسیتی در ریاضیات برای پژوهشگران، معلمان و توسعه‌دهندگان آزمون به جهت رسیدن به هدف‌ها و غایت‌های آموزش‌وپرورش حیاتی

است. در پژوهش‌های جدید مانند پژوهش دیندیال[1] (۲۰۰۸؛ به نقل از مینائی و غفاری، ۱۳۹٤) بر این موضوع تأکید شده که این تصور و ایده قالب که دختران در مقایسه با پسران در ریاضیات از توانایی کمتری برخوردارند باید کنار گذاشته شود. مربیان تعلیم‌وتربیت باید دانش‌آموزان دختر را تشویق به مشارکت در موضوعات و فعالیت های مرتبط با ریاضی کنند تا بدین‌وسیله اعتمادبه‌نفس و علاقه بیشتری به ریاضیات پیدا کنند. پژوهش‌ها نشان می‌دهند که فرم سؤال می‌تواند تفاوت‌های جنسیتی را کاهش یا افزایش دهد، بنابراین، پرورش‌دهندگان آزمون‌ها باید تلاش کنند تا سؤال‌ها را به‌گونه‌ای طراحی کنند که تفاوت‌های جنسیتی را به حداقل برساند. با شناسایی عوامل تأثیرگذار بر تفاوت-های جنسیتی، مربیان، معلمان، سیاست‌گذاران تعلیم‌وتربیت کشور خواهند توانست برنامه‌های درسی جدید و خلاقانه طراحی و تدوین کنند یا برنامه‌های موجود را بهبود بخشند تا نظام و محیط آموزشی کشورمان بالنده و غنی شود. تشخیص سوگیری یا DIF جنسیتی به پرورش‌دهندگان آزمون در توسعه سنجش منصفانه کمک خواهد کرد و منابع مشتق شده از دستاوردهای آزمون روا و معتبر بیشتر خواهد شد. همچنین یافته‌های DIF می‌تواند به سازندگان در طراحی و توسعه برنامه‌های درسی متناسب با گروهه و همچنین تسهیل در تفسیر نتایج آزمون کمک کند (یان[2]، ۲۰۰۵).

از مشکلات روش‌شناسی در این پژوهش می‌توان به استفاده از رویکرد تشخیصی شناختی بر روی سؤال‌های تیمز اشاره کرد که برای هدف‌های تشخیصی طراحی نشده‌اند و متکی بر مفروضه تک‌بعدی بودن هستند. همچنین تبدیل داده‌های چندارزشی به داده‌های دوارزشی است. پیشنهاد می‌شود برای انجام تحلیل‌های مشابه، نخست به ساخت آزمون‌های تشخصی مختلف بر اساس رویکرد CDA اقدام شود. همچنین پیشنهاد می‌شود از دیگر مدل‌های تشخیصی شناختی در بررسی کارکرد افتراقی در آزمون‌های جدید تیمز یا آزمون‌های مشابه مانند پرلز نیز استفاده شود.

---

1. Dindyal
2. Yan

منابع

افضلی، افشین (۱۳۹۳). *مدل‌سازی تشخیصی شناختی ریاضیات پایه اول دبیرستان بر اساس روش سلسله‌مراتبی صفات*. رساله دکتری، دانشگاه علامه طباطبایی.

افضلی، افشین؛ دلاور، علی؛ فلسفی‌نژاد، محمدرضا؛ فرخی، نورعلی؛ برجعلی، احمد (۱۳۹۳). کاربرد مدل‌های تشخیصی شناختی در تعیین ماهیت تفاوت در عملکرد دانش‌آموزان دختر و پسر پایه اول دبیرستان در ریاضیات. *مجله دست آوردهای روان‌شناختی*، ٤(۲)، ۸۹-۱۰٤.

اکبری ورمزیار، شیرین (۱۳۹۳) *برآورد پارامتر و شناسایی کنش افتراقی سؤالات آزمون‌های سرنوشت‌ساز با فرمت مختلط با استفاده از مدل‌های کلاسیک و نظریه‌های جدید اندازه‌گیری*. رساله دکتری، دانشگاه علامه طباطبایی، دانشکده روان‌شناسی و علوم تربیتی، گروه سنجش و اندازه‌گیری.

امبرستون، سوزان؛ رایس، استیو (۱۳۸۸). *نظریات جدید روان‌سنجی برای روان‌شناسان*؛ ترجمه حسن پاشا شریفی، ولی‌الله فرزاد، مجتبی حبیبی، بلال ایزانلو. انتشارات رشد.

