



بررسی کنش افتراقی پرسش‌ها و عملکرد در آزمون: مقایسه رگرسیون لجستیک، مدل رش و منتل-هنزل

حسین کرمی*

(نویسنده مسئول)

استادیار، گروه زبان و ادبیات انگلیسی، دانشگاه تهران،

تهران، ایران

Email: hkarami@ut.ac.ir



علی خودی**

دکتری گروه زبان و ادبیات انگلیسی، پردیس کیش، دانشگاه تهران،

کیش، ایران

Email: Ali.khodi@ut.ac.ir



چکیده

یکی از ابزارهای بررسی عملکرد آزمون، «کنش افتراقی پرسش‌ها» است (Differential item functioning). این روش، می‌تواند عوامل تاثیرگذار بر عملکرد آزمودنی‌ها را پیدا کرده و از بروز سوگیری در آزمون جلوگیری کند. در دو دهه گذشته، روش‌های بسیاری برای تشخیص پیشنهاد شده است. شمار روش‌های تشخیص عملکرد افتراقی پرسش، گاهی باعث سردرگمی پژوهشگران می‌شود. از سوی دیگر، امکان مقایسه یافته‌های پژوهش‌هایی که با روش‌های مختلف به بررسی کنش افتراقی پرسش را پرداخته‌اند، دشوار می‌سازد. پژوهش حاضر به بررسی و سنجش نتایج به دست آمده از سه روش تشخیص کنش افتراقی پرسش پرداخته است: مدل رش، رگرسیون لجستیک و منتل-هنزل (MH). داده‌های استفاده شده در بررسی‌ها، برگرفته از آزمون توانش انگلیسی دانشگاه تهران (UTEPT) است که آزمون با اهمیت ویژه‌ای است و سالانه برای داوطلبان دکترا برگزار می‌شود. تجزیه و بررسی کنش افتراقی یکنواخت با سه روش برگرفته نشان داد که پرسش‌ها در عملکرد خود تفاوت‌های زیادی ندارند. نتایج تحلیل رگرسیون لجستیک، دو پرسش را برای وجود کنش افتراقی پیدا کرد که مشابه روش منتل-هنزل است. همچنین پرسش‌هایی به عنوان نشانگرهای کنش افتراقی قوی در مدل رش شناسایی شده بودند، همان پرسش‌ها بودند که در دو مدل دیگر نیز معرفی شده بودند. نتایج پژوهش حاضر نشان می‌دهد که استفاده از روش‌های مختلف برای بررسی وجود کنش افتراقی پرسش، الزاماً نتایج متفاوت ناگزیری را در پی ندارد و می‌توان از هر یک از روش‌های استفاده شده در این پژوهش بهره گرفت.

اطلاعات مقاله

تاریخ ارسال: ۱۳۹۹/۰۹/۱۹

تاریخ پذیرش: ۱۳۹۹/۱۰/۱۳

تاریخ انتشار: زمستان ۱۳۹۹

نوع مقاله: علمی پژوهشی

کلید واژگان:

کنش افتراقی، روایی، عدالت

آزمون، تبعیض

کلیه حقوق محفوظ است ۱۳۹۹

شناسه دیجیتال DOI: 10.22059/jflr.2021.315079.783

کرمی، حسین، خودی، علی. (۱۳۹۹). بررسی کنش افتراقی پرسش‌ها و عملکرد در آزمون: مقایسه رگرسیون لجستیک، مدل رش و منتل-هنزل. پژوهش‌های زبان‌شناختی در زبان‌های خارجی، دوره ۱۰، شماره ۴، زمستان ۱۳۹۹، از صفحه ۸۴۲ تا ۸۵۳.

Khodi, Ali, Karami, Hossein (2021). Differential Item Functioning and Test Performance: a Comparison Between the Rasch Model, Logistic Regression and Mantel-Haenszel. *Journal of Foreign Language Research*, 10 (4), 842-853.

DOI: 10.22059/jflr.2021.315079.783

* دکتر کرمی متخصص رشته آموزش زبان انگلیسی بوده و به‌طور تخصصی در زمینه بررسی روایی آزمون‌ها و سنجش و ارزیابی پژوهش می‌نماید.

** علی خودی دانش‌آموخته رشته آموزش زبان انگلیسی از دانشگاه تهران بوده و در حیطه پژوهشی آزمون‌سازی فعالیت می‌نماید.



Differential Item Functioning and Test Performance: a Comparison Between the Rasch Model, Logistic Regression and Mantel-Haenszel



Hossein Karami*
 (corresponding author)
 Assistant Professor, University of Tehran,
 Tehran, Iran.
 Email: hkarami@ut.ac.ir



Ali Khodi**
 PhD, University of Tehran, Kish international Campus,
 Kish, Iran
 Email: Ali.khodi@ut.ac.ir

ABSTRACT

Differential item functioning (DIF) is considered to be one of the tools for the examination of test fairness. This method is capable of finding the factors affecting the subjects' performance and prevent the occurrence of bias in the test. A plethora of methods for detecting Differential Item Functioning (DIF) has been suggested during the last couple of decades. The multiplicity of methods for diagnosing DIF might be a confusing issue for applied researchers and might lead to complications in the comparability of the findings of various DIF studies which have utilized different DIF detection techniques. This study aimed to investigate the comparability of results from three widely used DIF detection techniques: the Rasch model, Logistic Regression, and Mantel-Haenszel (MH). The data comes from an administration of the University of Tehran English Proficiency Test (UTEPT) which is a high-stakes test administered annually to PhD candidates. DIF analysis through the three techniques indicated that the three methods did not have significant differences in their performance. The Mantel-Hansel model flagged two items having DIF just similar to the findings of logistic regression model. Likewise, the items that were detected as strong-DIF items in Rasch model were the same as items detected by the two aforementioned models. Therefore, it might be concluded that use of different DIF detection techniques does not necessarily lead to flagging different items.

DOI: 10.22059/jflr.2021.315079.783

© 2021 All rights reserved.

ARTICLE INFO

Article history:
 Received:
 9th, December, 2020
 Accepted:
 2nd, January, 2021
 Available online:
 Winter 2021

Keywords:

Differential Item
 Functioning, fairness, bias,
 validity

Khodi, Ali, Karami, Hossein (2021). Differential Item Functioning and Test Performance: a Comparison Between the Rasch Model, Logistic Regression and Mantel-Haenszel. *Journal of Foreign Language Research*, 10 (4), 842-853.
 DOI: 10.22059/jflr.2021.315079.783

* Dr. Karami is an assistant professor of Applied Linguistics. His main research interests include various aspects of language testing and assessment.

