پژوهش‌های زبان‌شناختی در زبان‌های خارجی
شاپای چاپی: ٤١٢٣-٢٥٨٨ شاپای الکترونیکی:٧٥٢١-٢٥٨٨
www.jflr.ut.ac.ir

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

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

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 دهۀ 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 پژوهش¬هایی که با روش‌های مختلف به‌بررسی کنش افتراقی پرسش را پرداخته‌اند، دشوار می‌سازد. پژوهش حاضر به‌بررسی و سنجش نتایج به‌دست‌آمده از سه‌روش 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 نشان‌داد 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 به‌عنوان 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، الزاما نتایج متفاوت ناگزیری را در پی ندارد و می‌توان از هر یک از روش‌های استفاده شده در این پژوهش بهره گرفت.

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. *پژوهش های زبان‌شناختی در زبان‌های خارجی*، (۱۰) ۴، ۸۴۲-۸۵۳.
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 زمینۀ بررسی روایی آزمون‌ها و سنجش و ارزیابی پژوهش می‌نماید.
** علی خودی دانش‌آموخته رشته آموزش زبان انگلیسی از دانشگاه تهران بوده و در حیطۀ پژوهشی آزمون‌سازی فعالیت می‌نماید.

JOURNAL OF FOREIGN LANGUAGE RESEARCH
PRINT ISSN: 2588-4123 ONLINE ISSN: 2588-7521
ww.Jflr.ut.ac.ir

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

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

JOURNAL OF FOREIGN LANGUAGE RESEARCH, Volume 10, Number 4, Winter 2021, Page 842 to 853

* 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 in 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 یک‌آزمون بر عهدهٔ 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 نامرتبط (construct-irrelevant factors) است، بدین معنا که عملکرد در آزمون، نباید تحت تأثیر عواملی قرار گیرد که متفاوت از سازهٔ 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 به‌یک 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 (کلاوزر 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 مقایسهٔ 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 منتل–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 زمینه های 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 کله سی اوغلو، Acar & Kelecioglu، 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 سازه‌ای ممکن است کاربرد و قابلیت آزمون را کاهش دهد. بنابراین ، آزمون‌ها به‌طور کلی و موارد آزمایشی به‌طور خاص باید بتوانند توانایی افراد را اندازه گیری کنند، بدون اینکه ویژگی‌های متفرقه‌ای از شرکت کنندگان، کارکرد آنان را در سنجش سازه مورد نظر تحت تاثیر 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.، ۲۰۱٤) با این حال باید به‌یادداشت که وجود کنش افتراقی لزوما به‌معنای وجود تبعیض در سنجش نیست و درحقیقت شرط لازم، اما ناکافی برای آن است (مکنامارا و روور، McNamara, & Roever، ۲۰۰٦)

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، ۲۰۰٦). کنش افتراقی به‌عنوان احتمال مشروط برای دستیابی به‌پاسخ‌های صحیح تعریف می‌شود و زمانی اتفاق می‌افتد که یک پرسش موارد بیشتری از آن چیزی

پژوهش‌های زبان‌شناختی در زبان‌های خارجی، دوره ۱۰، شماره ٤، زمستان ۱۳۹۹، از صفحه ۸٤۲ تا ۸۵۳

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 به‌دو دسته یکنواخت (uniform)و غیر یکنواخت (non-uniform) طبقه بندی می‌شود و به‌طور سنتی با مقایسه پاسخ‌های داده شده به‌پرسش‌ها توسط افراد با توانایی مشابه اما متعلق به‌گروه‌های متفاوت به‌نام 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 به‌پرسشی 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 نظریهٔ کلاسیک آزمون (Classical Test Theory) است، همچون رگرسیون لجستیک یا منتل-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 نظریهٔ پرسش و پاسخ (IRT) است شناخته می‌شوند (کمیلی و شپارد، Camili &Shepard، ۱۹۹٤).