امینی، فاطمه (۱۳۸۵). *بررسی رابطه مهارت‌های عمومی و فعالیت‌های آموزشی معلمان و نمرات کسب‌شده دانش‌آموزان سال چهارم ابتدایی ایران در آزمون بین‌المللی پیشرفت تحصیلی تیمز ۲۰۰۷*. پایان‌نامه کارشناسی ارشدT.دانشگاه علامه طباطبایی، دانشکده روان‌شناسی و علوم تربیتی، گروه فناوری آموزشی.

رحیمی‌نژاد، عباس؛ خداپناهی، محمدکریم (۱۳۸۵). رابطه انگیزش پیشرفت و خودپنداشت با پیشرفت تحصیلی دانش‌آموزان پایه هشتم در درس علوم بر اساس تحلیل نتایج «تیمز_ار». *فصلنامه روان‌شناسان ایرانی*، ۲۰ (۷) ۲۰۷-۲۱۸.

شولتز، دوان؛ شولتز، سیدنی (۲۰۰۵). *نظریه‌های شخصیت*؛ ترجمه یحیی سیدمحمدی (۱۳۸٦). انتشارات ویرایش.

صادقی، مریم (۱۳۹۲). کارکرد افتراقی (DIF) سؤال‌های آزمون ریاضی پایه چهارم و هشتم تیمز ۲۰۱۱ *با استفاده از مدل‌های IRT و کلاسیک*. پایان‌نامه کارشناسی ارشد، دانشگاه آزاد اسلامی واحد تهران مرکز.

غفاری، زهرا (۱۳۹۳). *کارکرد افتراقی سؤال‌های پایه هشتم آزمون ریاضی تیمز ۲۰۱۱ در بین دانش‌آموزان دختر و پسر با استفاده از رویکرد نظریه سؤال – پاسخ (IRT)*. پایان‌نامه کارشناسی ارشد، دانشگاه آزاد اسلامی، واحد تهران.

فتوحی، لیلا (۱۳۸۷). *بررسی کارکرد افتراقی سؤال در سؤالات کنکور کارشناسی ارشد رشته روان‌شناسی سال ۱۳۸٤*. پایان‌نامه کارشناسی ارشد، دانشگاه علامه طباطبایی.

کریمی، عبدالعظیم (۱۳۸۸). مجموعه سؤال‌های علوم ریاضیات TIMSS. پژوهشگاه مطالعات آموزش‌وپرورش.

کریمی، عبدالعظیم (۱۳۸۹). مجموعه سؤال‌های قابل‌انتشار فیزیک و ریاضیات تیمز بین‌المللی پیشرفته ۲۰۰۸. مرکز ملی بین‌المللی تیمز و پرلز، پژوهشگاه مطالعات آموزش‌وپرورش.

گرامی‌پور، مسعود (۱۳۹۳). ارزیابی توان آماری تحلیل رگرسیون لجستیک در آشکارسازی کنش افتراقی سؤال‌های آزمون. *فصلنامه مطالعات اندازه‌گیری و ارزشیابی آموزشی*،٤(۸)، ۱۸۷-۲۱۱.

گرامی‌پور، مسعود؛ فلسفی‌نژاد، محمدرضا؛ دلاور، علی؛ فرخی، نورعلی (۱۳۹۱). مقایسهٔ روش‌های تحلیل عاملی تأییدی (CFA) و نسبت درست‌نمایی مبتنی بر مدل پرسش –پاسخ (IRT) در ردگیری کنش افتراقی سؤالات آزمون‌های سرنوشت‌ساز. *فصلنامهٔ اندازه‌گیری*، ۹(٤)، ۱۰۵- ۱۲۲.

محسن‌پور، مریم؛ گویا، زهرا؛ شکوهی یکتا، محسن؛ کیامنش، علیرضا؛ بازرگان، عباس (۱۳۹٤). سنجش تشخیصی صلاحیت‌های سواد ریاضی. *فصلنامه نوآوری‌های آموزشی*، ۱٤(٤) ۸-۳۵.

ملاهادی، محسن (۱۳۸۸). ارزشیابی به‌وسیله آزمون‌های چندگزینه‌ای. *مجله راهبردهای آموزشی*، ۲(٤) ۱۸۲-۲۰۵.

