شناسه دیجیتال (Doi): 10.22059/JFLR-201712-439 (R3)
شاپای چاپی (issn): ۲۵۸۸-٤۱۲۳
نشانی اینترنتی مجله: http//jflr.ut.ac.ir، پست الکترونیک: jflr@ut.ac.ir

کارکرد دوگانهٔ سوال، بر اساس رویکرد مبتنی بر نظریهٔ تشخیصی طبقه‌بندی در آزمون خواندن و درک مفاهیم کنکور منحصرا زبان

حمدالله راوند*
استادیار گروه زبان انگلیسی دانشگاه ولی عصر(عج) رفسنجان،
رفسنجان، ایران

(تاریخ دریافت: ۹٦/۰۹/۳۰، تاریخ تصویب: ۹۷/۰٦/۳۱، تاریخ چاپ: مهر ۱۳۹۷)

چکیده

کارکرد دوگانه سوال (DIF) وقتی اتفاق می‌افتد که آزمون‌شوندگان با سطح توانایی برابر از سازهٔ مورد سنجش، عملکرد متفاوتی در هر کدام از سوالات یک آزمون داشته باشند. مطالعهٔ حاضر به بررسی DIF در سوالات خواندن و درک مفاهیم آزمون منحصرا زبان ورودی دانشگاه‌های ایران می پردازد. علاوه بر این، در مطالعهٔ حاضر روش DIF مبتنی بر مدل‌های تشخیصی طبقه‌بندی و روش منتل هنزل پرداخته شده است. بدین منظور پاسخ ۱۰۰۰۰ نفر از داوطلبان آزمون مذکور، با استفاده از بسته‌های CDM وdifR در نرم‌افزار R استفاده گردید. نتایج نشان داد که در روش تشخیصی طبقه‌بندی، یک سوال و در روش سنتی منتل هنزل دو سوال، دارای DIF متوسط شناخته شدند که به نظر می‌رسد تهدیدی برای روایی سازهٔ آزمون مورد نظر محسوب نمی‌شود. همچنین می‌توان نتیجه گرفت هنگامی که نیمرخ خصیصه‌ای به‌عنوان متغیر جورکردنی استفاده می‌شود تعداد سوالات کمتری به‌عنوان DIF شناخته می‌شوند.

واژه‌های کلیدی: کارگرد دوگانه سوال، نیمرخ خصیصه‌ای، مدل‌های تشخیصی طبقه‌بندی، روش منتل هنزل.

* نویسنده مسئول :E-mail: ravand@vru.ac.ir

۱- کارکرد دوگانه سوال

آزمون منصف،[1] آزمونی است که برای تمامی آزمون‌شوندگان، فارغ از این‌که به چه گروهی وابسته‌اند سازهٔ یکسانی را می‌سنجد. درصورتی‌که سازه‌هایی غیر از سازهٔ مورد نظر، بر عملکرد افراد تاثیر گذار باشد و به گفته مسیک[2] (۱۹۹۶) **واریانس نامربوط**[3] در نمرات پدیدآید، و عملکرد افراد از گروه‌های مختلف (برای مثال زن در مقابل مرد) با نمرات کل یا توانایی یکسان در این سازه‌ها یا **ابعاد ثانوی**[4] متفاوت باشد، اصطلاحا می‌گویند **کارکرد دوگانهٔ سوال** (DIF[5]) اتفاق افتاده است. بنابراین براساس تعریف مرسوم، DIF وقتی اتفاق می‌افتد که سازهٔ یا سازه‌های دیگری، غیر از سازهٔ اصلی آزمون، بر عملکرد آزمون‌شوندگان در هر سوال تاثیر گذار باشد و آزمون‌شوندگان با سطح مالکیت (یا توانایی) برابر از سازهٔ اصلی عملکرد متفاوتی در سازه‌های ثانویه داشته باشد. در تعریف بالا یک نکته مهم مغفول مانده است: درصورتی‌که سازه‌های ثانویه مربوط به سازهٔ مورد سنجش باشد وحتی ممکن است در ساختار سوال دیده شده باشد، عملکرد متفاوت افراد همتراز، در این خصییه های ثانوی ،**کارکرد دوگانهٔ خصیصه** (DAF[6]) نامیده می‌شود. درحالی که DIF برای روایی سازهٔ تهدید محسوب می‌شود، DAF به روایی سازهٔ کمک می کند(واکر و برتواس[7]، ۲۰۰۱). وجود DIF امکان مقایسهٔ افراد متعلق به گروههای مختلف را از بین می برد. این مسئله از جنبه **تعمیم‌پذیری**[8] روایی سازهٔ (مسیک، ۱۹۹۶) حایز اهمیت است. معمولاً هر کدام از سازه‌های تاثیرگذار بر عملکرد یک **بعد**[9] را تشکیل می‌دهد. با توجه به تعریف بالا DIF می‌تواند در نتیجه **چند بعدیتی**[10] در یک سوال اتفاق بیفتد. معمولا سوالات آزمون‌ها طوری طراحی می‌شود که فقط یک بعد غالب را اندازه بگیرند و سوالاتی نیزکه ابعاد دیگری را اندازه می‌گیرند از منظر مدل اندازه‌گیری کلاسیک و همچنین نظریهٔ سوال پاسخ، تحت تاثیر واریانس

1. Fair test
2. Messick
3. Construct-irrelevant variance
4. Secondary dimensions
5. Differential item functioning
6. Differential attribute functioning
7. Walker & Beretvas
8. Generalizability aspect
9. Dimension
10. Multidimensionality

نامربوط و تهدیدی برای روایی سازهٔ محسوب می‌شود. اما معمولا سازه‌های مورد مطالعهٔ در تعلیم و تربیت، به‌طور عام و گسترهٔ آموزش زبان به‌طور خاص، چند بعدی‌اند. به‌عنوان مثال در حال حاضر باور غالب بر این است که سازهٔ خواندن و درک مطلب چند بعدی است و از **ریزمهارت**[1] های مختلفی تشکیل شده است (بقایی و راوند[2]، ۲۰۱۶). ضرورت‌های **سنجش اصیل**[3] حکم می‌کند که هر کدام از سوالات یک آزمون، به تبعیت از ساختار سازهٔ مورد سنجش، محدود به اندازه‌گیری محتوی و فرایند خاص نباشد (رامبورگ و ویلسون[4]، ۱۹۹۵). بنابراین در یک آزمون اصیل، پاسخگویی به هر کدام از سوالات، مستلزم تسلط به چندین خصیصه یا فرایند است. بر اساس آنچه گفته شد، می‌توان نتیجه گرفت آزمون‌های شناختی، معمولا چند بعدی‌اند. گاهی اوقات سوالات طوری طراحی می‌شوند که بیش از یک بعد را بسنجد، در نتیجه واریانس ایجاد شده در نمرات از سوی این ابعاد واریانس مربوط می‌باشد. با توجه به اینکه این ابعاد، وجوه مختلف یک سازهٔ است، یا ممکن است خود سازه‌های متفاوتی باشند که عمدا توسط طراحان سوالات در ساختار آزمون دیده شده‌اند، عملکرد متفاوت گروه‌های مختلف در این سوالات، طبیعتا نباید DIF تلقی شود. ولی رویکردهای فعلی به DIF (رویکرد نظریهٔ اندازه‌گیری کلاسیک ورویکرد نظریهٔ سوال پاسخ) قادر نیستند بین سازه‌ها یا خصیصه‌هایی که در نمرات واریانس مربوط و واریانس نامربوط ایجاد می کنند تمایز قایل شوند.

۲- فرایند تشخیص DIF

تحلیل DIF برای اولین بار در دههٔ ۱۹۶۰ در پاسخ به نگرانی عمومی دربارهٔ **سوگیری**[5] آزمون‌های شناختی علیه آزمون‌شوندگان اقلیت انجام شد (انگف[6]، ۱۹۹۳). برای انجام تجزیه و تحلیل DIF درآغاز آزمون‌شوندگان به گروه‌هایی با سطح نزدیک به برابر از سازهٔ مورد سنجش تقسیم می‌شوند. این معیار طبقه‌بندی یا سطح بندی **متغیر جورکردنی**[7] نامیده می‌شود. . برای

1. Subskill
2. Baghaei & Ravand
3. Authentic assessment
4. Ramborg & wilson
5. Biasedness
6. Angoff
7. Matching variable

بررسی DIF بایسته است که آزمون‌شوندگان بر اساس یک متغیر گروه‌بندی که همان متغیر جورکردنی حداقل به دو گروه تقسیم شوند. بیشتر، گروهی که دارای افراد بیشتری‌ست **گروه مرجع**[۱] و گروهی را که در اقلیت اند **گروه کانونی**[۲] می نامند آزمون‌شوندگانی که بنابرمتغیر جورکردنی در گروه n قرار می‌گیرند، چنانچه پرسش‌ها عادلانه باشند، باید همگان شانس برابری برای پاسخ به هر کدام از سوال ها را داشته باشند. به‌عبارت دیگر درجهٔ دشواری سوال برای آزمون‌شوندگان گروه n فارغ از اینکه به چه جنسی (زن یا مرد) دارند، نباید به‌طور معنی‌داری متفاوت باشد.