برای مثال، در پژوهشی انجام شدهٔ ژو و آریادوست (Zhu & Aryadoust، ۲۰۲۰) تاثیر زبان مادری شرکت‌کنندگان آزمون و ارتباط آن را با کنش افراقی مورد بررسی قرار گرفته است. یافته‌های آن نشان‌داد پرسش‌های آزمون، متاثر 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., ۲۰۱٦). یافته‌ها حاکی از آن است که با وجود این اثر در ٤ 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) به‌بررسی وجود کنش افتراقی برگرفته از اثرات میزان سطح و توانش زبانی، سن، ملیت و سطح و طبقهٔ اجتماعی پرداختند. بدین منظور از دو روش منتل هنزل و نظریه پرسش و پاسخ بهره گرفتند. یافته‌های آنان حاکی از آن بود که از باب ملیت افراد، پرسش‌ها دارای کنش افتراقی‌اند. 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 نمرهٔ کل آزمون عملکرده و نیاز به‌سنجش 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 زبان‌انگلیسی دانشگاه تهران (UTEPT) انتخاب شدند. شرکت‌کنندگان آزمون دانشجویان مشغول به‌تحصیل در مقطع دکتری رشته‌های مختلف در دانشگاه تهران بودند که بازهٔ سنی بیشتر؛ در رده سنی ۲۵ تا ٤۰ ساله بوده است. شرکت‌کنندگان در پژوهش حاضر بر اساس سوابق تحصیلی خود به‌دو گروه علوم انسانی و علوم و فناوری تقسیم شدند(هرگروه ۱۵۰۰ نفر). لزوم انتخاب تعداد برابر شرکت‌کنندگان برای هر دو گروه به‌دلیل پیشگیری ازاثرگذاری اندازه نمونه بر یافته‌های پژوهش است.

ابزار پژوهش

متقاضیان دوره‌های دکتری دانشگاه تهران بایسته‌است که در آزمون سنجش توانش ومهارت زبان‌های خارجی به‌نام "آزمون مهارت‌انگلیسی 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. همهٔ 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 یک‌نمرهٔ 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 به‌مفاهیم 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 به‌روش‌های ساده مبتنی بر شاخص های دشواری پرسش‌ها (نمودار دلتا) یا شیوه های پیچیده مبتنی بر نظریه پرسش و پاسخ (IRT) اشاره نمود. رویکردهای مبتنی بر نظریه سوال–پاسخ به‌دلیل ظرافت مفهومی خود ، از جمله روش‌های تشخیص کنش افتراقی‌اند که به‌طور گسترده استفاده می‌شوند. در مطالعه حاضر، به‌طور خاص از سه‌روش برای تجزیه و بررسی داده‌ها استفاده شده است: مدل رش، رگرسیون لجستیک و منتل–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 نمره‌کل، متغیرگروه‌بندی و همکنشی بین این دو هستند. روش منتل–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 (به طور معمول نمره آزمون مشاهده شده) به‌دست می‌آید تا ارزیابی شود که احتمال موفقیت دریک 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 می شود 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 خی۲ (Chi-square) ارزیابی شود.

در بحث شیوهٔ سنجش کنش افتراقی پرسش‌ها با نظریه رگرسیون لجستیک می‌توان عنوان داشت که در سه‌حالت قابل تبین است: در حالت اول، مدل فقط دربردارنده نمرهٔ واقعی (true score)است در مدل دوم، بررسی در بردارنده نمرهٔ واقعی و متغیر دسته‌بندی کننده است و در مدل سوم متغییری به‌نام همکنشی هر دو متغییر (ذکر شده درحالت دوم) به بررسی‌ها اضافه می‌شود. تفاوت بین حالت اول و سوم از طریق نمرهٔ خی۲ (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 بارهٔ حجم نمونه نیز می‌توان از مربع ضریب آر(R2) استفاده کرد (جودوین و گریل، Jodoin & Gierl، ۲۰۰۱). در این معیار مقادیر کمتر از۰,۰۳۵ نوع الف کنش افتراقی بوده و قابل چشم‌پوشی است، بین ۰,۰۳۵ تا ۰,۰۷۰ نوع ب کنش افتراقی و میزان متوسط است و بیش از۰,۰۷۰ نوع ج وجود کنش افتراقی بوده و مقادیر زیادی را نشان می‌دهد .

در بارهٔ سنجش کنش افتراقی با روش منتل–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، Zieky، ۱۹۹۳) 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.