** Ali Khodi is a PhD holder of applied linguistics who is interested in language testing.

۱. مقدمه

این انگاره که اعتبار (Validity) مهمترین ملاحظه در گسترش و بهره‌گیری از آزمون است، به‌واقعیتی بدیهی تبدیل شده است (چپل، [Chapelle, ۲۰۲۰](#)). بنابراین، اطمینان از معتبر بودن عملکرد یک آزمون بر عهده طراحان و توسعه‌دهندگان آزمون و کاربران آن است (بکمن، [Bachman, ۱۹۹۰](#)). یکی از تهدیدات و موانع اصلی، علیه اعتبار آزمون وجود واریانس برگرفته از عوامل بی‌ربط و سازه‌های نامرتب (construct-irrelevant factors) است، بدین معنا که عملکرد در آزمون، نباید تحت تأثیر عواملی قرار گیرد که متفاوت از سازه مورد نظر آزمون است. در غیر این صورت، آزمون، داری سوگیری خواهد بود (کرمی، [Karami, ۲۰۱۳](#)). در این راستا از روشی آماری که به‌طور گسترده برای تشخیص سوگیری پرسش‌ها در آزمون استفاده می‌شود، عملکرد کنش افتراقی پرسش‌ها (DIF) است.

کنش افتراقی هنگامی رخ می‌دهد که شرکت‌کنندگان در آزمون با سطح توانایی برابر، اما از دو گروه مختلف، شانس متفاوتی در پاسخ‌گویی به یک پرسش را داشته باشند. با این حال، نمی‌توان عنوان کرد که کنش افتراقی برابر و همسان، به‌معنای وجود تبعیض (bias) در سنجش است. بلکه می‌توان متصور شد که پیش‌شرطی برای به‌وجود آمدن تبعیض در سنجش است. بنابراین، وجود کنش افتراقی، شرط اساسی، اما ناکافی برای سوگیری در سنجش است. شیوه‌های مختلفی برای تشخیص کنش افتراقی در پژوهش‌های پیشین ارائه شده است (کلوزر و مازور، [Clauser & Mazor, ۱۹۹۸](#)؛ کرمی، [Karami, ۲۰۱۲](#)). آنچه مطرح است، تفاوت‌های عملکردی این شیوه‌های سنجش کنش افتراقی در بررسی‌های انجام شده است. بنابراین، بایسته است که چگونگی و امکان مقایسه این شیوه‌های واکاوی را در پروژه‌های مختلف بسنجیم و بررسی کنیم. در همین راستا، پژوهش می‌کوشد به‌بررسی عملکرد سه‌روش تشخیص کنش افتراقی پرسش‌ها بپردازد: مدل رش، رگرسیون لجستیک و متل-هنزل. همچنین در پژوهش حاضر اثرگذاری این شیوه‌های سنجش کنش افتراقی بر روایی و پایایی داده‌ها و یافته‌های آزمون مورد ارزیابی قرار می‌گیرد.

۲. مروری بر پژوهش‌های پیشین

بررسی ویژگی‌های روانشناختی دانش‌آموزان در زمینه‌های مختلف، مانند آموزش؛ بیشتر با استفاده از آزمون انجام می‌شود. هدف از این آزمایش‌ها ارزیابی افراد از نظر توانایی مورد نظر است که سازه‌ی مد نظر سنجش نام دارد (Construct of measurement) (آکار و کله‌سی اوغلو، [Acar & Kelecioğlu, ۲۰۱۰](#)) از آنجا که مشخصات توانایی‌های افراد باید از طریق یک ابزار اندازه‌گیری واجد شرایط (به‌عنوان مثال آزمون‌ها) انجام شود، وجود هر عامل یا فاکتور غیرمرتبط سازه‌ی ممکن است کاربرد و قابلیت آزمون را کاهش دهد. بنابراین، آزمون‌ها به‌طور کلی و موارد آزمایشی به‌طور خاص باید بتوانند توانایی افراد را اندازه‌گیری کنند، بدون اینکه ویژگی‌های متفرقه‌ای از شرکت‌کنندگان، کارکرد آنان را در سنجش سازه مورد نظر تحت تأثیر قرار دهد (یوآر، کلیسیاوغلو، و دوغان، [Uyar, Kelecioğlu, & Doğan, ۲۰۱۷](#)).

وجود این حساسیت به‌این دلیل است که افراد با توانایی برابر اما از گروه‌های مختلف باید بتوانند با همان درجه و شانس برابر به‌سوال‌های آزمون پاسخ دهند. اگر در یک آزمون پرسش‌هایی وجود داشته باشد که احتمال میزان پاسخگویی به‌آن توسط افراد از جنس، سن یا گروه قومی خاص بیش از سایرین باشد، احتمال وجود تبعیض یا کنش افتراقی غیرقابل تصور نیست (کامرون و همکاران، [Camero et al., ۲۰۱۴](#)) با این حال باید به‌یادداشت که وجود کنش افتراقی لزوماً به‌معنای وجود تبعیض در سنجش نیست و درحقیقت شرط لازم، اما ناکافی برای آن است (مکنامارا و روور، [McNamara, & Roever, ۲۰۰۶](#)).

فرایندهای تعیین وجود تبعیض در عملکرد پرسش‌ها را می‌توان از راه بررسی پاسخ‌های افراد با توانایی یکسان، اما گروه‌های مختلف (به‌عنوان مثال نژاد، جنسیت و غیره) انجام داد. بنابراین، در این بررسی احتمال پاسخ‌های صحیح داده‌شده به‌یک پرسش به‌عنوان کنش افتراقی آن (DIF) تعریف شده است (استینبرگ و تیسن، [Steinberg, & Thissen, 2006](#)). کنش افتراقی به‌عنوان احتمال مشروط برای دستیابی به‌پاسخ‌های صحیح تعریف می‌شود و زمانی اتفاق می‌افتد که یک پرسش موارد بیشتری از آن چیزی

را که باید بسنجد اندازه‌گیری می‌کند که آن متغیر اضافی اندازه‌گیری تعریف شده است و به‌عنوان "پارامتر مداخله گر" (nuisance) نامیده می‌شود (آکرمان، Ackerman، ۱۹۹۲). کنش افتراقی پرسش‌ها به دو دسته یکنواخت (uniform) و غیر یکنواخت (non-uniform) طبقه بندی می‌شود و به‌طور سنتی با مقایسه پاسخ‌های داده شده به پرسش‌ها توسط افراد با توانایی مشابه اما متعلق به گروه‌های متفاوت به نام گروه مرجع (reference group) یا گروه آزمایش (focal group) بررسی می‌شود (مکنامارا و روور، McNamara, & Roever, ۲۰۰۶). کنش افتراقی یکنواخت نیز به‌صورت احتمال پاسخ صحیح به پرسشی تعریف شده است که به‌طور یکنواخت برای همگی افراد با سطح توانایی متفاوت در گروه الف از افراد در گروه ب بیشتر باشد (زومبو، Zumbo، ۲۰۰۳) در حالی که کنش افتراقی غیریکنواخت که گاهی به‌عنوان کنش متقاطع (crossing) نیز شناخته می‌شود به‌حالتی اشاره می‌کند که یک پرسش برای افراد با سطح توانایی متفاوت در یک گروه، عملکرد متفاوتی را نشان می‌دهد.