سه رویکرد متفاوت به تجزیه وتحلیل DIF وجود دارد. تفاوت این روش‌ها در نوع متغیر جورکردنی است که بر اساس آن آزمون‌شوندگان به n گروه که هر کدام سطح متفاوتی از متغیر جورکردنی را دارند تقسیم می‌شوند.

۱. رویکرد نظریهٔ اندازه‌گیری کلاسیک[۳]

۲. رویکرد نظریهٔ سوال پاسخ[۴]

۳. رویکرد مدل‌های تشخیصی طبقه‌بندی[۵]

در رویکرد مبتنی بر نظریهٔ کلاسیک، از نمرهٔ کل برای جورکردن آزمون‌شوندگان استفاده می‌شود، ولی در رویکرد مبتنی بر نظریهٔ سوال پاسخ، از توانایی برآورد شده برای جورکردن استفاده می‌شود. بر اساس دو رویکرد بالا، درصورتی‌که پرسش‌ها طوری طراحی شده باشند که همزمان چند بعد را اندازه‌گیری کنند و همچنین، درصورتی‌که عملکرد گروه‌های **جور شده/همتراز شده**[۶] در هر کدام از سوالات متفاوت باشد، سوال مورد نظر به‌عنوان موردی از DIF تلقی می‌شود. اما در رویکرد مدل‌های تشخیصی طبقه‌بندی ابتدا زیرمهارت‌ها یا ابعادی که توسط آزمون سنجش می‌شوند، مشخص می‌شود. دراین رویکرد ابعاد مشخص شده، **ابعاد مربوط**[۷] وابعاد یا زیرمهارتهای احتمالی دیگری که در پاسخگویی به سوالات آزمون دخیل

1. Reference group
2. Focal group
3. Classical test theory
4. Item response theory
5. Diagnostic classification models
6. Matched groups
7. Construct-relevant dimensions

باشند و موجب ایجاد واریانس نامربوط می‌شوند، **ابعاد نامربوط**[1] تلقی خواهند شد. با توجه به تعریفی که در بالا ارایه شد، عملکرد متفاوت گروه‌های مرجع و کانونی در هر کدام از ابعاد مربوط نباید موجب DIF شود. اما در عمل، در دو رویکرد کلاسیک و سوال پاسخ عملکرد متفاوت گروه‌های مختلف، باعث گزارش سوال به‌عنوان سوال دارای DIF می‌شود. این مسئله نشان می‌دهد که تعریفی که در ابتدا از DIF ارائه شد ناقص است زیرا در آن تمامی سازه‌ها یا ابعاد ثانوی به‌عنوان **سازهٔ مخالف**[2] نامربوط در نظر گرفته می‌شوند. به همین دلیل محققان بین DIF و DAF تمایز قائل شده‌اند (میلوسکی و بارون[3]، ۲۰۰۲) در واقع DAF در نتیجه کارکرد متفاوت صفت یا خصیصه‌ای اتفاق می‌افتد که در ساختار سوال دیده شده است. بنابراین DAF بیانگر نقاط قوت و ضعف آزمون‌شوندگان در از خصیصه‌های مشخص شده می‌باشد و به هیچ وجه نباید به‌عنوان **سوگیری** سوال تلقی شود. تعریف DIF نیاز به بازنگری دارد، به‌طوری که DIF در هنگامی که روی می‌دهد که سازه‌های ثانوی سازهٔ مخالف باشند. درصورتی‌که سازه‌های ثانویه سازهٔ موافق باشند DAFرخ می‌دهد. بایسته یاداوری است در مقاله حاضر، از اصطلاح DIF برای اشاره به مفهوم عملکرد متفاوت سوال که نشان از سوگیری سوال دارد، استفاده می‌شود ولی از اصطلاح DAF برای اشاره به پدید ه ای که نشانگر نقاط ضعف یا قوت آزمون‌شوندگان است به کار می‌رود.

۳- رویکرد مبتنی بر مدل‌های تشخیصی طبقه‌بندی

در رویکرد مبتنی بر مدل‌های تشخیصی طبقه‌بندی، ابتدا خصیصه‌ها یا سازه‌های مورد سنجش توسط هر آزمون به‌طور کلی و هر سوال به‌طور خاص طی یک فرآیند کیفی (و گاهی کمی) مشخص می‌شود. سپس آزمون‌شوندگان بر اساس تسلط یا عدم تسلط، به هر کدام از خصیصه‌ها به طبقات مختلفی دسته‌بندی می‌شوند. به‌طور معمول تعداد این طبقات، تابعی از تعداد خصیصه‌ها است. به‌عنوان مثال، اگر عملکرد موفق در یک آزمون، منوط به تسلط به سه خصیصه باشد، تعداد طبقاتی که افراد در آن‌ها طبقه‌بندی می‌شوند $۲^۳=۸$ است، اما اگر درصورتی‌که سوالات منصف باشند، انتظار می‌رود، احتمال جواب درست به آن سوالات برای افراد یک طبقهٔ خاص با هم برابر باشد، فارغ از این که این افراد از نظر جنسیت، ملیت، رشتهٔ

1. Construct-irrelevant dimensions
2. Irrelevant construct
3. Milewski & Baron

تحصیلی، و غیره به چه گروهی تعلق دارند. به‌عبارت دیگر در رویکرد مبتنی بر مدل‌های تشخیصی طبقه‌بندی متغیر جورکردنی **نیمرخ خصیصه‌ای**[1] آزمون‌شوندگان است. نیمرخ خصیصه‌ای از تعدادی صفر و یک به تعداد خصیصه‌های مورد سنجش توسط یک آزمون تشکیل شده‌اند. مثلاً نیمرخ $\alpha = [1,1,0]$ مربوط به یکی از آزمون‌شوندگان، گویای این است که فرد مورد نظر مربوط به طبقه ای ست که در آن افراد به خصیصه اول ودوم مسلط اند ولی به خصیصهٔ سوم تسلط ندارند. در رویکرد اول و دوم، یعنی رویکرد مبتنی بر نظریهٔ کلاسیک و سوال پاسخ خصیصه‌های احتمالا دو تا از خصیصه‌ها در متغیر جورکردنی یا در نظر گرفته نمی‌شوند ویا به‌عنوان بخشی از یک **نمرهٔ مرکب**[2] که حاصل میانگین غیر وزنی نمرهٔ همه سوالات است در نظر گرفته می‌شوند. در این دور رویکرد درصورتی‌که عملکرد آزمون‌شوندگان جورشده ولی متعلق به گروه‌های مختلف در هر کدام از این دو خصیصه متفاوت باشند DIF پدید می آید، اما در رویکرد مبتنی بر روش تشخیصی طبقه‌بندی ابعاد یا خصیصه‌های مختلفی در متغیر جورکردنی لحاظ شده است، عملکرد متفاوت گروه‌های مختلف در این خصیصه‌ها تفاوت های **مربوط به سازهٔ**[3] تلقی می‌شود. بنابراین انتظار این است ،با توجه به اینکه طیف وسیع‌تری از خصیصه‌های دخیل در پاسخگویی به سوالات یک آزمون در نمره جورکردنی دیده می‌شوند، در روش مبتنی بر مدل‌های تشخیصی طبقه‌بندی نسبت به دو رویکرد دیگر سوالات کمتری به‌عنوان سوالات دارای DIF شناخته شوند.

با توجه به آنچه که گفته شد، مطالعهٔ حاضر با استفاده از رویکرد تشخیصی طبقه‌بندی، به بررسی کارکرد متفاوت سوالات خواندن و درک مفاهیم کنکور ورودی دوره کارشناسی رشته‌های زبان انگلیسی در دانشگاه‌های دولتی (که به اختصار کنکور منحصرا به زبان نامیده می‌شود) برای داوطلبان زن و مرد می پردازد. به‌طور خاص، مطالعهٔ حاضر از روش DIF مبتنی بر مدل طبقه‌بندی شناختی دینا ارائه شده توسط دلاتوری و ناناداکومار[4] (۲۰۱٤) استفاده می‌کند. همچنین مطالعهٔ حاضر به مقایسه DIF مبتنی بر مدل‌های شناختی طبقه‌بندی و DIF منتل هنزل مبتنی بر نظریهٔ کلاسیک آزمون می پردازد. در روش منتل هنزل مورد استفاده در مطالعهٔ حاضر از نمرهٔ کل به‌عنوان متغیر جورکردنی استفاده می گردد.