٤. نتایج و بحث و بررسی

با بررسی‌های انجام شده و واکاوی داده‌های پژوهش به‌شیوه‌های رگرسیون لجستیک، منتل–هنزل و مدل رش

شاخص‌های کیفی سنجش پرسش‌ها به‌دست 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 به‌تبیین 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 نشان‌داده شده و اینکه ترتیب هر پرسش در آزمون برای شرکت‌کنندگان به‌چه صورت بوده است. در ستون دوم (Measure) به‌میزان و درجه سختی هر پرسش در واحد لوجیت (logit) اشاره شده است (لینکر،Lincare، b ۲۰۱۰). به‌طور 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 به‌صورت کلی باید بین ۱,۳+ و۱,۳- باشد، اگر چنین نبود به‌ستون بعدی آن یعنی ZSTD که اشاره به‌مربع نمره زی (Z-square) دارد رجوع کرده و در این ستون Z باید عددی بین۲- تا ۲+ باشد، در غیر این صورت آیتم نرمال نیست و از تناسب در مدل خارج است (Overfit)(لینکر، Lincare، a۲۰۱۰) که در این پژوهش چنین پرسش یافت نشد. قوانین مشابهی برای ستون ۵ و ٦ نیز قابل اعمال است. در ستون شماره ۷ و ۸ به‌همبستگی‌های اندازه گیری نقطه ای (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 تبین شده است.