برای تشخیص و بررسی کنش افتراقی شیوه‌های متفاوتی بررسی و به‌کار برده می‌شود که برخی از آنان براساس نظریه کلاسیک آزمون (Classical Test Theory) است، همچون رگرسیون لجستیک یا متل-هنزل؛ در حالی که برخی دیگر از شیوه‌های سنجش کنش افتراقی حالت پارامتریک داشته و با نام‌هایی همچون آزمون خی ۲ لرد (Lord's chi-square test) یا نسبت احتمال (likelihood ratio) که بر اساس نظریه پرسش و پاسخ (IRT) است شناخته می‌شوند (کمیلی و شپارد، Camili & Shepard, ۱۹۹۴).

برای مثال، در پژوهشی انجام شده ژو و آریادوست (Zhu & Aryadoust, ۲۰۲۰) تاثیر زبان مادری شرکت‌کنندگان آزمون و ارتباط آن را با کنش افتراقی مورد بررسی قرار گرفته است. یافته‌های آن نشان‌داد پرسش‌های آزمون، متأثر از زبان مادری زبان‌آموزان نبوده است. در پژوهشی دیگر، اثر کنش افتراقی پرسش‌ها بین شرکت‌کنندگان آلمانی و انگلیسی سنجیده شده است (فیشر و همکاران، Fischer et al., ۲۰۱۶). یافته‌ها حاکی از آن است که با وجود این اثر در

۴ پرسش آزمون، میزان آن قابل چشم پوشی است. با بررسی بانک پرسش‌های که برای آزمون ورودی موسسه مورد استفاده قرار می‌گرفت، النا اولیوری و همکاران (Elena Oliveri et al., 2018) به بررسی وجود کنش افتراقی برگرفته از اثرات میزان سطح و توانش زبانی، سن، ملیت و سطح و طبقه اجتماعی پرداختند. بدین منظور از دو روش متل هنزل و نظریه پرسش و پاسخ بهره گرفتند. یافته‌های آنان حاکی از آن بود که از باب ملیت افراد، پرسش‌ها دارای کنش افتراقی اند. در پژوهشی دیگر نیز توسط چن، لیو و زومبو (Chen, Liu, & Zumbo, 2020) روش جدید از سنجش کنش افتراقی پیشنهاد شده است که بر اساس نمره کل آزمون عملکرده و نیاز به سنجش و ارزیابی بیشتری دارد.

۳. روش انجام پژوهش

شرکت‌کنندگان

شرکت‌کنندگان این پژوهش ۳۰۰۰ تن از متقاضیانی بودند (از هر دو جنسیت مرد و زن) که از بین کل شرکت‌کنندگان آزمون توانش مهارت زبان انگلیسی دانشگاه تهران (UTEPT) انتخاب شدند. شرکت‌کنندگان آزمون دانشجویان مشغول به تحصیل در مقطع دکتری رشته‌های مختلف در دانشگاه تهران بودند که بازه سنی بیشتر؛ در رده سنی ۲۵ تا ۴۰ ساله بوده است. شرکت‌کنندگان در پژوهش حاضر بر اساس سوابق تحصیلی خود به دو گروه علوم انسانی و علوم و فناوری تقسیم شدند (هر گروه ۱۵۰۰ نفر). لزوم انتخاب تعداد برابر شرکت‌کنندگان برای هر دو گروه به دلیل پیشگیری از اثرگذاری اندازه نمونه بر یافته‌های پژوهش است.

ابزار پژوهش

متقاضیان دوره‌های دکتری دانشگاه تهران بایسته‌است که در آزمون سنجش توانش و مهارت زبان‌های خارجی به نام "آزمون مهارت انگلیسی دانشگاه تهران" (UTEPT) شرکت‌نموده و نتایج آن را به‌دپارتمان مربوطه ارائه‌دهند. هدف اصلی این آزمون شناسایی افرادی است که سطح مهارت انگلیسی مناسب و قابل قبولی دارند. این آزمون از سه‌بخش شامل دستور زبان، خواندن و درک مفاهیم و واژگان

تشکیل شده است. همه پرسش‌ها در قالب چند گزینه‌ای ارائه شده‌اند و بخش خواندن و درک مفاهیم شامل متونی است که بی‌درنگ پس از آن‌ها شماری پرسش درک مطلب به شرکت‌کنندگان ارائه می‌شود. شمار پرسش‌های درک مطلب برای هر بخش متفاوت است. در این آزمون یک‌نمره کل به داوطلبان گزارش می‌شود که به‌سادگی مجموع نمراتی است که در سه خرده آزمون کسب کرده‌اند. در این مطالعه ما به بررسی بخش دستور زبان این آزمون پرداخته‌ایم که شامل پرسش‌های مرتبط به مفاهیم دستور زبان انگلیسی می‌باشد.

تحلیل داده‌ها

طی سه دهه گذشته انواع مختلفی از شیوه‌های تخصیص کنش افتراقی پرسش‌ها ارائه شده است که از جمله می‌توان به روش‌های ساده مبتنی بر شاخص‌های دشواری پرسش‌ها (نمودار دلتا) یا شیوه‌های پیچیده مبتنی بر نظریه پرسش و پاسخ (IRT) اشاره نمود. رویکردهای مبتنی بر نظریه سوال-پاسخ به دلیل ظرافت مفهومی خود، از جمله روش‌های تشخیص کنش افتراقی‌اند که به‌طور گسترده استفاده می‌شوند. در مطالعه حاضر، به‌طور خاص از سه روش برای تجزیه و بررسی داده‌ها استفاده شده است: مدل رش، رگرسیون لجستیک و متل-هنزل.

در رگرسیون لجستیک، پاسخ پرسش‌ها به‌عنوان متغیر وابسته در نظر گرفته می‌شود که باید از طریق سایر متغیرها پیش‌بینی شود. متغیرهای مورد نظر در این شیوه نمره کل، متغیر گروه‌بندی و هم‌کنشی بین این دو هستند. روش متل-هنزل یک رویکرد تشخیص کنش افتراقی غیرپارامتریک است که بر اساس ایده نسبت شانس (odd ratio) استوار است. نظریه نسبت شانس با "جمع آوری اطلاعات در بین سطوح متفاوت متغیر در حال سنجش (به‌طور معمول نمره آزمون مشاهده شده) به‌دست می‌آید تا ارزیابی شود که احتمال موفقیت در یک پرسش خاص برای گروه مرجع در مقایسه با سایر شرکت‌کنندگان در گروه آزمایش چقدر متفاوت است" (سیرسی و ریوس، Sireci & Rios، ۲۰۱۳، ص ۱۷۵)

همچنین برای ارزیابی کنش افتراقی، سه مدل رگرسیون را می‌توان تعیین کرد که مدل اول فقط نمره کل را شامل می‌شود مدل دوم هم نمره کل و عامل گروه‌بندی، و مدل آخر این دو عامل به‌اضافه مدت تعامل. تفاوت بین مدل اول و