1. Attribute profile
2. Composite score
3. Construct-relevant differences
4. Hou, de la Torre & Nandakumar

مطالعهٔ حاضر از چند جهت می‌تواند دارای اهمیت باشد: (۱) به لحاظ عملی می‌تواند در کل، در زمینهٔ روایی سازهٔ کنکور زبان و به‌طور خاص وجود احتمالی DIF در سوالات این آزمون که یکی از آزمون‌های سرنوشت ساز برای پذیرش متقاضیان در رشته‌های کارشناسی زبان انگلیسی دانشگاه‌های دولتی است، روشنگری نماید. با توجه به اینکه این آزمون برای اهداف گزینش دانشجو صورت می‌گیرد که مستلزم مقایسهٔ عملکرد متقاضیان با یکدیگر است، وجود DIF مقایسه پذیری متقاضیان را محدود و در نتیجه، روایی سازهٔ آزمون را تهدید می کند. حصول اطمینان از روایی سازهٔ این چنین آزمونی از جنبهٔ تعمیم‌پذیری، اهمیت بسیار دارد. درصورت برقرار بودن جنبهٔ تعمیم‌پذیری روایی سازهٔ، می‌توان اطمینان حاصل کرد که آزمون مورد نظر برای همگی متقاضیان دارای توانایی تقریبا برابر، فارغ از اینکه به چه گروهی وابسته اند سازهٔ یا خصیصه یکسانی را می‌سنجد. (۲) به لحاظ روش‌شناختی، مطالعهٔ حاضر به مقایسهٔ روش DIF منتل هنزل[1] مبتنی بر نظریهٔ کلاسیک آزمون و روش مبتنی بر مدل‌های تشخیصی طبقه‌بندی می‌پردازد انتظار می‌رود در روش مبتنی بر مدل‌های شناختی طبقه‌بندی سوالات کمتری به‌عنوان سوال دارای کمر گزارش شود، زیرا در این روش متغیر جورکردنی طیفی از خصیصه‌ها را که توسط آزمون اندازه‌گیری می‌شود در بر میگیرد.

٤- مطالعات DIF در آزمون‌های ورودی دانشگاه‌های ایران

درحالیکه مطالعات DIF زیادی دربارهٔ آزمون‌های سرنوشت ساز مطرح دنیا از قبیل آزمون تافل (بیلی[2]، ۱۹۹۹؛ وال و هراک[3]، ۲۰۰۸ ؛ لی، برلاند، و موراکی[4] ۲۰۰۵؛ ریان و بکمن[5] ۱۹۹۲؛ آلدرمن و هلند[6] ۱۹۸۱؛ برلاند، لی، نجاریان، و موراکی[7]، ۲۰۰٤؛ لیو،شدل، مالوی،و کنگ[8]، ۲۰۱۱) و **آزمون استعدادتحصیلی**[9] (بریجمان و وندلر[10] ۱۹۹۱؛ کانارک[11] ۱۹۸۸ ؛ کرلی

1. Mantel-Haenszel
2. Bailey
3. Wall & Horák
4. Lee, Breland & Muraki
5. Ryan & Bachman
6. Alderman & Holland
7. Breland, Lee, Najarian & Muraki
8. Liu, Schedl, Malloy & Kong
9. Scholastic aptitude test
10. Bridgeman & Wendler
11. Kanarek

و اشمیت[1] ۱۹۹۳؛ کارلتون و هاریس[2]، ۱۹۹۲؛ دورانز و کولیک[3] ۱۹۸۶؛ اشمیت و دوران[4] ۱۹۹۰؛ لورنس، کرلی، مک هیل[5] ۱۹۸۸؛ لورنس کرلی[6] ۱۹۸۹) وحتی آزمون‌های داخلی مهمی همچون**آزمون بسندگی زبان انگلیسی دانشگاه تهران**[7] (کرمی[8]، ۲۰۱۱؛ علوی، رضایی، و امیریان[9]، ۲۰۱۲؛ رضایی و شعبانی[10]، ۲۰۱۰؛ صالحی و طیبی[11] ۲۰۱۱؛ فیدالگو، علوی، امیریان[12]، ۲۰۱٤؛ امیریان، علوی، و فیدالگو[13]، ۲۰۱٤) انجام شده است، بررسی آزمون ورودی دانشگاه‌ها در ایران توجه کافی را به خود جلب نکرده است. آزمون‌های ورودی دانشگاه‌های ایران که هر ساله خیل عظیمی از داوطلبان ورود به مقاطع کارشناسی، کارشناسی ارشد، و دکتری دانشگاه‌ها را گزینش می کنند، به لحاظ ویژگی‌های روانسنجی به‌طور کل و مطالعات DIF به‌طور خاص به‌طور بسیار پراکنده مورد بررسی قرار گرفته‌اند. براتی، کتابی و احمدی[14] (۲۰۰٦) به بررسی DIF در آزمون ورودی دوره‌های کارشناسی دانشگاهها پرداختند. در مطالعهٔ دیگری راوند و فیروزی[15] (۲۰۱٦) با استفاده از مدل راش، به بررسی روایی سازهٔ آزمون ورودی دوره‌های کارشناسی ارشد رشته‌های زبان انگلیسی پرداختند که بخشی از این مطالعهٔ شامل بررسی جنبه تعمیم‌پذیری روایی سازهٔ آزمون مورد نظر از طریق تجزیه و تحلیل DIF می‌شود. همچنین در یک مطالعهٔ دیگر احمدی و دارابی[16] (۲۰۱٦) DIF در سوالات زبان انگلیسی ورودی دورهٔ دکتری رشته‌های زبان انگلیسی دانشگاه‌ها را بررسی کردند. بنابراین بررسی جنبه های مختلف ویژگی‌های روانسنجی آزمون‌های ورودی

1. Curley & Schmitt
2. Carlton &Harris
3. Dorans & Kulick
4. Schmitt & Dorans
5. Lawrence, Curley & McHale
6. Lawrence & Curley
7. University of Tehran's English Proficiency Test
8. Karami
9. Alavi, Rezaee & Amirian
10. Rezaee & Shabani
11. Salehi & Tayebi
12. Fidalgo, Alavi & Amirian
13. Amirian, Alavi & Fidalgo
14. Barati, Ketabi & Ahmadi
15. Ravand & Firoozi
16. Ahmadi & Darabi

دانشگاه‌های ایران به مطالعات بیشتری نیاز دارد، که مطالعهٔ حاضر کوششی برای رفع این نیاز است.

۵- مدل‌های تشخیصی طبقه‌بندی

مدل‌های تشخیصی طبقه‌بندی بر خلاف مدل کلاسیک آزمون و نظریهٔ سوال پاسخ که آزمون‌شوندگان را به‌طور معمول در امتداد یک سازهٔ تک بعدی رتبه‌بندی می کنند، آزمون‌شوندگان را بر اساس تسلط یا عدم تسلط در سازه‌های چند بعدی، طبقه‌بندی می‌کنند. نمرات کل به‌دست آمده از نظریهٔ کلاسیک و یا برآوردهای توانایی ارایه شده توسط نظریهٔ سوال پاسخ می‌توانند برای اهداف گزینش افراد برای حضور در یک دورهٔ خاص یا ادامه تحصیل در موسسات یا دانشگاه‌ها مفید واقع شوند. طبقه‌بندی افراد بر اساس نتایج به‌دست آمده از مدل‌های تشخیصی طبقه‌بندی می‌توانند در جهت فراهم سازی بازخورد ریزساختار برای انجام مداخله‌های دقیق‌ترو در نتیجه آموزش و یادگیری بهتر مفید واقع شود. به‌عبارت دیگر، می‌توان گفت مدل کلاسیک آزمون و نظریهٔ سوال پاسخ **سنجش یادگیری**[1] انجام می‌دهند ولی مدل‌های تشخیصی طبقه‌بندی **سنجش در جهت یادگیری**[2] انجام می دهند (یانگ[3] ۲۰۰۵). درصورت عملکرد ناموفق آزمون‌شوندگان در یک آزمون خاص نظریهٔ های کلاسیک و سوال پاسخ دربارهٔ علت عدم موفقیت اطلاعاتی در اختیار نمی‌گذارند، اما مدل‌های طبقه‌بندی با توجه به اینکه وضعیت تسلط یا عدم تسلط هر یک از آزمون‌شوندگان را در هر کدام از خصیصه‌های اندازه‌گیری شده توسط آزمون مشخص می کنند، می‌توانند درصورت عدم موفقیت در هر کدام از سوالات دلایل عدم موفقیت را ارائه دهند و در نتیجه می‌توان بر اساس نتایج، آموزش هدفمند و دقیق‌تربرای برطرف کردن نقاط ضعف آزمون‌شوندگان ارائه کرد. در سال‌های اخیر، مدل‌های تشخیصی طبقه‌بندی توجه زیادی به خود جلب کرده اند. اما پژوهش های انجام شده دربارهٔ این مدل‌ها بیشتر جنبهٔ روش‌شناختی داشته و کمتر به جنبهٔ کاربردی آن‌ها پرداخته شده است. مدل‌های تشخیصی طبقه‌بندی را بر اساس اینکه خصیصه‌های مورد نیاز برای یک آزمون چگونه با هم تعامل می کنند تا منجر به جواب درست