جدول شماره ۱: پیشفرض‌های مدل رش و تناسب داده ها و مدل

Item	Measure	Infit MNSQ	Infit ZSTD	Outfit MNSQ	Outfit ZSTD	PMC Observed	PMC Expected	DIF Contrast
۲۰	۰,٦۳	۱,۲۷	۹,۹	۱,۳٦	۹,۹	۰,۱٤	۰,٤۰	-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	۱,۲٤	۹,۹	۰,۲۳	۰,٤۰	-۰,۳۲
۳٤	۰,٤۲	۱,۱۸	۹,۹	۱,۲٤	۹,۹	۰,۲۲	۰,٤۰	-۰,۲۰
۳۲	۰,٤۸	۱,۱۲	۸,۱	۱,۱۷	۸,۰	۰,۲۸	۰,٤۰	-۰,۳۳
۱	۰,٤۵	۱,۱۱	۷,۷	۱,۱۵	٦,۷	۰,۲۹	۰,٤۰	-۰,۲۸
۲۵	۱,۷۷	۱,۰۱	۰,۰۲	۱,۱۳	۳,۳	۰,۳۵	۰,۳۷	۰,۰۲
۱۱	۰,۷۹	۱,۰۹	۵,٦	۱,۱۲	۵,۳	۰,۳۱	۰,٤۰	-۰,۱۳
۱۸	-۰,۹۷	۱,۰۱	۰,۰۵	۱,۱۱	۲,٦	۰,۳۱	۰,۳٤	-۰,۲۹
۱٤	-۰۷۹	۱,۰۳	۱,۰٦	۱,۱۰	۲,٦	۰,۳۱	۰,۳۵	-۰,۰٤
۳۱	-۰,٦۵	۱,۰۲	۱,۰۱	۱,۰۹	۲,۷	۰,۳۲	۰,۳٦	۰,۰۰
۱۲	-۰,۸۹	۱,۰۲	۰,۹	۱,۰۸	۲,۰	۰,۳۱	۰۳٤	۰,۰۵
۳۵	-۰,٤٤	۰,۹۸	-۰,۹	۱,۰٤	۱,٤	۰,۳۸	۰,۳۷	۰,۰۹
۲۷	۱,۳۷	۰,۹۷	-۱,٤	۱,۰٤	۱,۲	۰,٤۰	۰,۳۹	۰,۰٦
۳	-۰,۰۲	۱,۰۲	۱,٤	۱,۰۲	۰,۹	۰,۳۷	۰,۳۹	۰,۰۰
۵	۰,۵۲	۱,۰۱	۰,٦	۱,۰۱	۰,۳	۰,۳۹	۰,٤۰	-۰,۳٤
۳۳	۰,۵۳	۱,۰۰	۰,۳	۱,۰۰	۰,۱	۰,٤۰	۰,٤۰	-۰,۰٦
۱۵	۰,۳۳	۱,۰۰	-۰,۳	۰,۹۹	-۰,۵	۰,٤۰	۰,۳۹	۰,۰٤
۱۳	-۱,۱۵	۰,۹۹	-۰,۲	۰,۹۵	-۱,۱	۰,۳۳	۰,۳۲	۰,۰۰
٦	-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	۱,٦٦	۰,۸۹	−۵,۰	۰,۹٦	−۱,۱	۰,٤٦	۰,۳۸	۰,۰۰
۲۹	۰,۲۵	۰,۹۵	−۳,۸	۰,۹۱	−٤,۲	۰,٤۵	۰,۳۹	−۰,۰۷
۲۳	۰,۳۲	۰,۹۵	−۳,۹	۰,۹٤	−۲,۹	۰,٤٤	۰۳۹.	۰,۱٦
۲٤	۰,۰۸	۰,۹۵	−۳,۸	۰,۹۳	−۲,۹	۰,٤٤	۰,۳۹	۰,۱۱
۷	۰,۱۹	۰,۹٤	−٤,۲	۰,۹۳	−۳,۱	۰,٤۵	۰,۳۹	۰,۱۷
۳۰	۰,۲٦	۰,۹٤	−٤,٦	۰,۹۱	−٤,۳	۰,٤۵	۰,۳۹	۰,۱۳
۱۹	−۰,٦۷	۰,۹	−۳,۲	۰,۹۲	−۲,٤	۰,٤۲	۰,۳۵	۰,۲۳
۲۱	−۱,۱۱	۰,۹۲	−۳,۲	۰,۸۵	−۳,٤	۰,٤۱	۰,۳۳	۰,۲۲
۱٦	−۰,۳۲	۰,۹۱	−۵,٦	۰,۸۵	−۵,٤	۰,٤٦	۰,۳۷	۰,۳۵
٤	−۰,۲۸	۰,۹۱	−٦,۰۰	۰,۸٤	−٦,۰	۰,٤۷	۰,۳۸	۰,۲۹
۲٦	−۱,۵۰	۰,۹۸	−۳,۷	۰,۷٦	−٤,٦	۰,٤۲	۰,۳۰	۰,۵۵
۱۷	−۰,۳۵	۰,۸۸	−۷,٦	۰,۸۱	−۷,۱	۰,٤۹	۰,۳۷	۰,۰۰
۲	−۰,٦۱	۰,۸۷	−٦,۸	۰,۷۸	−۷,۰	۰,٤۹	۰,۳٦	۰,٦۷
۸	−۰,۲۲	۰,۸٦	−۹,۰۰	۰,۸۱	−۷,۷	۰,۵۱	۰,۳۷	۰,۱۸
۲۸	−۰,۸٦	۰,۸٦	−٦,۵	۰,۷٦	−٦,۷	۰,٤۸	۰,۳٤	۰,۲۱

سپس در گام بعدی برای مقایسه قدرت تشخیص پرسش‌ها دارای کنش افتراقی از سه‌شیوهٔ 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 شیوهٔ منتل–هنزل نشانگرهای خی۲ منتل–هنزل(Mantel-Haenszel chi-square) و منتل–هنزل لور یا الگوریتم طبیعی آلفا MH) (LORو خطای استاندارد منتل–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 داده‌ها (unidimentionality)، شرکت‌کنندگان در دو گروه آزمایش و مرجع با یکدیگر تطبیق داده شدند و پرسش‌ها مشابه یافته‌های روش منتل–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 اندازهٔ کنش افتراقی هر پرسش (DIF size)سوال‌های شماره ۲ و ۱۰ انتخاب شدند. علی‌رغم وجود میزان احتمال معنا داری برای برخی از پرسش‌ها نظیر ۵و ۱٦و ۲٦ به‌دلیل انکه سایز کنش افتراقی معنی دار نبود آنان به‌عنوان پرسش‌های با سطح قابل چشم پوشی 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 بارهٔ انتخاب و تشخیص پرسش‌ها با سطح اندک و قابل چشم‌پوشی از نظر کنش افتراقی به‌دلیل آنکه تست‌های آماری متفاوتی در هرشیوه 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.