سوم باید از طریق آزمون χ^2 (Chi-square) ارزیابی شود. در بحث شیوه سنجش کنش افتراقی پرسش‌ها با نظریه رگرسیون لجستیک می‌توان عنوان داشت که در سه حالت قابل تبیین است: در حالت اول، مدل فقط دربردارنده نمره واقعی (true score) است در مدل دوم، بررسی در بردارنده نمره واقعی و متغیر دسته‌بندی کننده است و در مدل سوم متغیری به‌نام هم‌کنشی هر دو متغیر (ذکر شده در حالت دوم) به بررسی‌ها اضافه می‌شود. تفاوت بین حالت اول و سوم از طریق نمره χ^2 (Chi square) قابل بررسی است که اگر میزان معناداری را نشان داد می‌توان در نظر داشت که کنش افتراقی یکنواخت یا غیریکنواخت در این آزمون وجود دارد. گام دوم، بررسی تفاوت مدل اول و دوم است که نشانگر وجود کنش افتراقی یکنواخت است، در حالی که مقایسه مدل دوم و سوم بیانگر وجود کنش افتراقی غیر یکنواخت است. در باره حجم نمونه نیز می‌توان از مربع ضریب آر (R2) استفاده کرد (جودوین و گریل، Jodoin & Gierl، ۲۰۰۱). در این معیار مقادیر کمتر از ۰,۳۵، نوع الف کنش افتراقی بوده و قابل چشم‌پوشی است، بین ۰,۳۵ تا ۰,۷۰، نوع ب کنش افتراقی و میزان متوسط است و بیش از ۰,۷۰، نوع ج وجود کنش افتراقی بوده و مقادیر زیادی را نشان می‌دهد.

در باره سنجش کنش افتراقی با روش متل-هنزل نیز اگر شاخص سنجش (MH D-DIF) مثبت باشد، حاکی از آن است که آن پرسش برای گروه مرجع دشوارتر از شرکت‌کنندگان در گروه آزمایش بوده است و اگر آن شاخص منفی باشد، نشانگر آن است که این پرسش برای گروه آزمایش دشوارتر بوده است. بر اساس معیارهای سازمان ETS (زیکی، ۱۹۹۳، Zieky) پرسش‌ها نوع الف، کنش افتراقی را از خود نشان می‌دهند، اگر میزان قطعی شاخص MH D-DIF کمتر از یک باشد و نوع ج کنش افتراقی را نشان می‌دهند، اگر آن شاخص بیشتر از ۱,۵ باشد یا به‌طور معناداری از یک بیشتر باشد و سایر پرسش‌ها با شاخص‌های متفاوت از آنچه اشاره شد از نوع ب کنش افتراقی‌اند.

۴. نتایج و بحث و بررسی

با بررسی‌های انجام شده و واکاوی داده‌های پژوهش به‌شیوه‌های رگرسیون لجستیک، متل-هنزل و مدل رش

شاخص‌های کیفی سنجش پرسش‌ها به دست آمد و در جداول زیر ارائه شده است. در این بخش، مقایسه یافته‌های هر سه روش صورت پذیرفته و بررسی شده است. از آن رو که استفاده از مدل‌های بررسی کنش افتراقی بر مبنای نظریه کلاسیک آزمون پیش‌فرض‌های خاصی ندارد، نخست، در این بخش به تبیین تناسب مدل و داده‌ها و مفروضات نظریه رش می‌پردازیم. و سپس نتایج یافته‌های هر سه مدل ارائه می‌شود.

در جدول شماره یک در ستون اول (item) شماره پرسش در آزمون نشان داده شده و اینکه ترتیب هر پرسش در آزمون برای شرکت‌کنندگان به چه صورت بوده است. در ستون دوم (Measure) به میزان و درجه سختی هر پرسش در واحد لوجیت (logit) اشاره شده است (لینکر، [Lincare](#)، ۲۰۱۰b). به طور مثال در این آزمون سخت‌ترین پرسش شماره ۹ بوده و ساده‌ترین پرسش شماره ۳ بوده است. رابطه مستقیمی بین قدر مطلق عدد نشان‌داده شده با درجه سختی

هر پرسش وجود دارد. در بارهٔ ستون سوم و چهارم باید عنوان داشت که در بخش اول (MNSQ) باید توجه داشته باشیم که این عدد به صورت کلی باید بین ۱،۳+ و ۱،۳- باشد، اگر چنین نبود به ستون بعدی آن یعنی ZSTD که اشاره به مربع نمره زی (Z-square) دارد رجوع کرده و در این ستون Z باید عددی بین ۲- تا ۲+ باشد، در غیر این صورت آیتم نرمال نیست و از تناسب در مدل خارج است (Overfit) (لینکر، [Lincare](#)، ۲۰۱۰a) که در این پژوهش چنین پرسش یافت نشد. قوانین مشابهی برای ستون ۵ و ۶ نیز قابل اعمال است. در ستون شماره ۷ و ۸ به همبستگی‌های اندازه‌گیری نقطه‌ای (point-measure correlation) پرداخته شده که نشانگر رابطه بین عملکرد افراد در پرسش و توانایی کلی افراد بر مبنای کل آزمون است. در ستون مورد انتظار (Expected) میزان همبستگی مورد انتظار پس از برازش داده‌ها در مدل نشان داده شده در حالی که در ستون مشاهده شده (observed) همبستگی رخ داده در بررسی تبیین شده است.

جدول شماره ۱: پیش‌فرض‌های مدل رش و تناسب داده‌ها و مدل

Item	Measure	Infit	Infit	Outfit	Outfit	PMC	PMC	DIF
		MNSQ	ZSTD	MNSQ	ZSTD	Observed	Expected	Contrast
۲۰	۰،۶۳	۱،۲۷	۹،۹	۱،۳۶	۹،۹	۰،۱۴	۰،۴۰	-۰،۳۲
۱۰	۰،۱۷	۱،۱۹	۹،۹	۱،۲۵	۹،۹	۰،۲۱	۰،۳۹	-۰،۷۷
۲۲	۰،۸۳	۱،۱۷	۹،۹	۱،۲۴	۹،۹	۰،۲۳	۰،۴۰	-۰،۳۲
۳۴	۰،۴۲	۱،۱۸	۹،۹	۱،۲۴	۹،۹	۰،۲۲	۰،۴۰	-۰،۲۰
۳۲	۰،۴۸	۱،۱۲	۸،۱	۱،۱۷	۸،۰	۰،۲۸	۰،۴۰	-۰،۳۳
۱	۰،۴۵	۱،۱۱	۷،۷	۱،۱۵	۶،۷	۰،۲۹	۰،۴۰	-۰،۲۸
۲۵	۱،۷۷	۱،۰۱	۰،۰۲	۱،۱۳	۳،۳	۰،۳۵	۰،۳۷	۰،۰۲
۱۱	۰،۷۹	۱،۰۹	۵،۶	۱،۱۲	۵،۳	۰،۳۱	۰،۴۰	-۰،۱۳
۱۸	-۰،۹۷	۱،۰۱	۰،۰۵	۱،۱۱	۲،۶	۰،۳۱	۰،۳۴	-۰،۲۹
۱۴	-۰،۷۹	۱،۰۳	۱،۰۶	۱،۱۰	۲،۶	۰،۳۱	۰،۳۵	-۰،۰۴
۳۱	-۰،۶۵	۱،۰۲	۱،۰۱	۱،۰۹	۲،۷	۰،۳۲	۰،۳۶	۰،۰۰
۱۲	-۰،۸۹	۱،۰۲	۰،۹	۱،۰۸	۲،۰	۰،۳۱	۰،۳۴	۰،۰۵
۳۵	-۰،۴۴	۰،۹۸	-۰،۹	۱،۰۴	۱،۴	۰،۳۸	۰،۳۷	۰،۰۹
۲۷	۱،۳۷	۰،۹۷	-۱،۴	۱،۰۴	۱،۲	۰،۴۰	۰،۳۹	۰،۰۶
۳	-۰،۰۲	۱،۰۲	۱،۴	۱،۰۲	۰،۹	۰،۳۷	۰،۳۹	۰،۰۰
۵	۰،۵۲	۱،۰۱	۰،۶	۱،۰۱	۰،۳	۰،۳۹	۰،۴۰	-۰،۳۴
۳۳	۰،۵۳	۱،۰۰	۰،۳	۱،۰۰	۰،۱	۰،۴۰	۰،۴۰	-۰،۰۶
۱۵	۰،۳۳	۱،۰۰	-۰،۳	۰،۹۹	-۰،۵	۰،۴۰	۰،۳۹	۰،۰۴
۱۳	-۱،۱۵	۰،۹۹	-۰،۲	۰،۹۵	-۱،۱	۰،۳۳	۰،۳۲	۰،۰۰
۶	-۰،۲۰	۰،۹۹	-۰،۵	۰،۹۷	-۱،۰	۰،۳۹	۰،۳۸	۰،۱۰