1. Assessment of learning
2. Assessment for learning
3. Jang

برای سوالات شود، به سه گروه تقسیم می‌شوند(راوند، ۲۰۱۶): (۱) **مدل‌های غیرجبرانی**[1]، (۲) **مدل‌های جبرانی**[2]، و (۳) **مدل‌های افزایشی یا جمع پذیر**[3]. در مدل‌های جبرانی تسلط به یک خصیصه می‌تواند عدم تسلط به خصیصه‌های دیگر را جبران کند. برای یک مثال اگر برای جواب درست به یک سوال سه خصیصه مورد نیاز باشد، تسلط به هرکدام از خصیصه‌ها می‌تواند عدم تسلط به یک یا دو خصیصهٔ دیگر را جبران کنند. به‌عبارت دیگر در مدل‌های جبرانی فرقی نمی‌کند که فرد به خصیصهٔ اول، دوم، سوم، یا هر سه خصیصه مسلط باشد. در هر کدام از چهار حالت گفته شده احتمال پاسخ صحیح به سوال برابر و بالا وجوددارد. در عوض در مدل‌های غیر جبرانی، پاسخ صحیح به سوال، نیازمند تسلط به همه خصیصه‌های مورد نیاز برای آن سوال می‌باشد. عدم تسلط به یکی، مثل عدم تسلط به همگی خصیصه‌ها است. در مدل‌های افزایشی هر کدام از خصیصه‌های مورد نیاز به‌طور مستقل احتمال پاسخ صحیح را افزایش می‌دهد و تسلط به خصیصه‌های بیشتر احتمال پاسخ درست را بیشتر می کند. تفاوت مدل‌های جبرانی و افزایشی در این است که در مدل‌های جبرانی خصیصه‌ها به‌عنوان جایگزین همدیگر تلقی می‌شوندوفرض براین است که این خصیصه‌ها راه‌های جایگزین برای رسیدن به جواب یکسان می باشند. ولی در مدل‌های افزایشی، خصیصه‌های مورد نیاز، برای هر سوال جایگزین یکدیگر نیستند. در مطالعهٔ حاضر از مدل دینا (یانکر و سیتسما[4]، ۲۰۰۱) که یک مدل غیر جبرانی می‌باشد جهت مطالعهٔ DIF استفاده میگردد.

۵-۱- مدل دینا

مدل دینا یک مدل تشخیصی طبقه‌بندی غیر جبرانی است. دینا یکی از ساده‌ترین مدل‌های تشخیصی طبقه‌بندی است که برای هر سوال آزمون‌شوندگان را به دو گروه تقسیم می کند: (۱) گروهی که به همهٔ خصیصه‌های موردنیاز برای پاسخگویی به سوال مورد نظر، مسلط هستند و در نتیجه، انتظار می‌رود به آن سوال پاسخ صحیح دهند و (۲) گروهی که کمینه بر یکی از خصیصه‌های مورد نیاز مسلط نیستند و بنابراین انتظار می‌رود که سوال را اشتباه جواب دهند بر اساس مدل دینا و از **منظر قطعیت**[5] افرادی که همهٔ خصیصه‌های مورد نیاز یک سوال را دارند،

1. Non-compensatory models
2. Compensatory models
3. Additive models
4. Junker & Sijtsma
5. Deterministically

باید بتوانند سوال را درست جواب دهند، اما از **منظر احتمال**[1] ممکن است که آزمون‌شوندگان دچار **لغزش**[2] شوند و سوال را اشتباه جواب دهند. همچنین در مدل دینا از نگاه قطعیت کسی که کمینه یکی از خصیصه‌های مورد نیاز برای یک سوال را نداشته باشد، باید به آن سوال جواب اشتباه بدهد، ولی از نگاه احتمال ممکن است فرد مورد نظر با استفاده از **حدس زدن**[3] به سوال مورد نظر درست جواب دهد. بنابراین مدل دینا برای هر سوال دو پارامتر محاسبه می کند: یکی پارامتر لغزش و دیگری پارامتر حدس زدن. بایسته یاداوری است که تفاوت بین پارامترهای لغزش و حدس زدن، نشان دهندهٔ درجهٔ تمیز سوال است که به‌عنوان شاخص مهم کیفیت سوال مورد استفاده قرار می‌گیرد.

تابع پاسخ سوال بر اساس مدل دینا را می‌توان با رابطه ریاضی زیر نمایش داد

$$P(X_{ij}=1\,|\,\xi_{ij})=\left(1-s_j\right)^{\xi_{ij}}\, g_j^{\,1-\xi_{ij}}$$

در این رابطه $P(X_{ij}=1\,|\,\xi_{ij})$ احتمال جواب درست توسط فرد jباوضعیت تسلط ξ به سوال i، s_j احتمال لغزش، و g_j احتمال حدس زدن است. درصورتی‌که فردبه همه خصیصه‌های مورد نیاز برای یک سوال مسلط باشد ($\xi_{ij}=1$)، احتمال پاسخ صحیح به آن سوال برابراست با احتمال عدم لغزش ($1-s_j$):

$$P(X_{ij}=1\,|\,\xi_{ij}=1)=\left(1-s_j\right)^{1} g_j^{\,1-1}=1-s_j$$

اما هنگامی که فرد حداقل یکی از خصیصه‌های مورد نیاز برای سوال را ندارد($\xi_{ij}=0$)، احتمال پاسخ درست به سوال برابر است با حدس زدن (g_j):

$$P(X_{ij}=1\,|\,\xi_{ij}=0)=\left(1-s_j\right)^{0} g_j^{\,1-0}=g_j$$

۵-۲-۱- DIF مبتنی بر مدل دینا

هو، د ل توره، و ناندا کومار[4] (۲۰۱٤) ضمیمه ای بر مدل دینا معرفی کردند که قادر است سوالات دارای DIF را در چارچوب مدل‌های تشخیصی طبقه‌بندی مشخص کند. تا به امروز

1. Stochastically
2. Slipping
3. Guessing
4. Hou, de la Torre & Nandakumar

مطالعات محدودی درباره DIF در چارچوب مطالعات تشخیصی طبقه‌بندی انجام شده است (لی[1]، ۲۰۰۸، میلوسکی و بارون، ۲۰۰۲، ژنگ[2]، ۲۰۰۶). هدف بعضی از این مطالعات (میلوسکی و بارون، ۲۰۰۲) یافتن سوالات دارای سوگیری نبوده است. میلوسکی و بارون (۲۰۰۲) درصدد بودند تا از طریق تجزیه وبررسی DIF نقاط ضعف و قوت گروهی از دانش آموزان یک مدرسه را کشف کنند. در این مطالعهٔ، از نمرهٔ کل به‌عنوان متغیر جورکردنی استفاده شده است. ژنگ (۲۰۰۶) به مقایسه عملکرد دو متغیر جورکردنی نمره کل و نیمرخ خصیصه‌ای با استفاده از روش سنتی منتل هنزل، روش سیب تست (شیلی و استوت[3]،۱۹۹۳) پرداخت. نتایج مطالعهٔ نشان داد که در هر دو روش منتل هنزل و روش سیب تست نیمرخ عملکرد خصیصه‌ای همتراز و حتی بهتر از متغیر جورکردنی نمرهٔ کل می‌باشد. لی (۲۰۰۸) با گسترش مطالعهٔ میلوسکی و بارون (۲۰۰۲) و با استفاده از ضمیمه **مرتبه بالاتر مدل دینا**[4] به بررسی همزمان DIF و DAF پرداخت. مطالعات مذکور، دارای محدودیت‌های زیراست: (۱) برآورد نیمرخ خصیصه‌ای افراد و پارامترهای سوالات ممکن است، تحت تاثیر سوالات دارای DIF آلوده شده باشند. در این روش‌ها ابتدا باید **پاکیزه سازی**[5] به‌عنوان یکی از مراحل انجام DIF در دستور کار قرار گیرد. (۲) روش‌های ارائه شده دراین مطالعات می‌توانند فقط به بررسی **DIF همسان**[6] بپردازد.