مینائی، اصغر (۱۳۹۱). *مدل پردازی تشخیصی شناختی (CDM) سؤال‌های ریاضیات تیمز ۲۰۰۷ در دانش‌آموزان پایه هشتم ایران با استفاده از مدل یکپارچه با پارامتر پردازی مجدد (RUM) و مقایسه مهارت‌های ریاضی دانش‌آموزان دختر و پسر*. پایان‌نامه منتشرنشده دکتری، دانشگاه علامه طباطبایی، دانشکده روان‌شناسی و علوم تربیتی.

مینائی، اصغر (۱۳۹۲). سنجش مقایسه‌پذیری سازه و تحلیل کارکرد افتراقی سؤال‌ها (DIF) و بلوک‌های (DTF) آزمون علوم پایه هشتم تیمز ۲۰۰۷ در بین دانش‌آموزان ایران و آمریکا. *فصلنامه اندازه‌گیری تربیتی*، ۱۱(٤) ۱۱۰-۱٤٦.

مینائی، اصغر؛ غفاری، زهرا (۱۳۹٤). کارکرد افتراقی سؤال‌های پایه هشتم آزمون ریاضی تیمز ۲۰۱۱ در بین دانش‌آموزان دختر و پسر با استفاده از رویکرد نظریه سؤال – پاسخ (IRT). *فصلنامه اندازه‌گیری*، ۲۱(٦)، ۲۲-۳۹.

Birnbaum, A., Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores. Some latent trait models and their use in inferring an examinee's ability.* Addison-Wesley, Reading, MA.

Chiu, C. Y. (2013). Statistical Refinement of the Q-matrix in Cognitive Diagnosis. *Applied Psychological Measurement, 37*(8), 598-618.

Choi, K. M., Lee, Y. S., & Park, Y. S. (2015). What CDM Can Tell About What Students Have Learned: An Analysis of TIMSS Eighth Grade Mathematics. *Eurasia Journal of Mathematics, Science & Technology Education, 11*(6), 1563-1577.

Davier, M. (2005). A general diagnostic model applied to language testing data. *British Journal of Mathematical & Statistical Psychology, 61*(2), 287-307.

De La Torre, J. (2011). generalized DINA model framework. *Psychometrika, 76*(2) 79-199.

De la Torre, J., & Lee, Y. S. (2013). Evaluating the Wald test for item-level comparison of saturated and reduced models in cognitive diagnosis. *Journal of Educational Measurement, 50*(4), 355-373.

Dindyal, J. (2008). An overview of the gender factor in mathematics in TIMSS-2003 for the Asia-Pacific region. *ZDM, 40*(6), 993-1005.

Embretson, S. E. & Reise, S. P. (2000). Item Response Theory for psychologists. Lawrence Erlbaum Associates.

Ercikan, K. & Koh, K. (2005). Examining the construct comparability of the English and French versions of TIMSS. *International Journal of Testing, 5*(1), 23-35.

Ercikan, K. & McCreith, T. (2002). Effects of adaptations on comparability of test items and test scores. In *Secondary analysis of the TIMSS data* (pp. 391-405)

Ercikan, K. (1998). Translation effects in international assessments. *International Journal of Educational Research, 29*(6), 543-553

George, A. C., & Robitzsch, A. (2014). Multiple group cognitive diagnosis models, with an emphasis on differential item functioning. *Psychological Test & Assessment Modeling, 56*(4), 405-432.

Haertel, E. H. (1989). Using restricted latent class models to map the skill structure of achievement items. *Journal of Educational Measurement, 26*(4), 301-321.

Hartz, S. M. (2002). *A Bayesian framework for the unified model for assessing cognitive abilities: Blending theory with practicality*.

Henson, R. A., Templin, J. L., & Willse, J. (2009). Defining a family of cognitive diagnosis models using log-linear models with latent variables. *Psychometrika*, *74*(2), 191-210.

Horn, J., & McArdle J. (1992). A practical and theoretical guide to measurement invariance in aging research. *Experimental Aging Research*, *18*(3), 117-144.

Hou, L., De La Torre, J., & Nandakumar, R. (2014). Differential item functioning assessment in cognitive diagnostic modeling: Application of the Wald test to investigate DIF in the DINA model. *Journal of Educational Measurement, 51*(1) 98-125.

Huber, P. J. (2011). *Robust statistics* (pp. 1248-1251). Springer Berlin Heidelberg.

Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement*, *25*(3), 258-272.