Item	Measure	Infit MNSQ	Infit ZSTD	Outfit MNSQ	Outfit ZSTD	PMC Observed	PMC Expected	DIF Contrast
۹	۱,۶۶	۰,۸۹	-۰,۰	۰,۹۶	-۱,۱	۰,۴۶	۰,۳۸	۰,۰۰
۲۹	۰,۲۵	۰,۹۵	-۳,۸	۰,۹۱	-۴,۲	۰,۴۵	۰,۳۹	-۰,۰۷
۲۳	۰,۳۲	۰,۹۵	-۳,۹	۰,۹۴	-۲,۹	۰,۴۴	۰,۳۹	۰,۱۶
۲۴	۰,۰۸	۰,۹۵	-۳,۸	۰,۹۳	-۲,۹	۰,۴۴	۰,۳۹	۰,۱۱
۷	۰,۱۹	۰,۹۴	-۴,۲	۰,۹۳	-۳,۱	۰,۴۵	۰,۳۹	۰,۱۷
۳۰	۰,۲۶	۰,۹۴	-۴,۶	۰,۹۱	-۴,۳	۰,۴۵	۰,۳۹	۰,۱۳
۱۹	-۰,۶۷	۰,۹	-۳,۲	۰,۹۲	-۲,۴	۰,۴۲	۰,۳۵	۰,۲۳
۲۱	-۱,۱۱	۰,۹۲	-۳,۲	۰,۸۵	-۳,۴	۰,۴۱	۰,۳۳	۰,۲۲
۱۶	-۰,۳۲	۰,۹۱	-۰,۶	۰,۸۵	-۰,۴	۰,۴۶	۰,۳۷	۰,۳۵
۴	-۰,۲۸	۰,۹۱	-۶,۰۰	۰,۸۴	-۶,۰	۰,۴۷	۰,۳۸	۰,۲۹
۲۶	-۱,۵۰	۰,۹۸	-۳,۷	۰,۷۶	-۴,۶	۰,۴۲	۰,۳۰	۰,۵۵
۱۷	-۰,۳۵	۰,۸۸	-۷,۶	۰,۸۱	-۷,۱	۰,۴۹	۰,۳۷	۰,۰۰
۲	-۰,۶۱	۰,۸۷	-۶,۸	۰,۷۸	-۷,۰	۰,۴۹	۰,۳۶	۰,۶۷
۸	-۰,۲۲	۰,۸۶	-۹,۰۰	۰,۸۱	-۷,۷	۰,۵۱	۰,۳۷	۰,۱۸
۲۸	-۰,۸۶	۰,۸۶	-۶,۵	۰,۷۶	-۶,۷	۰,۴۸	۰,۳۴	۰,۲۱

کنش افتراقی هر پرسش (DIF size) سوال‌های شماره ۲ و ۱۰ انتخاب شدند. علی‌رغم وجود میزان احتمال معنا داری برای برخی از پرسش‌ها نظیر ۵ و ۱۶ و ۲۶ به دلیل آنکه سائز کنش افتراقی معنی دار نبود آنان به‌عنوان پرسش‌های با سطح قابل چشم‌پوشی از کنش افتراقی نشان داده شدند و دسته بندی گردیدند.

می‌توان نتیجه گرفت که علی‌رغم بهره‌بری از شیوه‌های مختلف و روش‌های آماری متفاوت، هر سه شیوه در پژوهش حاضر در مورد نشانه‌گذاری پرسش‌ها با سطح کنش افتراقی متوسط و بالا مشترکات زیادی داشته و عملکرد یکسانی از خود ارائه دادند. در باره انتخاب و تشخیص پرسش‌ها با سطح اندک و قابل چشم‌پوشی از نظر کنش افتراقی به دلیل آنکه تست‌های آماری متفاوتی در هر شیوه به‌کار برده شده است یافته اندکی از هم تفاوت دارد و نیازمند مبنای نظری برای تفسیر است. آنچه مبرهن و واضح است این است که قدرت هر سه روش از منظر تشخیص مشابه و تا حد زیادی یکسان بوده است.

سپس در گام بعدی برای مقایسه قدرت تشخیص پرسش‌ها دارای کنش افتراقی از سه شیوه گفته شده در بخش پیشین بهره برده شد و نتایج آن در جدول شماره ۲ نشان داده شده است. در بحث بررسی کنش افتراقی با شیوه متتل-هنزل نشانگرهای χ^2 متتل-هنزل (Mantel-Haenszel chi-square) و لور یا الگوریتم طبیعی آلفا (MH LOR) و خطای استاندارد متتل-هنزل (LOR SE) سنجیده شد و بر اساس سطح معنادار بودن نشانگر MH CHI وجود کنش افتراقی بررسی و دسته‌بندی شد. در این بررسی پرسش‌های شماره ۲ و ۱۰ دارای کنش افتراقی با سطح متوسط تشخیص داده شدند.