مدل DIF دینا محور ارائه شده توسط هو و همکاران (۲۰۱٤) با داشتن ویژگی‌های زیر برای رفع محدودیت‌های فوق برآمده است (۱) در مدل DIF دینا مدرج سازی سوالات برای گروه‌های مرجع و کانونی به‌طور مجزا انجام می‌گیرد بنابراین احتمال آلوده شدن نیمرخ خصیصه‌ای و پارامترهای سوالات دراثر وجود سوالات DIF از بین می‌رود. (۲) این روش قادر به کشف همزمان DIF همسان و غیر همسان است.

DIF در چارچوب مدل‌های تشخیصی طبقه‌بندی می‌تواند از طریق رابطه ریاضی زیر نمایش داده شود:

$$\Delta_{j\alpha_l} = P(\mathrm{X}_j = 1 \mid \alpha_l)_F - P(\mathrm{X}_j = 1 \mid \alpha_l)_R$$

1. Li
2. Zhang
3. Shealy & Stout
4. Higher-order DINA
5. Purification
6. Uniform DIF

$\Delta_{j\alpha_l} < 0$ نشان دهنده DIF در سوال jبرای آزمون‌شوندگان طبقه‌بندی شده در نیمرخ خصیصه‌ای α_l ، $P(X_j = 1|\alpha_l)_F$ احتمال جواب درست به سوال j برای آزمون‌شوندگان گروه کانونی و $P(X_j = 1|\alpha_l)_R$ احتمال جواب درست به سوال j برای آزمون‌شوندگان گروه مرجع می‌باشد.

اگر $\Delta_{j\alpha_l} > 0$ DIF به نفع گروه کانونی و اگر $\Delta_{j\alpha_l} < 0$ DIF به نفع گروه مرجع می‌باشد.

برخلاف فرایند تشخیص DIF در روش‌های سنتی که دشواری سوال را در گروه‌های مرجع وکانونی مقایسه می کنند، DIF مبتنی بر دینا، پارامترهای لغزش و حدس زدن را در گروه‌های مرجع و کانونی مقایسه می‌کند . در این روش با استفاده از **آزمون والد**[1] فرض صفر زیر آزمون می‌شود:

$$H_0 = \begin{matrix} s_{Fj} - s_{Rj} = 0 \\ g_{Fj} - g_{Rj} = 0 \end{matrix}$$

s_{Fj} و s_{Rj} به ترتیب لغزش در سوال jبرای گروه‌های کانونی ومرجع و g_{Fj} و g_{Rj} به ترتیب حدس زدن در سوال jبرای گروه‌های کانونی ومرجع‌اند. درصورتی‌که هر کدام از دو پارامتر لغزش و حدس زدن تفاوت معناداری در گروه‌های کانونی و مرجع داشته باشند، سوال مورد نظر به‌عنوان سوال دارای DIF شناخته می‌شود.

به‌طور خلاصه فرایند انجام DIF مبتنی بر مدل دینا به‌شرح ذیل است:

۱) با استفاده از مدل دینا نیمرخ خصیصه‌ای آزمون‌شوندگان برآورد می‌شود.

۲) آزمون‌شوندگان دارای نیمرخ خصیصه‌ای یکسان در طبقات یکسان قرار می‌گیرند.

۳) در هرطبقه پارامترهای لغزش و حدس زدن در هر سوال به‌طور جداگانه برای گروه‌های مرجع و کانونی محاسبه می‌شوند.

٤) تفاوت پارامترهای لغزش و حدس زدن بین گروه‌های مرجع و کانونی به لحاظ معنی‌داری بررسی می‌شود.

۵) درصورتی‌که هر کدام (یا هر دو) از پارامترهای مذکور، یعنی لغزش وحدس زدن، بین دو گروه کانونی و مرجع تفاوت معنی‌داری داشته باشند، سوال مورد نظر به‌عنوان سوال دارای DIF شناخته می‌شود.

1. Wald Test

مطالعهٔ حاضر در راستای پاسخ به دو سوال زیر گام برمی دارد:

۱) آیا سوالات خواندن و درک مفاهیم بخش تخصصی کنکور منحصر به زبان، دارای **DIF جنسیت**[1]اند؟

۲) آیا تعداد سوالات گزارش شده به‌عنوان DIF جنسیت در دو روش منتل هنزل (هنگامی که نمرهٔ کل به‌عنوان متغیر جورکردنی استفاده می‌شود) و روش مبتنی بر مدل‌های شناختی طبقه‌بندی یکسان است؟

۵-۲-۲- روش

در مطالعهٔ حاضر پاسخ‌های ۱۰۰۰۰ نفر از آزمون‌شوندگان به سوالات خواندن و درک مفاهیم آزمون کنکور منحصر به زبان، در سال ۱۳۹۲ مورد استفاده قرار گرفت. حدود ۷۰٪ نمونهٔ مورد مطالعهٔ، زن بودند. آزمون کنکور ورودی دانشگاه‌ها هرساله توسط سازمان سنجش برگزار می‌شود. آزمون کنکور منحصر به زبان در دو بخش عمومی و تخصصی برگزار می‌شود آزمون تخصصی شامل ۷۰ سال به‌شرح زیر است دستور زبان(۱۰ سوال)، واژگان (۱۵ سوال)، ساختار جمله (۵ سوال)، کاربردهای زبان (۱۰ سوال)، سوالات کلوز (۱۰ سوال)، و سوالات خواندن و درک مفاهیم (۲۰ سوال). تمامی سوالات به صورت چهار گزینه ای بوده و برای پاسخگویی به سوالات این بخش، آزمون‌شوندگان ۷۰ دقیقه زمان داشتند. در مطالعهٔ حاضر پاسخ ۱۰ هزار نفر از داوطلبان کنکور منحصر به زبان به ۲۰ سوال خواندن و درک مفاهیم بخش تخصصی آزمون مذکور مورد مطالعهٔ قرار گرفت.

٦- ماتریس کیو

ماتریس کیو[2] (تاتسوکا[3]، 1983) یکی از اجزای مهم یک مطالعهٔ، تشخیص طبقه‌بندی است. ماتریس کیو دارای تعدادی سطر، به اندازهٔ تعداد سوالات و تعدادی ستون، به اندازهٔ تعداد خصیصه‌های مورد نیاز برای پاسخگویی به سوالات هر آزمون می‌باشد. در واقع ماتریس کیو مشخص می‌کند کدام خصیصه‌ها برای پاسخگویی به هر کدام از سوالات یک آزمون مورد

1. Gender DIF
2. Q-matrix
3. Tatsuoka

نیاز می‌باشد. درمطالعهٔ حاضر از ماتریس کیو تدوین شده توسط همتی، بقایی، و بمانی[1] (۲۰۱۶) استفاده شد. در مطالعهٔ مذکور، همتی و همکاران یک مطالعهٔ تشخیصی طبقه‌بندی روی آزمون خواندن و درک مفاهیم استفاده شده در مطالعهٔ حاضر انجام دادند. با استفاده از **نظرکارشناسان**[2] پنج خصیصه برای جواب دادن به سوالات این آزمون لازم تشخیص داده شد: (۱) **استنباط**[3]، (۲) **استخراج اطلاعات صریح**[4]، (۳) **فهم معنی کلمات با استفاده از متن**[5]، (٤) **تشخیص مرجع ضمایر**[6]، و (۵) **ارزیابی گزینه‌ها**[7]. جدول شمارهٔ یک بخشی از ماتریس کیو مورد استفاده در مطالعهٔ حاضر نشان می‌دهد.

جدول شمارهٔ ۱. بخشی از ماتریس کیوی مورد استفاده (برگرفته از همتی و همکاران، ۲۰۱۶)

Item	A1	A2	A3	A4	A5
1	1	0	0	0	1
2	0	1	0	0	1
3	0	0	1	0	0
4	1	1	0	0	0
5	1	0	0	0	0

همانطور جدول شمارهٔ ۱ نشان می‌دهد، سوال شمارهٔ یک، به‌عنوان مثال، خصیصه‌های شمارهٔ یک و پنج را می‌سنجد.