جدول شماره ۲: مقایسه توانش مدل های تحلیل در تشخصی کنش افتراقی پرسش‌ها

سوال	متتل=هنزل				رگرسیون لجستیک				مدل رش			
نشانگر	LOR SE	MH LOR	MH CHI	ETS	DIF χ^2	1st R^2	2nd R^2	3rd R^2	J & G	DIF size	Welch Prob.	ETS
۱	۲٫۸۵۱	۰٫۱۳٤۷	۰٫۰۷۸۳	A	٤٫۰٦٦	۰٫۱۰	۰٫۱۰	۰٫۱۰	A	-۰٫۱۳	۰٫۰۰۰٤	A
۲	۳٤٫٤۳٤	۰٫٥٦٥٥	۰٫۰۹٦٤	B	۳۳٫۳۳٤	۰٫۳٦	۰٫۳۷	۰٫۳۷	B	۰٫٥٦	۰٫۰۰۰۰	B
۳	۰٫۲٦۹۷	۰٫۰٤٦۱	۰٫۰۸۲۲	A	۳٫٦۲۲	۰٫۱۸	۰٫۱۸	۰٫۱۸	A	۰٫۰٦	۱٫۰۰۰	A
٤	٤٫٤۷۹	۰٫۱۹۲۸	۰٫۰۸۹۲	A	٥٫۲۰۸	۰٫۳۱	۰٫۳۲	۰٫۳۲	A	۰٫۲۰	۰٫۰۰۰٥	A
٥	۱٦٫٤۱٤	۰٫۳٤۰۰	۰٫۰۸۳۱	A	۲۲٫۷۲٥	۰٫۲۰	۰٫۲۰	۰٫۲۰	A	-۰٫۳٥	۰٫۰۰۰	A
٦	۱٫٥۱٥	۰٫۱۰۷٦	۰٫۰۸٤٥	A	۲٫۷٥۹	۰٫۲۱	۰٫۲۱	۰٫۲۱	A	۰٫۱۰	۰٫۲۰٤٤	A
۷	۱٫۲۲۲	۰٫۰۹٦۹	۰٫۰۸٤٤	A	**۳٫۱۱۳	۰٫۲۷	۰٫۲۷	۰٫۲۷	A	۰٫۰۹	۰٫۰۳۸۰	A
۸	۰٫۰۰۱٦	۰٫۰۰۰٥	۰٫۰۹۲۱	A	۰٫٤	۰٫۳۷	۰٫۳۷	۰٫۳۷	A	-۰٫۱۰	۰٫۰۲۸۱	A
۹	۰٫۰۰۷۰	۰٫۰۱۳٦	۰٫۱۰۱۱	A	۳۰٫٦٤۲	۰٫۲۸	۰٫۲۸	۰٫۳۰	A	-۰٫۰۱	۱٫۰۰۰	A
۱۰	٤۱٫۸۱۸	۰٫٥۱۱٥	۰٫۰۷۸۹	B	٤٤٫۱۷	۰٫۰٥	۰٫۰۷	۰٫۰۷	B	-۰٫٥۱	۰٫۰۰۰۰	B
۱۱	۰٫۰۰۷	۰٫۰۰۱۲	۰٫۰۸۰۳	A	٥٫۷۹۲	۰٫۱۲	۰٫۱۲	۰٫۱۲	A	-۰٫۰۱	۰٫۱۲۰۱	A
۱۲	۱٫۰٦٦۲	۰٫۰۹۹٤	۰٫۰۹۲٤	A	۱٫۲٦۸	۰٫۱٥	۰٫۱٥	۰٫۱٥	A	۰٫۱۰	۰٫٥۷۰۳	A
۱۳	۰٫۰۰۰۱	۰٫۰۰٤۲	۰٫۰۹۹۸	A	۰٫۳۸	۰٫۱۸	۰٫۱۸	۰٫۱۸	A	۰٫۰۱	۱٫۰۰۰	A
۱٤	۰٫۰۲۷۲	۰٫۰۱۹۰	۰٫۰۹۰٦	A	۰٫۰٦۱	۰٫۱٤	۰٫۱٤	۰٫۱٤	A	۰٫۰۳	۰٫۸۰٥۹	A
۱٥	۰٫۱٤۹۰	۰٫۰۳٥۱	۰٫۰۸۲۲	A	۰٫٥٥۹	۰٫۲۱	۰٫۲۱	۰٫۲۱	A	۰٫۰۳	۰٫۰٦۱٤٤	A
۱٦	۸٫۹٤۷۷	۰٫۲۷۳۹	۰٫۰۹۰۱	A	۱۰٫٥٦	۰٫۳۱	۰٫۳۱	۰٫۳۱	A	-۰٫۲۸	۰٫۰۰۰۰	A