در بررسی مشابه، با استفاده از رگرسیون لجستیک بر اساس فرضیه یک‌بعدی بودن داده-ها (unidimensionality)، شرکت‌کنندگان در دو گروه آزمایش و مرجع با یکدیگر تطبیق داده شدند و پرسش‌ها مشابه یافته‌های روش متتل-هنزل مجدد به‌عنوان پرسش‌هایی که کنش افتراقی از خود نشان داده‌اند انتخاب شدند. سرانجام، با بررسی داده‌ها از راه مدل رش و بررسی میزان احتمال وجود کنش افتراقی (Welch Prob.) و اندازه

جدول شماره ۲: مقایسه توانش مدل های تحلیل در تشخیص کنش افتراقی پرسش ها

سوال	نشانگر	متل=هنزل				رگرسیون لجستیک				مدل رش		
		ETS	MH CHI	MH LOR	LOR SE	ETS	$1^{st} R^2$	$2^{nd} R^2$	$3^{rd} R^2$	J & G	DIF size	Welch Prob.
۱	۲,۸۵۱	A	۰,۰۷۸۳	۰,۱۳۴۷	۲,۸۵۱	A	۰,۱۰	۰,۱۰	۰,۱۰	A	-۰,۱۳	۰,۰۰۰۴
۲	۳۴,۴۳۴	B	۰,۰۹۶۴	۰,۰۵۶۵	۳۴,۴۳۴	B	۰,۳۶	۰,۳۷	۰,۳۷	B	۰,۵۶	۰,۰۰۰۰
۳	۰,۲۶۹۷	A	۰,۰۸۲۲	۰,۰۴۶۱	۰,۲۶۹۷	A	۰,۱۸	۰,۱۸	۰,۱۸	A	۰,۰۶	۱,۰۰۰
۴	۴,۴۷۹	A	۰,۰۸۹۲	۰,۱۹۲۸	۴,۴۷۹	A	۰,۳۱	۰,۳۲	۰,۳۲	A	۰,۲۰	۰,۰۰۰۰۵
۵	۱۶,۴۱۴	A	۰,۰۸۳۱	۰,۳۴۰۰	۱۶,۴۱۴	A	۰,۲۰	۰,۲۰	۰,۲۰	A	-۰,۳۵	۰,۰۰۰۰
۶	۱,۰۵۱۵	A	۰,۰۸۴۵	۰,۱۰۷۶	۱,۰۵۱۵	A	۰,۲۱	۰,۲۱	۰,۲۱	A	۰,۱۰	۰,۲۰۴۴
۷	۱,۲۲۲	A	۰,۰۸۴۴	۰,۰۹۶۹	۱,۲۲۲	A	۰,۲۷*	۰,۲۷	۰,۲۷	A	۰,۰۹	۰,۰۳۸۰
۸	۰,۰۰۱۶	A	۰,۰۹۲۱	۰,۰۰۰۵	۰,۰۰۱۶	A	۰,۳۷	۰,۳۷	۰,۳۷	A	-۰,۱۰	۰,۰۲۸۱
۹	۰,۰۰۷۰	A	۰,۱۰۱۱	۰,۰۱۳۶	۰,۰۰۷۰	A	۰,۲۸	۰,۲۸	۰,۲۸	A	-۰,۰۱	۱,۰۰۰
۱۰	۴۱,۸۱۸	B	۰,۰۷۸۹	۰,۰۵۱۱۵	۴۱,۸۱۸	B	۰,۰۵	۰,۰۷	۰,۰۷	B	-۰,۵۱	۰,۰۰۰۰
۱۱	۰,۰۰۷	A	۰,۰۸۰۳	۰,۰۰۱۲	۰,۰۰۷	A	۰,۱۲	۰,۱۲	۰,۱۲	A	-۰,۰۱	۰,۱۲۰۱
۱۲	۱,۰۶۶۲	A	۰,۰۹۲۴	۰,۰۹۹۴	۱,۰۶۶۲	A	۰,۱۵	۰,۱۵	۰,۱۵	A	۰,۱۰	۰,۰۷۰۳
۱۳	۰,۰۰۰۱	A	۰,۰۹۹۸	۰,۰۰۴۲	۰,۰۰۰۱	A	۰,۱۸	۰,۱۸	۰,۱۸	A	۰,۰۱	۱,۰۰۰
۱۴	۰,۰۲۷۲	A	۰,۰۹۰۶	۰,۰۱۹۰	۰,۰۲۷۲	A	۰,۱۴	۰,۱۴	۰,۱۴	A	۰,۰۳	۰,۸۰۵۹
۱۵	۰,۱۴۹۰	A	۰,۰۸۲۲	۰,۰۳۵۱	۰,۱۴۹۰	A	۰,۲۱	۰,۲۱	۰,۲۱	A	۰,۰۳	۰,۰۶۱۴۴
۱۶	۸,۹۴۷۷	A	۰,۰۹۰۱	۰,۲۷۳۹	۸,۹۴۷۷	A	۰,۳۱	۰,۳۱	۰,۳۱	A	-۰,۲۸	۰,۰۰۰۰
۱۷	۲,۹۴۱۲	A	۰,۰۹۲۶	۰,۱۶۲۹	۲,۹۴۱۲	A	۰,۳۵	۰,۳۵	۰,۳۵	A	-۰,۱۶	۱,۰۰۰
۱۸	۸,۳۲۱۲	A	۰,۰۹۶۲	۰,۲۸۲۱	۸,۳۲۱۲	A	۰,۱۵	۰,۱۵	۰,۱۵	A	-۰,۲۸	۰,۰۰۱۷
۱۹	۲,۴۸۶۶	A	۰,۰۹۲۹	۰,۱۵۰۹	۲,۴۸۶۶	A	۰,۲۶	۰,۲۶	۰,۲۶	A	۰,۱۵	۰,۰۰۹۲
۲۰	۰,۰۰۰۳	A	۰,۰۷۶۴	۰,۰۰۱۵	۰,۰۰۰۳	A	۰,۰۱	۰,۰۱	۰,۰۱	A	-۰,۱۰	۰,۰۰۰۱
۲۱	۰,۴۳۹۶	A	۰,۱۰۳۷	۰,۰۷۴۳	۰,۴۳۹۶	A	۰,۲۷	۰,۲۷	۰,۲۷	A	۰,۰۹	۰,۰۲۳۶
۲۲	۱,۹۰۹۰	A	۰,۰۷۸۳	۰,۱۱۱۰	۱,۹۰۹۰	A	۰,۰۶	۰,۰۶	۰,۰۶	A	-۰,۱۰	۰,۰۰۰۱
۲۳	۲,۰۹۰۷	A	۰,۰۸۴۰	۰,۱۲۴۹	۲,۰۹۰۷	A	۰,۲۶	۰,۲۶	۰,۲۶	A	۰,۱۱	۰,۰۴۳۱
۲۴	۰,۴۳۳۱	A	۰,۰۸۵۰	۰,۰۵۹۶	۰,۴۳۳۱	A	۰,۲۶	۰,۲۶	۰,۲۶	A	۰,۰۵	۰,۱۷۶۴
۲۵	۱,۲۶۸۳	A	۰,۰۹۴۶	۰,۱۱۰۵	۱,۲۶۸۳	A	۰,۱۶	۰,۱۶	۰,۱۶	A	۰,۱۲	۰,۸۰۶۶
۲۶	۸,۹۰۲۳	A	۰,۱۱۹۶	۰,۳۶۲۳	۸,۹۰۲۳	A	۰,۳۲	۰,۳۳	۰,۳۳	A	۰,۳۸	۰,۰۰۰۰
۲۷	۰,۸۳۷۲	A	۰,۰۹۰۱	۰,۰۸۶۶	۰,۸۳۷۲	A	۰,۲۱	۰,۲۱	۰,۲۱	A	۰,۱۰	۰,۰۵۱۳۰
۲۸	۰,۰۵۸۶	A	۰,۱۰۲۶	۰,۰۳۰۱	۰,۰۵۸۶	A	۰,۳۷	۰,۳۷	۰,۳۷	A	-۰,۴۰	۰,۰۲۰۷
۲۹	۳,۲۷۹۵	A	۰,۰۸۴۶	۰,۱۵۶۴	۳,۲۷۹۵	A	۰,۲۷	۰,۲۷	۰,۲۷	A	-۰,۱۶	۰,۳۷۳۵
۳۰	۱,۰۰۲۱	A	۰,۰۸۵۱	۰,۰۸۹۰	۱,۰۰۲۱	A	۰,۲۷	۰,۲۷	۰,۲۷	A	۰,۰۷	۰,۱۰۴۳
۳۱	۰,۰۲۲۹	A	۰,۰۸۹۰	۰,۰۱۷۵	۰,۰۲۲۹	A	۰,۱۶	۰,۱۶	۰,۱۶	A	۰,۰۲	۱,۰۰۰
۳۲	۷,۶۰۱۸	A	۰,۰۷۹۱	۰,۲۲۱۵	۷,۶۰۱۸	A	۰,۱۱	۰,۱۱	۰,۱۱	A	-۰,۲۲	۰,۰۰۰۰
۳۳	۰,۳۷۴۶	A	۰,۰۸۱۸	۰,۰۵۳۵	۰,۳۷۴۶	A	۰,۲۰	۰,۲۰	۰,۲۰	A	-۰,۰۷	۰,۴۵۱۲
۳۴	۰,۰۹۴۸	A	۰,۰۷۷۳	-۰,۰۲۶۸	۰,۰۹۴۸	A	۰,۰۵	۰,۰۵	۰,۰۵	A	۰,۰۳	۰,۰۱۰۹
۳۵	۰,۶۹۶۹	A	۰,۰۸۷۳	۰,۰۷۶۷	۰,۶۹۶۹	A	۰,۲۱	۰,۲۱	۰,۲۱	A	۰,۰۷	۰,۲۶۳۲