۷- تجزیه و تحلیل

در مطالعهٔ حاضر از بسته آماری CDM (روبیچ، کایفر، جورج، و اونلو[8]، ۲۰۱۷) در نرم افزار رایگان R برای تجزیه و بررسی داده‌ها استفاده شد. بسته CDM از **روش براورد حداکثر**

1. Hemmati, Baghaei & Bemani
2. Expert judgement
3. Inferencing
4. Extracting explicit information
5. Understanding word meaning from context
6. Identifying pronominal references
7. Evaluating response options
8. Robitzsch, Kiefer, George & Uenlue

درست نمایی[1] با استفاده از الگوریتم EM برای برازش مدل استفاده می‌کند. لازم به ذکر است در حال حاضر فقط دو بستهٔ آماری از جمله CDM و GDINA (ما و د ل توره۲، ۲۰۱۷) قادر به انجام مطالعات DIF بر مبنای مدل‌های تشخیصی طبقه‌بندی می‌بینید. همچنین برای محاسبهٔ DIF بر اساس رویکرد سنتی از بستهٔ difR (ماگیس، بلند، رایشه، ۲۰۱۶) از مجموعه بسته‌های نرم افزار R استفاده شد.

۸- نتایج، بحث، و نتیجه گیری

نتایج تجزیه تحلیل DIF در جدول شمارهٔ دو ارائه شده است. ستون چهارم در این جدول نتایج آزمون معنی‌داری تفاوت بین پارامترهای لغزش و حدس زدن را برای گروه‌های زن و مرد نشان می‌دهد با توجه به نتایج مندرج در این ستون در هفت عدد از سوالات پارامتر لغزش و/ یا حدس زدن بین گروه‌های زن و مرد تفاوت معنی‌داری نشان می‌دهد ($p < .05$). با توجه به حجم بالای نمونه و احتمال معنی‌دار شدن تفاوت های ناچیز، شاخص اندازه تأثیر برای تفاوت های بین پارامترها محاسبه شد، که در ستون آخر جدول آورده شده است. جودوین و جیرل[3] (۲۰۰۱) اندازه تاثیر ۰۵۹. مرز بین DIF ناچیز و ملایم و ۰۸۸. را مرز بین DIF ملایم و شدید می‌داند. با توجه به معیار پیشنهاد شده توسط جودوین و جیرل، از بین ۷ سوال که پارامترهای آن بین زنان و مردان تفاوت معنی‌داری دارد، تنها یک سوال یعنی سوال هفتم دارای DIF ملایم است.

بنابراین، سوال اول تحقیق را می‌توان این طور پاسخ داد که اگرچه بر اساس آزمون معنی‌داری هفت سال دارای DIF جنسیت شناخته شد، ولی اندازهٔ تاثیر نشان داد که فقط یکی از این سوالات دارای DIF آن هم از مقولهٔ ملایم بود. با توجه به اینکه تنها یک سوال و آن هم دارای DIF متوسط بود، می‌توان نتیجه گرفت آزمون کنکور منحصر به زبان از جنبهٔ تعمیم‌پذیری روایی سازهٔ به نظر می‌رسد، مشکلی نداشته باشد. با توجه به اینکه آزمون مذکور جهت گزینش دانشجو ودر نتیجه مقایسه عملکرد آنان به کار می‌رود، اطمینان از آزمون مورد نظر از جنبه تعمیم‌پذیری بیانگر این است که پارامترهای لغزش و حدس زدن که از شاخص‌های کیفیت سوال می باشند برای گروه‌های زن و مرد تقریبا مشابه‌اند.

1. Maximum likelihood estimation
2. Ma & de la Torre
3. Jodoin & Gierl,

جدول شماره ۲. نتایج DIF بر اساس مدل دینا

Item		X2	df	p	p.holm	UA
1		0.97	2	0.61	1	0.01
2		0.03	2	0.98	1	0
3		2.82	2	0.24	1	0.02
4		0.3	2	0.86	1	0
5		9.39	2	0.01	0.16	0.01
6		4.02	2	0.13	1	0.02
7		28.9	2	0	0	0.06
8		1.85	2	0.4	1	0.01
9		20.43	2	0	0	0.05
10		3.76	2	0.15	1	0.02
11		7.91	2	0.02	0.29	0.03
12		5.47	2	0.06	0.84	0.02
13		0.36	2	0.84	1	0.01
14		7.79	2	0.02	0.29	0.01
15		0.65	2	0.72	1	0
16		0.44	2	0.8	1	0
17		18.24	2	0	0	0.03
18		1.81	2	0.4	1	0.02
19		8.27	2	0.02	0.26	0.02
20		4.71	2	0.09	1	0.02

برای پاسخ‌گویی به سوال دوم ازروش منتل هنزل با متغیر جورکردنی نمرهٔ کل استفاده شد. نتایج نشان داد که در این روش هم هفت سوال دارای DIF جنسیت انـد. بـا توجـه بـه انـدازهٔ نمونهٔ مطالعهٔ حاضر، شاخص انـدازهٔ تـأثیر بـرای تفـاوت پارامترهـای دشـواری سـوالات نیـز

محاسبه شد. دوران و هلند[1] (۱۹۹۳) DIF را بر حسب اندازهٔ تأثیر به سه مقوله تقسیم کردند نوع A یا نوع قابل چشم پوشی، نوع B یا متوسط ونوع C یا شدید. DIF نوع A وقتی حادث می‌شود که کای اسکوئر در سطح ۰/۰۵ به‌طور معنی‌دار نباشد واندازه تاثیر کمتر از یک باشد. DIF نوع C وقتی حادث می‌شود که کای اسکوئر در سطح ۰/۰۵ معنی‌دار باشد و اندازه تاثیر از یک و نیم بیشتر باشد. سوالاتی که درجه DIF آن ها نه در مقوله A و نه در مقوله C قرار می‌گیرد در مقولهٔ B دسته‌بندی می‌شوند. از میان هفت سوال که بر اساس آزمون معنی‌داری به‌عنوان سوال دارای DIF جنسیت گزارش شده بودند پنج سوال دارای DIF نوع A یا قابل چشم پوشی و دو سوال دیگر به‌عنوان سوالات دارای DIF متوسط گزارش شدند. قابل ذکر است که در روش تشخیصی طبقه‌بندی و منتل هنزل با نمرهٔ کل به‌عنوان متغیر جورکردنی سوالات متفاوتی به‌عنوان سوالات دارای DIF معرفی شدند در روش تشخیصی طبقه‌بندی سوال هفتم و در روش منتل هنزل سوالات پنجم و نهم به‌عنوان سوالات دارای DIF معرفی شدند.

با توجه به نتایج به‌دست آمده، جواب سوال دوم مبنی بر اینکه "آیا تعداد سوالات دارای DIF در روش منتل هنزل از روش تشخیصی طبقه‌بندی بیشتر است؟" مثبت می‌باشد. نتایج تحقیق حاضر نشان می‌دهد که تعداد سوالات دارای DIF وقتی که متغیر جورکردنی نمره کل می‌باشد، نسبت به وقتی که متغیر جورکردنی نیمرخ خصیصه‌ای می‌باشد، بیشتر است. این نتیجه را اینطور می‌توان توجیه کرد که وقتی که متغیر جورکردنی نمرهٔ کل می‌باشد، ابعاد مختلف سازهٔ اندازه‌گیری شده توسط آزمون در آن دیده نشده درصورتی‌که آزمون‌شوندگان از گروه‌های مختلف عملکرد متفاوتی در این ابعاد داشته باشند، سوال مورد نظر با عنوان سوال دارای DIF شناخته می‌شود به همین دلیل این احتمال وجود دارد که سوالات بیشتری به‌عنوان سوال دارای DIF شناخته شود. در حالی که در روش تشخیصی طبقه‌بندی خصیصه‌های مختلفی که درجوابگویی به سوالات یک آزمون تاثیرگذارند از قبل مشخص می‌شوند و در متغیر جورکردنی دیده می‌شوند و بنابراین عملکرد متفاوت آزمون‌شوندگان در هر کدام از این خصیصه‌ها، باعث نمی‌شود که سوال مورد نظر به‌عنوان سوال DIF معرفی گردد. زیرا این خصیصه‌ها همگی خصیصه مربوط در نظر گرفته می‌شوند و واریانس ناشی از وجود این

1. Dorans & Holland

خصیصه‌ها به‌عنوان واریانس مربوط در نظر گرفته می‌شود. نتایج مطالعهٔ حاضر همگام است با نتایج به‌دست آمده توسط لی(۲۰۰۸) همگام است. لی در یک مطالعهٔ شبیه سازی دریافت که در روش تشخیصی طبقه‌بندی **خطای نوع یک**[1] کمتر و **قدرت**[2] بیشتری نسبت به منتل هنزل با نمرهٔ کل به‌عنوان متغیر جورکردنی وجود دارد. هر چند که لی در بخش دوم مطالعهٔ مذکور با استفاده از داده‌های واقعی دریافت که، نسبت به مطالعهٔ شبیه سازی ، دو روش از تفاوت کمتری برخوردارند.