۱۷	۲٫۹٤۱۲	۰٫۱٦۲۹	۰٫۰۹۲٦	A	٦٫۷٤۹	۰٫۳٥	۰٫۳٥	۰٫۳٥	A	-۰٫۱٦	۱٫۰۰۰	A
۱۸	۸٫۳۲۱۲	۰٫۲۸۲۱	۰٫۰۹٦۲	A	۸٫۱٥٥	۰٫۱٥	۰٫۱٥	۰٫۱٥	A	-۰٫۲۸	۰٫۰۰۱۷	A
۱۹	۲٫٤۸٦٦	۰٫۱٥۰۹	۰٫۰۹۲۹	A	۲٫٤۹٦	۰٫۲٦	۰٫۲٦	۰٫۲٦	A	۰٫۱٥	۰٫۰۰۹۲	A
۲۰	۰٫۰۰۰۳	۰٫۰۰۱٥	۰٫۰۷٦٤	A	۲٫۱٥۹	۰٫۰۱	۰٫۰۱	۰٫۰۱	A	-۰٫۱۰	۰٫۰۰۰۱	A
۲۱	۰٫٤۳۹٦	۰٫۰۷٤۳	۰٫۱۰۳۷	A	۲٫٤۱۰	۰٫۲۷	۰٫۲۷	۰٫۲۷	A	۰٫۰۹	۰٫۰۲۳٦	A
۲۲	۱٫۹۰۹۰	۰٫۱۱۱۰	۰٫۰۷۸۳	A	٤٫۲٦۳	۰٫۰٦	۰٫۰٦	۰٫۰٦	A	-۰٫۱۰	۰٫۰۰۰۱	A
۲۳	۲٫۰۹۰۷	۰٫۱۲٤۹	۰٫۰۸٤۰	A	۳٫۷۳۳	۰٫۲٦	۰٫۲٦	۰٫۲٦	A	۰٫۱۱	۰٫۰٤۳۱	A
۲٤	۰٫٤۳۳۱	۰٫۰٥۹٦	۰٫۰۸٥۰	A	۳٫۱۰٥	۰٫۲٦	۰٫۲٦	۰٫۲۷	A	۰٫۰٥	۰٫۱۷٦٤	A
۲٥	۱٫۲٦۸۳	۰٫۱۱۰٥	۰٫۰۹٤٦	A	۱٫۹۸۲	۰٫۱٦	۰٫۱٦	۰٫۱٦	A	۰٫۱۲	۰٫۸۰٦٦	A
۲٦	۸٫۹۰۲۳	۰٫۳٦۲۳	۰٫۱۱۹٦	A	۹٫۹۲۸	۰٫۳۲	۰٫۳۳	۰٫۳۳	A	۰٫۳۸	۰٫۰۰۰۰	A
۲۷	۰٫۸۳۷۲	۰٫۰۸٦٦	۰٫۰۹۰۱	A	*٥٫۸۰٦	۰٫۲۱	۰٫۲۱	۰٫۲۱	A	۰٫۱۰	۰٫٥۱۳۰	A
۲۸	۰٫۰٥۸٦	۰٫۰۳۰۱	۰٫۱۰۲٦	A	۰٫۱۹۸	۰٫۳۷	۰٫۳۷	۰٫۳۷	A	-۰٫٤۰	۰٫۰۲۰۷	A
۲۹	۳٫۲۷۹٥	۰٫۱٥٦٤	۰٫۰۸٤٦	A	۸٫۱۹٥	۰٫۲۷	۰٫۲۷	۰٫۲۷	A	-۰٫۱٦	۰٫۳۷۳٥	A
۳۰	۱٫۰۰۲۱	۰٫۰۸۹۰	۰٫۰۸٥۱	A	۱٫۸۷۷	۰٫۲۷	۰٫۲۷	۰٫۲۸	A	۰٫۰۷	۰٫۱۰٤۳	A
۳۱	۰٫۰۲۲۹	۰٫۰۱۷٥	۰٫۰۸۹۰	A	٥٫۹۸٦	۰٫۱٦	۰٫۱٦	۰٫۱٦	A	۰٫۰۲	۱٫۰۰۰	A
۳۲	۷٫٦۰۱۸	۰٫۲۲۱٥	۰٫۰۷۹۱	A	۱۱٫۳۲۷	۰٫۱۱	۰٫۱۱	۰٫۱۱	A	-۰٫۲۲	۰٫۰۰۰۰	A
۳۳	۰٫۳۷٤٦	۰٫۰٥۳٥	۰٫۰۸۱۸	A	۲٫۰٥۲	۰٫۲۰	۰٫۲۰	۰٫۲۰	A	-۰٫۰۷	۰٫٤٥۱۲	A
۳٤	۰٫۰۹٤۸	-۰٫۰۲٦۸	۰٫۰۷۷۳	A	۱٫۱۸۲	۰٫۰٥	۰٫۰٥	۰٫۰٥	A	۰٫۰۳	۰٫۰۱۰۹	A
۳٥	۰٫٦۹٦۹	۰٫۰۷٦۷	۰٫۰۸۷۳	A	۰٫۹٦۱	۰٫۲۱	۰٫۲۱	۰٫۲۱	A	۰٫۰۷	۰٫۲٦۳۲	A