با توجه با یافته‌های جدول شماره ۲ تا حدی می‌توان نتیجه گرفت که برای پرسش‌های ساده و یا پرسش‌های بسیار سخت شیوهٔ متئل-هنزل به‌صورت سخت‌گیرانه‌تری کنش افتراقی پرسش‌ها را تبیین نموده است. یکی دیگر از یافته‌های این پژوهش آن است که درصد پرسش‌ها داری کنش افتراقی در هر سه شیوه میزان برابری دارد. شباهت و توانش دو مدل متئل-هنزل و رگرسیون لجستیک را می‌توان در تشابه مدل آماری زیربنایی آنان یافت، اما تشخیص پرسش‌های مشابه، توسط مدل رش حاکی از آن است که دو روش دیگر هم توانایی قابل قبولی برای پیدا کردن پرسش‌های دارای کنش افتراقی دارند (همبلتون و راجرز، [Hambleton & Rogers](#), ۱۹۸۹).

تاثیر وجود کنش افتراقی بر عدالت آزمون را می‌توان از منظر تفاوت‌های کارکردی شرکت‌کنندگان آزمون سنجید. وجود پرسش‌هایی با ویژگی کنش افتراقی بدان مفهوم است که در مواردی سازه‌های غیرمرتبطی به‌پاسخ افراد تاثیر می‌گذارد و ملاک توانایی شرکت‌کنندگان آزمون نیست. در آزمون حاضر، به‌دلیل وجود تعداد اندک پرسش‌ها با ویژگی کنش افتراقی (۲ سوال) می‌توان بیان کرد که ۵,۷ درصد پرسش‌ها امکان تاثیرپذیری از عوامل نامرتبطی برای شرکت‌کنندگان وجود دارد که در برخی از موارد به‌نفع آنان و در برخی اوقات به‌ضرر آنان می‌باشد.

پرواضح است که به‌صورت کلی بخش بسیار اعظمی از آزمون حاضر هیچ‌گونه تبعیضی را برای شرکت‌کنندگان قائل نبوده و عدالت آزمون در بیش از ۹۵ درصد پرسش بخش دستور زبان به‌خوبی وجود داشته است. لیکن باید در نظر داشت که مفهوم عدالت آزمون مفهومی چند بعدی است و وجود یا فقدان آن متأثر از عوامل مختلفی است که از زمان طراحی آزمون باید آنان را کنترل کرد و سرانجام در زمان تفسیر نتایج، اثرگذاری آن عوامل را در نظر گرفت.

در بحث روایی و وجود کنش افتراقی نیز آنچه به‌دست آمده، حاکی از آن است که به‌طور کلی بخش دستور زبان جزء بخش‌هایی از آزمون‌های زبان است که کمترین میزان کنش افتراقی را نشان می‌دهد و از روایی خوبی برخوردار است (مکنامارا، روور، [McNmara & Rover](#), ۲۰۰۶) در همین پژوهش نیز مشاهده می‌شود که تعداد بسیار اندکی از

پرسش‌ها، کنش افتراقی از خود نشان دادند و سایر پرسش‌ها عملکرد دقیقی در سنجش سازه مد نظر خود داشتند. در رابطه با تاثیر پیشینه و رشته تحصیلی نیز باید در نظر گرفت که در بخش دستور زبان مشارکت و اثرگذاری خاصی مشاهده نشد (هیل، [Hale](#), ۱۹۸۸).

۵. نتیجه‌گیری

هدف از این پژوهش بررسی و تبیین قدرت هر یک از شیوه‌های سنجش و تشخیص کنش افتراقی پرسش‌ها آزمون بود که پیش‌تر معرفی شدند (مدل رش، رگرسیون لجستیک، متئل هنزل). هدف دیگر ارائه شده در این پژوهش، اثرگذاری این شیوه‌ها بر روایی و صحت تحلیل‌های آماری انجام شده بود که در خلال آنچه عنوان شده بدان نیز اشاره شد.

بر اساس آنچه در جدول شماره ۲ در بحث مقایسه شیوه‌های سنجش کنش افتراقی مطرح شد در تشخیص پرسش‌های با سطح کنش افتراقی متوسط و یا بالا (که البته در اینجا چنین پرسش‌های وجود نداشت) مدل متئل هنزل کارآمدی مناسب از خود ارائه می‌دهد.