عملکرد تقریبا یکسان دو روش تشخیصی طبقه‌بندی و منتل هنزل به هنگام استفاده از داده‌های واقعی رااینطور می‌توان توجیه کرد که احتمالا ماتریس کیوی تهیه شده برای آزمون ممکن است دقیق نباشد که پارامترهای لغزش نسبتا بزرگ سوالات آزمون مورد مطالعهٔ می‌تواند دلیلی بر این مدعا باشد. در واقع یکی از محدودیت‌های اساسی استفاده از روش‌های DIF که از نیمرخ خصیصه‌ای به‌عنوان متغیر جورکردنی استفاده می کنند، این است که درصورتی‌که در تعیین خصیصه‌های مورد نیاز برای هر آزمون به‌طور کل وهر سوال آن آزمون به‌طور خاص، دقت لازم صورت نگرفته باشد (تمامی خصیصه‌های مورد نیاز جهت عملکرد موفق در سوالات یک آزمون در نظر گرفته نشده باشد) و درصورتی‌که عملکرد افراد از گروه‌های مختلف در این سوالات متفاوت باشد، این سوالات به‌عنوان سوالات دارای DIF شناخته می‌شوند. خصیصه‌های که در ماتریس کیو مثلا برای یک آزمون خواندن در نظر گرفته می‌شوند می‌توانند **مربوط به آزمون**[3]، و **مربوط به متن**[4] باشد. از خصیصه‌های مربوط به آزمون می‌توان به راهبردهای مربوط به **شم آزمون دهی**[5] از قبیل استفاده از راهبرد حذف گزینه‌های نادرست در سوالات چند گزینه‌ای، پیش بینی جواب قبل از نگاه کردن به گزینه‌ها، انتخاب یک گزینه به دلیل در برداشتن کلمه یا عبارتی از متن، راهبردهای مدیریت از قبیل تنظیم سرعت خواندن برای افزایش درک مطلب، در نظر گرفتن زمان اختصاص داده شده به هر سوال اشاره کرد. خصیصه‌های مربوط به متن شامل های خصیصه‌های در نظر گرفته شده در مطالعهٔ حاضراست. به‌عنوان مثال دیگر از این خصیصه‌ها به موارد زیر می‌توان اشاره کرد. خلاصه

1. Type I error
2. Power
3. Test-related
4. Text-related
5. Test wiseness

کردن ایده های مطرح شده در متن، دانش واژگان، دانش دستور زبان و غیره. بیشتر مطالعات تشخیصی طبقه‌بندی انجام شده (یانگ[1]، ۲۰۰۹ ، لی ، اچ.[2]، ۲۰۱۱، راوند[3]، ۲۰۱۶) روی خواندن و درک مفاهیم، فقط خصیصه‌های مربوط به متن را در ماتریس کیو آورده اند. همچنین با توجه به اینکه روش‌های فعلی تهیهٔ ماتریس کیو بر اساس نظریهٔ متخصصان می‌باشد و روش‌های تجربی محدودی برای این امکان وجود دارد، احتمال دارد ماتریس های تدوین شده کامل و دقیق نباشند. به این معنی کامل که همه خصیصه‌های مورد نیاز برای آزمون در نظر گرفته نشده باشند و دقیق یعنی اینکه از میان خصیصه‌های در نظر گرفته شده برای عملکرد موفق در یک آزمون، خصیصه‌های مورد نیاز برای هر سوال به‌طور دقیق مشخص نشده باشد.

در زمینهٔ سوالات DIF در هر دو روش این نکته حایز اهمیت است که برای هر سه سوال در ماتریس کیو، خصیصهٔ استنباط در نظر گرفته شده بود. بررسی این سوالات نشان می‌دهد که در سوال پنجم، بر اساس متنی دربارهٔ اکو توریسم، از آزمون‌شوندگان سوال شده بود که جملهٔ پیشنهادی در کدام بند متن قرار می‌گیرد. سوال هفتم نیز بر اساس متن اکوتوریسم بود. در این سوال در خصوص نگرش نویسنده پرسیده شده بود. سوال نهم بر اساس متنی دربارهٔ هوش هیجانی و انجام آزمایشی براساس این هوش طراحی شده بود. این سوال که با عبارت " بر طبق متن..." شروع می‌شود راجع به اطلاعاتی که به‌طور صریح در متن قید شده سوال می کند. به نظر می‌رسد که خصیصهٔ دیگری غیر از خصیصهٔ استنباط برای پاسخگویی این سوال نیاز است که در ماتریس کیو دیده نشده است. اما سوال اینست که اگر ماتریس کیو در خصوص سوال نهم کامل نیست، چرا روش مبتنی بر مدل دینا به آن واکنشی نشان نداده و سوال به‌عنوان DIF معرفی نشده؟ همچنین در خصوص سوال پنجم با توجه به اینکه کلمات دشوار زیادی در متن وجود دارد، به نظر می‌رسد خصیصهٔ دانش واژگان نیز برای این سوال در ماتریس کیو باید دیده شود. اما اگر ماتریس کیو دقیق نیست و خصیصهٔ ثانویه ای در میان است که در متغییر جورکردنی (در اینجا نیمرخ خصیصه‌ای) دیده نشده و توضیع نمرات داطلبان مرد و زن در این خصیصه متفاوت است، چرا روش منتل هنزل به این مسئله

1. Jang
2. Li & H.
3. Ravand

حساسیتی نشان نداده؟ برای حل این تناقض شاید بهتر باشد با استفاده از **تکنیک برون فکنی اندیشه**[1] از گروهی از آزمون‌شوندگان زن و مرد خواست تا در ضمن جوابگویی به سوالات خواندن مورد مطالعهٔ به برون گویی افکار خود بپردازند تا از این طریق تصویر کامل تری از خصیصه‌های مورد نیاز برای جوابگویی به سوالات خواندن آزمون کنکور منحصر به زبان به‌دست آورد.

بنابرآنچه که گفته شد می‌توان گفت هرچندکه در روش‌هایی همچون مدل‌های طبقه‌بندی شناختی که از نیمرخ خصیصه‌ای به‌عنوان متغییر جورکردنی استفاده می‌شود نسبت به روش‌هایی که از نمرهٔ کل برای این منظور استفاده می کنند تعداد سوال کمتری ممکن است به‌عنوان سوال دارای DIF معرفی شوند، باز هم در این روش هنگامی که یک سوال به‌عنوان DIF گزارش می‌شود، معلوم نیست علت کدام یک از موارد زیر می‌باشد: آیا DIF در نتیجهٔ عملکرد متفاوت گروه‌های مختلف در یک خصیصه نامربوط می‌باشد؟، آیا DIF در نتیجه عملکرد متفاوت گروه‌های مختلف در یک خصیصه مربوط است، ولی در ماتریس کیو دیده نشده است؟. فقط درصورتی‌که پاسخ سوال اول مثبت باشد، می‌توان گفت سوالات DIF دارای سوگیری‌اند و در نتیجه تهدیدی برای روایی سازهٔ آزمون تلقی می‌شوند. در نتیجه برای اینکه روش DIF مبتنی بر مدل تشخیصی طبقه‌بندی دینا بتواند به ویژگی‌های وعده داده شده خود عمل کنند، باید اطمینان حاصل کرد تمامی های خصیصه‌های مربوط اعم از خصیصه‌های مربوط به متن و همچنین مربوط به توانش راهبردی از جمله راهبردهای **هدف گذاری**[2]، **برنامه ریزی**[3]، و **ارزیابی**[4] در تدوین ماتریس کیو دیده شده باشند که کار بسیار سخت و شاید بتوان گفت تقریبا ناممکن است. با توجه به محدودیت مذکور احتیاط لازم در استفاده از نتایج به‌دست آمده از روش DIF مبتنی بر مدل‌های طبقه‌بندی تشخیصی باید اعمال شود. آخر اینکه گاهی اوقات محققان مجبور می‌شوند در انتخاب بین دو مدل، از خواسته‌های ایده‌آل خود عدول کرده و مدلی را انتخاب کنند که با توجه به شرایط آن ها عملی تر باشد. با توجه به اینکه مدل‌های تشخیصی طبقه‌بندی تشخیصی به لحاظ نظری و عملی دارای پیچیدگی های خاصی اند استفاده از آنان برای بسیاری از محققان کار ساده‌ای نیست (راوند و روبیچ[5]،

1. Think-aloud protocol
2. Goal setting
3. Planning
4. Evaluation
5. Ravand & Robitzsch

۲۰۱۵). در چنین شرایطی استفاده از نمرهٔ کل به‌عنوان متغیر جورکردنی در چارچوب یکی از روش‌های سنتی تشخیص DIF برای بسیاری از پژوهشگران گزینه مناسب‌تری خواهد بود.