Lee, Y. S., Park, Y. S., & Taylan, D. (2011). A cognitive diagnostic modeling of attribute mastery in Massachusetts, Minnesota, and the US national sample using the TIMSS 2007. International Journal of Testing, 11(2), 144-177.

Leighton, J. P., & Gierl, M. J. (2007). Verbal reports as data for cognitive diagnostic assessment. *Cognitive diagnostic assessment for education: Theory and applications*, 146-172.

Penfield, R. D., & Algina, J. (2003). Applying the Liu-Agresti Estimator of the Cumulative Common Odds Ratio to DIF Detection in Polytomous Items. *Journal of Educational Measurement*, *40*(4), 353-370.

Ravand, H., & Robitzsch, A. (2015). Cognitive Diagnostic Modeling Using R. *Practical Assessment, Research, and Evaluation*: Vol. 20 , Article 11.

Robitzsch, A., Kiefer, T., George, A. C., & Uenlue, A. (2016). CDM: Cognitive diagnosis modeling. *R package version*, *3*

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

Su, Y., & Wang, W. (2005). Efficiency of the Mantel, generalized Mantel–Haenszel, and logistic discriminant function analysis methods in detecting differential item functioning for polytomous items. *Applied Measurement in Education*, *18*(4) 313-350.

Tay, L., Meade, A. W., & Cao, M. (2015). An overview and practical guide to IRT measurement equivalence analysis. *Organizational Research Methods, 18*(1), 3-46.

Templin, J. L. (2004). *Generalized linear mixed proficiency models*. Unpublished doctoral dissertation, University of Illinois at Urbana-Champaign.

Teresi, J. A., Ocepek-Welikson, K., Kleinman, M., Cook, K. F., et al. (2007). Evaluation measurement equivalence using the item response theoryloglikelihood ratio (IRTLR) method to assess differential item functioning .*Quality Life Research, 16(*l1), 43-68.

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.

Wu, A. D., Li, Z., & Zumbo, B. D. (2007). Decoding the meaning of factorial invariance and updating the practice of multi-group confirmatory factor analysis: A demonstration with TIMSS data. *Practical Assessment, Research & Evaluation*, *12*(3), 1-26.

Zheng, Y., Chiu, C. Y., & Douglas, J. A. (2016). *NPCD: Nonparametric methods for cognitive diagnosis*. R package version 1.0-5.

Zumbo, B. D. (1999). *A handbook on the theory and methods of differential item functioning (DIF)*. National Defense Headquarters.

Application of Cognitive Diagnostic Models (CDM) Study of the Differential Item Functioning of TIMSS Eighth Grade Mathematics 2007 Among Girls and Boy

Asghar Minai*
Jalil Younesi**
Behrooz Kazemi Dana***

Abstract:
Attempts were made in this research to make use of the capabilities of Cognitive Diagnosis Models (CDMs) for precise and more detailed investigation of the Differential Item Functioning in the mathematics test of the eighth grade (third grade of guidance school) (T. 2007). For analyzing the data and responding to the research questions, the non-compensative model DINA, with the two parameters of guessing and slipping, was used. First, the data were fitted to the model, and the parameters were estimated for the two female and male groups. The Differential Item Functioning was then examined using the Wald test. The results demonstrated that of the 44 investigated questions, 26 were anchoring questions, and 18 (40%) were ones with DIF, where 12 out of the 18 questions with DIF had linear DIF, and 6 had nonlinear DIF. Of the 18 questions with DIF, 7 had Differential Item Functioning in favor of the female students who had achieved mastery over the attributes necessary for responding correctly to the questions, and 10 in favor of the male ones. One question here had Differential Item Functioning in favor of neither the group of females nor the group of males who had achieved mastery over the attributes necessary for responding correctly to the questions. Furthermore, of the 18 questions with DIF, 9 have Differential Item Functioning in favor of the female students who have not achieved mastery over the attributes necessary for responding correctly to the questions, and 7 in favor of the male ones

Keywords: Differential Item Functioning (DIF), Cognitive Diagnosis Models (CDMs) ,TIMSS test, Non-compensatory Model, DINA Model.

* Associate Professor of Assessment and Measurement (Psychometrics), Allameh Tabataba'i University, Tehran, Iran
** Associate Professor of Assessment and Measurement (Psychometrics), Allameh Tabataba'i University, Tehran, Iran
*** Ph.D. Student of Assessment and Measurement (Psychometrics), University of Tahran, Tehran, Iran (Responsible Author: dana_behrooz@yahoo.com)