باید در نظر داشت که به‌طور کلی مدل رگرسیون لجستیک مدل عمومی‌تر نسبت به مدل متئل-هنزل است و در محاسبات متئل-هنزل فقط مدلی از کنش افتراقی یافت می‌شود که در تمام سطوح یک متغیر وجود داشته باشد و نتایج این پژوهش نشان‌داد که مدل متئل-هنزل و مدل رگرسیون لجستیک هر دو به‌اندازه مدل رش در تشخیص کنش افتراقی توانمندند. اما این مبحث که آیا در بحث تشخیص کنش افتراقی غیر یکنواخت چگونه عمل می‌کنند، نیازمند بحث و بررسی بیشتر است. از نظر بررسی‌های آماری و زمان تحلیل، مدل رگرسیون لجستیک از دو مدل دیگر پیچیده‌تر و زمان‌برتر می‌باشد. باید این نکته را عنوان کرد که برای تست‌های بسیار ساده و یا بسیار سخت فرض بر آن بوده که شیوه متئل هنزل کارآمد می‌باشد، اما برای تست‌هایی با میزان سختی متوسط از کارآمدی آن کاسته می‌شود. این بخش حداقل با توجه به‌درجه سختی پرسش‌ها که محاسبه گردید می‌توان ذکر نمود که این آزمون از درجه سختی متوسطی برخوردار بوده و ناکارآمدی از بابت اعمال شیوه متئل-هنزل در آن مشاهده نگردید.

یافته‌های این پژوهش برای طراحان آزمون، برگزارکنندگان آزمون و به‌ویژه تحلیل‌گران نتایج آزمون‌ها رویکردهای جدیدی در بهره‌بری از شیوه‌های مختلف سنجش‌کنش افتراقی را پیشنهاد می‌کند. لیکن در تعمیم‌پذیری این یافته‌ها باید دقت لازم مد نظر قرار داده شود، زیرا مبحث مورد سنجش در پژوهش حاضر که با نام دستور زبان از آن یاد می‌شود سازه‌ای است که در تحلیل‌های آماری به‌عنوان سازه تک بعدی (unidimensional) شناخته می‌شود و در تطابق مفروضات نظریه سوال-پاسخ می‌باشد. فلذا اگر سازه ای که مساله‌ی عدم وابستگی مکانی (local independence) برای آن مطرح نباشد با شیوه‌های ذکر شده سنجیده شود شاید منجر به کسب نتایج متفاوتی گردد. همچنین محتمل است که اگر در بحث مدل‌های نظریه سوال-

پاسخ فراتر از مدل رش گام برداریم و مدل‌های دوپارامتری و سه پارامتری را نیز در تحلیل‌ها لحاظ نماییم نتایج تغییر می‌یابد و حساسیت نظریه سوال-پاسخ افزایش خواهد یافت. در بحث نرم افزاری نیز ممکن است که بتوان در راستای افزایش دقت سنجش کوشید و با استفاده از نرم افزارهای مختلف برای اعمال این مدل‌ها اطمینان حاصل نمود که یافته برگرفته از اثربخشی خود شیوه‌ها و مفروضات مدل‌های بررسی شده است نه اثر نرم افزارهای مختلف استفاده شده در این پژوهش. همچنین وابستگی یافته‌ها به تعداد پرسش‌ها هر آزمون نیز نکته دیگری است که پیشنهاد میشود در سطوح مختلف مورد بررسی قرار گیرد.

منابع

- Acar, T., & Kelecioğlu, H. (2010). Comparison of Differential Item Functioning Determination Techniques: HGLM, LR and IRT-LR. *Educational Sciences: Theory and Practice*, 10(2), 639-649.
- Ackerman, P. L. (1992). Predicting individual differences in complex skill acquisition: Dynamics of ability determinants. *Journal of applied psychology*, 77(5), 598.
- Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.
- Cameron, I. M., Scott, N. W., Adler, M., & Reid, I. C. (2014). A comparison of three methods of assessing differential item functioning (DIF) in the Hospital Anxiety Depression Scale: ordinal logistic regression, Rasch analysis and the Mantel chi-square procedure. *Quality of life research*, 23(10), 2883-2888.
- Camilli, G. (2006). Test fairness. In R. Brennan (Ed.), *Educational measurement* (4th ed.) (pp. 221-256). New York: American Council on Education & Praeger series on higher education.
- Camilli, G., & Shepard, L. (1994). *Methods for identifying biased test items*. Newbury Park, CA Sage.
- Chapelle, C. A. (2020). Validity in language assessment. *The Routledge Handbook of Second Language Acquisition and Language Testing*, 11.
- Chen, Z., & Henning, G. (1985). Linguistic and cultural bias in language proficiency tests. *Language Testing*, 2(2), 155-163.

- Chen, M. Y., Liu, Y., & Zumbo, B. D. (2020). A propensity score method for investigating differential item functioning in performance assessment. *Educational and Psychological Measurement, 80*(3), 476-498.
- Clauser, E. B. & Mazor, M. K. (1998) Using Statistical Procedures to Identify Differentially Functioning Test Items. *Educational Measurement: Issues and Practice*.17, 31-44.
- Elena Oliveri, M., Lawless, R., Robin, F., & Bridgeman, B. (2018). An exploratory analysis of differential item functioning and its possible sources in a higher education admissions context. *Applied Measurement in Education, 31*(1), 1-16.
- Fischer, H. F., Wahl, I., Nolte, S., Liegl, G., Brähler, E., Löwe, B., & Rose, M. (2017). Language-related differential item functioning between English and German PROMIS Depression items is negligible. *International Journal of Methods in Psychiatric Research, 26*(4), e1530.
- Hale, G. (1988). Student major field and text content: Interactive effects on reading comprehension in the TOEFL. *Language Testing, 5*(1), 49-61.
- Hambleton, R. K., & Rogers, H. J. (1989). Detecting potentially biased test items: Comparison of IRT area and Mantel-Haenszel methods. *Applied Measurement in Education, 2*(4), 313-334.
- 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.
- Karami, H. (2012) An introduction to Differential Item Functioning. *International Journal of Educational and Psychological Assessment, 11*(2), 59-76.
- Karami, H. (2013) The quest for fairness in language testing. *Educational Research and Evaluation, 19*(2&3), 158-169.
- Linacre, J. M. (2010a). *A User's Guide to WINSTEPS®*. Retrieved May 2, 2010 from <http://www.winsteps.com/>.
- Linacre, J. M. (2010b). *Winsteps® (Version 3.70.0) [Computer Software]*. Beaverton, Oregon: Winsteps.com.
- McNamara, T. & C. Roever (2006). *Language Testing: The Social Dimension*. Malden, MA & Oxford: Blackwell.
- Sireci, S. G., & Rios, J. A. (2013). Decisions that make a difference in detecting differential item functioning. *Educational Research and Evaluation, 19*(2-3), 170-187.
- Steinberg, L., & Thissen, D. (2006). Using effect sizes for research reporting: Examples using item response theory to analyze differential item functioning. *Psychological Methods, 11*(4), 402.

Uyar, Ş., Kelecioğlu, H., & Doğan, N. (2017). Comparing differential item functioning based on manifest groups and latent classes. *Kuram ve Uygulamada Eğitim Bilimleri*, 17(6), 1977-2000.

Zieky, M. (1993). Practical questions in the use of DIF statistics in test development. In P.W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 337-348). Hillsdale, NJ: Lawrence Erlbaum Associates.

Zumbo, B. D. (2003). Does item-level DIF manifest itself in scale-level analyses? Implications for translating language tests. *Language testing*, 20(2), 136-147.

Zhu, X., & Aryadoust, V. (2020). An investigation of mother tongue differential item functioning in a high-stakes computerized academic reading test. *Computer Assisted Language Learning*, 1-25.