با توجه با یافته‌های جدول شماره ۲ تا حدی می‌توان نتیجه گرفت که برای پرسش‌های ساده و یا پرسش‌های بسیار سخت شیوهٔ منتل–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 سه‌شیوه میزان برابری دارد. شباهت و توانش دو مدل منتل–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 می گذارد و ملاک توانایی شرکت کنندگان 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 شرکت کنندگان 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 هیچ گونه 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 متاثر 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 بخش هایی از آزمون‌های زبان است که کمترین میزان کنش افتراقی را نشان می‌دهد و از روایی خوبی برخوردارست (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 حود 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 تبین قدرت هر یک از شیوه‌های سنجش و تشخیص کنش افتراقی پرسش‌ها آزمون بود که پیش‌تر معرفی شدند (مدل رش، رگرسیون لجستیک، منتل 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 منتل هنزل کارامدی مناسب از خود ارئه می‌دهد.

باید درنظر داشت که به‌طور کلی مدل رگرسیون لجستیگ مدل عمومی‌تر نسبت به‌مدل منتل–هنزل است و در محاسبات منتل–هنزل فقط مدلی از کنش افتراقی یافت می‌شود که در تمام سطوح یک متغییر وجود داشته باشد و نتایج این پژوهش نشان‌داد که مدل منتل–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 یکنواحت 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 منتل‌هنزل کارامد می‌باشد، اما برای تست هایی با میزان سختی متوسط از کارامدی 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 ناکارمدی از بابت اعمال شیوه منتل–هنزل در آن مشاهده نگردید.

یافته‌های این پژوهش برای طراحان آزمون، برگزارکنندگان آزمون و به‌ویژه تحلیل‌گران نتایج آزمون‌ها رویکردهای جدیدی در بهره‌بری از شیوه های مختلف سنجش کنش 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 به‌کسب نتایج متفاوتی گردد. همچنین محتمل است که اگر در بحث مدل‌های نظریه سوال-پاسخ فراتر از مدل رش گام برداریم و مدل های 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 افزار های مختلف برای اعمال این مدل ها اطمینان حاصل نمود که یافته برگرفته از اثربخشی خود شیوه‌ها و مفروضات مدل های بررسی شده است نه اثر نرم افزارهای مختلف استفاده شده در این پژوهش. همچنین وابستگی یافته‌ها به‌تعداد پرسش‌ها هر آزمون نیز نکته دیگری است که پیشنهاد میشود در سطوح مختلف مورد بررسی قرار گیرد.

منابع

Acar, T., & Kelecioglu, 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 Egitim 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.