۹- منابع

Ahmadi, A., Darabi, A. (2016). Gender differential item functioning on a national field-specific test: The case of PhD entrance exam of TEFL in Iran. *Iranian Journal of Language Teaching Research, 4(1)*, 63-82.

Alavi, S. M., Rezaee, A. & Amirian, S. M. R. (2011).Academic discipline DIF in an English language proficiency test. *Journal of English Language Teaching and Learning*, *5*(7): 39-65

Alderman, D. L., & Holland, P. W. (1981). *Item performance across native language groups on the Test of English as a Foreign Language* (TOEFL Research Report No. 9). Princeton, NJ: Educational Testing Service.

Amirian, S. M. R., Alavi, S. M., & Fidalgo, A. M. (2014). Detecting gender DIF with an English proficiency test in EFL context. *Iranian Journal of Language Testing*, *4*(2), 187-203.

Angoff, W. H. (1993). Perspectives on differential item functioning methodology. In P. W. H. H. Wainer (Ed.), *Differential item functioning* (pp. 3-23). Hillsdale, NJ, England: Lawrence Erlbaum Associates, Inc.

Bailey, K. (1999). Washback in language testing (TOEFL Monograph Series 15). Princeton, NJ: Educational Testing Service.

Barati, H., & Ahmadi, A., R. (2010). Gender-based DIF across the Subject Area: A Study of the Iranian National University Entrance Exam. *Journal of Teaching Language Skills, 2(3)*, 1-26.

Barati, H., Ketabi, S., Ahmadi, A.(2006).Differential item functioning in high stakes tests: the effect of field of study. *IJAL*, *19*(2), 27-42.

Breland, H., Lee, Y.-W., Najarian, M., & Muraki, E. (2004). *An analysis of the TOEFL CBT writing prompt difficulty and comparability of different gender groups* (TOEFL esearch Report No. 76). Princeton, NJ: Educational Testing Service.

Bridgeman, B. and Wendler, C. 1991. Gender differences in predictors of college mathematics performance and in college mathematics classes. *Journal of Educational Psychology, 83(2):* 275–284.

Carlton, S. T &Harris, A. M. (1992). Patterns of gender differences on mathematics items on the Scholastic Aptitude Test. *Applied Measurement in Education, 6(2)*, 137-151. doi: 10.1207/s15324818ame06023

Curley, W. and Schmitt, A. P. 1993: *Revising SAT-Verbal items to eliminate Differential Item Functioning*. College Board Report 93-2. New York: College Entrance Examination Board

Dorans, N. J., & Holland, P. W. (1993). DIF detection and description: Mantel-Haenszel and standardization. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 35–66). Hillsdale, NJ: Erlbaum

Dorans, N. J., & Kulick, E. (1986). Demonstrating the utility of the standardization approach to assessing unexpected differential item functioning on the Scholastic Aptitude Test. *Journal of Educational Measurement*, *23*, 355–368.

Fidalgo, A. M., Alavi, S. M. & Amirian, S. M. R. (2014). Strategies for testing statistical and practical significance in detecting DIF with logistic regression models. *Language Testing*, first published on April 11, 2014 as doi: 10.1177/0265532214526748 pp. 1-19

Hemmati, Baghaei, & Bemani (2016). Cognitive diagnostic modeling of L2 reading comprehension ability: providing feedback on the reading performance of Iranian candidates for the university entrance examination, *International Journal of Language Testing*, *6*, 92-100.

Hou, L., de la Torre, J. d., & 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.

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

Jang, E. E. (2009). Cognitive diagnostic assessment of L2 reading comprehension ability: Validity arguments for fusion model application to LanguEdge assessment. *Language Testing*, *26*, 31-73.

Jodoin, M. G., & Gierl, M. J. (2001). Evaluating type I error and power rates using an effect size measure with the logistic regression procedure for DIF detection. *Applied Measurement in Education, 14*(4), 329–349.

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

Kanarek, EA. (1988). *Gender differences in freshman performance and their relationship to use of the SAT in admissions.* Paper read at the annual meeting of the Regional Association for Institutional Research. *October, at Providence, RI*

Karami, H. (2011). Detecting gender bias in a language proficiency test. *International Journal of Language Studies, 5*(2).

Lawrence, I. M. and Curley, W. E. 1989: *Differential Item Functioning for males and females on SAT-Verbal Reading subscore items: follow-up study*. Educational Testing Service Research Report 89-22. Princeton, NJ: ETS

Lawrence, I. M. , Curley, W. E. and McHale, F. J. 1988: *Differential item functioning for males and females on SAT verbal reading subscore items.* Report No. 88-4. New York: College Entrance Examination Board.

Lee, Y.-W., Breland, H., & Muraki, E. (2005). Comparability of TOEFL CBT writing prompts for different native language groups. *International Journal of Testing, 5,* 131–158.

Li, F. (2008). *A modified higher-order DINA model for detecting differential item functioning and differential attribute functioning* (Doctoral dissertation). University of Georgia, Athens.

Li, H. (2011). A cognitive diagnostic analysis of the MELAB reading test. *Spaan Fellow*, 9, 17-46.

Liu, O. L., Schedl, M., Malloy, J., & Kong, N. (2009). *Does content knowledge affect TOEFL iBT™ reading performance? A confirmatory approach to differential item functioning* (Research Report No. RR-09-29). Princeton, NJ: Educational Testing Service.

Ma, W. & de la Torre, J. (2017). GDINA: The generalized DINA model framework. R package version 1.4.2. Retrived from https://CRAN.R-project.org/package=GDINA

Messick, S. (1996). Validity of performance assessments. In G. W. Phillips (Ed.), *Technical issues in large-scale performance assessment* (pp. 1-18). Washington, DC: U.S. Department of Education, National Center for Education Statistics.

Milewski, G. B., & Baron, P. A. (2002, April). *Extending DIF methods to inform aggregate report on cognitive skills*. Paper presented at the meeting of the National Council on Measurement in Education, New Orleans, LA.

Ravand, H. (2016). Application of a cognitive diagnostic model to a high-stakes reading comprehension test. Journal of Psychoeducational Assessment, 34(8), 782-799. http://dx.doi.org/10.1177/0734282915623053

Ravand, H., & Firoozi, T. (2016). Examining construct validity of the master's UEE using the Rasch model and the six aspects of the Messick's framework. *International Journal of Language Testing,* 6(1).

Ravand, H., & Robitzsch, A. (2015). Cognitive diagnostic modeling using R. *Practical Assessment, Research & Evaluation*, *20*(11), 1–12.

Rezaee, A., & Shabani, E. (2010). Gender differential item functioning analysis of the University of Tehran English Proficiency Test*. Pazhuhesh-e Zabanha-ye Khareji, 56,* 89-108.

Robitzsch, A., Kiefer, T., George, A. C., & Uenlue, A. (2017). CDM: Cognitive diagnosis modeling. R package version 6.0-101. https://CRAN.R-project.org/package=CDM

Ryan, K., & Bachman, L. (1992). Differential item functioning on two tests of EFL proficiency. *Language Testing, 9,* 12–29.

Salehi, M., & Tayebi, A. (2012). Differential item functioning (DIF) in terms of gender in the reading comprehension subtest of a high stakes test. *Iranian Journal of Applied Language Studies, l4*(1), 135-168

Schmitt, A. and Dorans, N. (1990). Differential item functioning for minority examinees on the SAT . *Journal of Educational Measurement* 27, 67-81 .

Shealy, R., & Stout, W. F. (1993a). An item response theory model for test bias. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 197–329). Hillsdale, NJ: Lawrence Erlbaum.

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

Wall, D., & Horák, T. (2008). *The impact of changes in the TOEFL examination on teaching and learning in Central and Eastern Europe—Phase 2, coping with change.* Princeton, NJ: ETS

Zhang, W. (2006). *Detecting differential item functioning using the DINA Model* (Doctoral dissertation). University of North Carolina, Greensboro.