فصلنامه مطالعات اندازه‌گیری و ارزشیابی آموزشی **Journal of Educational Measurement & Evaluation Studies**
سال چهارم، شمارهٔ ۸، زمستان ۱۳۹۳، صص ۲۱۱- ۱۸۷ **Vol. 4, No. 8, Winter 2015**

ارزیابی توان آماری تحلیل رگرسیون لوجستیک در آشکارسازی کنش افتراقی سؤال‌های آزمون

تاریخ دریافت مقاله: ۹۳/۰۳/۱۸

تاریخ پذیرش مقاله: ۹۳/۱۱/۱۷

مسعود گرامی‌پور *

چکیده

اگر چه تحلیل رگرسیون لوجستیک برای شناسایی سؤال‌های سودار آزمون‌های روان‌شناسی و علوم تربیتی معرفی شده است، اما تحقیقات اندکی به صورت تجربی توان آماری آن را مورد ارزیابی قرار داده است. هدف از تحقیق حاضر ارزیابی توان آماری تحلیل رگرسیون لوجستیک و بررسی عوامل مداخله‌گر در آشکارسازی کنش افتراقی سؤال‌های آزمون بود. برای پاسخگویی به سؤال‌های تحقیق از روش مطالعات شبیه‌سازی مونت کارلو استفاده شد. داده‌های مورد نیاز با استفاده از نرم افزار WINGEN و با توجه به عوامل مداخله‌گر شامل سه حجم نمونه متفاوت، دو نوع DIF هماهنگ یا ناهماهنگ ، چهار مقدار با شدت متفاوت DIF و سه سطح درصد سؤال‌های دارای DIF در ۷۲ شرایط مختلف آزمایشی با صد تکرار شبیه‌سازی شد. نتایج تحقیق حاضر نشان‌گر توان آماری مطلوب تحلیل رگرسیون لوجستیک در آشکارسازی کنش افتراقی سؤال است و پیشنهاد می‌شود این روش بیشتر برای تشخیص DIF هماهنگ و برای کارکرد دقیق در حجم‌های نمونه بسیار بزرگ استفاده شود.

واژگان کلیدی: تحلیل رگرسیون لوجستیک -کنش افتراقی سؤال- مطالعات شبیه‌سازی مونت کارلو

* عضو هیأت علمی دانشکده روان‌شناسی و علوم تربیتی دانشگاه خوارزمی (مسئول مکاتبات: mgramipour@yahoo. com)

مقدمه

متخصصان روانشناسی و علوم تربیتی اهمیت تغییرناپذیری اندازه‌گیری[۱] را به عنوان پیش‌نیازی برای مقایسه‌های گروهی گوشزد کرده‌اند (به طور مثال دراسگو[۲]، ۱۹۸٤؛ راجو[۳] ، لفیتی[٤] و برن[٥]، ۲۰۰۲؛ ریس[٦]، ویدامن[۷]و پوق[۸] ، ۱۹۹۳؛ وندربرگ[۹] ، ۲۰۰۲). هر چقدر که آزمون کارکرد متفاوتی برای گروه‌های مختلف داشته باشد، مقایسه‌های گروهی نیز غیر قابل اعتماد خواهد بود. بنابراین فقدان تغییرناپذیری اندازه‌گیری و کنش افتراقی سؤال (DIF[۱۰]) به عنوان تهدیدی جدی برای اعتبار[۱۱] آزمون تلقی می‌شود. به طوری که استانداردهای انجمن روان‌شناسی آمریکا (APA[۱۲]) به بررسی کنش افتراقی سؤال (DIF)؛ سؤال‌های آزمون برای ارزیابی منصفانه بودن آن تأکید دارد (برن و استوارت[۱۳]، ۲۰۰٦).

تعیین سؤال‌های سودار، گامی مهم در سنجش معتبر است. هر سؤال سودار دارای خطای نظام‌دار است که باعث نامعتبر شدن نتایج آزمون می‌شود (کمیلی[۱٤] و شپارد[۱٥]، ۱۹۹٤). اصطلاح کنش افتراقی سؤال به عنوان گواهی تجربی برای وجود یا عدم وجود سوگیری در سؤال مورد استفاده قرار می‌گیرد. هر سؤال موقعی دارای کنش افتراقی (DIF) است که افرادی با توانایی یکسان اما با گروهی متفاوت، احتمال یا

---

۱. measurement invariance

۲. Drasgow

۳. Raju

٤. Laffitte

٥. Byrne

٦. Reise

۷. Widaman

۸. Pugh

۹. Vanderberg

۱۰. Differential Item Functioning

۱۱. Validity

۱۲. American Psychological Association

۱۳. Stewart

۱٤. Camilli

15. Shepard

بختی متفاوت در دادن پاسخ درست به سؤال داشته باشند. این گروه‌ها در ادبیات کنش افتراقی (DIF) اصطلاحاً گروه‌های مرجع و کانونی[۱] نامیده می‌شوند. معمولاً گروه مرجع، گروه اکثریت و گروه کانونی، گروه اقلیت یا محروم در نظر گرفته می‌شود (سوامیناتان[۲] و راجرز[۳]، ۱۹۹۰). با کنترل توانایی در دو گروه اگر آزمودنی‌ها احتمال یا بخت متفاوت در پاسخ درست به یک سؤال داشته باشند، آن سؤال دارای کنش افتراقی است. باید توجه داشت که تفاوت زمانی معنی‌دار است که توانایی مورد نظر در دو گروه کنترل شده باشد. این بدین معنی است که تفاوت در عملکرد به تنهایی نشانه سوگیری سؤال نیست. ممکن است در بعضی مواقع آزمودنی‌ها در گروه‌های مختلف به واقع در توانایی مورد نظر تفاوت داشته باشند.

روش‌های مختلفی برای بررسی کنش افتراقی سؤال‌های آزمون وجود دارد. روش‌های آماری برای آشکارسازی کنش افتراقی (DIF) به دو دسته کلی تقسیم می‌شوند. مجموعه اول عموماً بر اساس نظریه کلاسیک آزمون (CTT[۴]) است، زیرا نمرات مشاهده شده آزمون به عنوان متغیر جفتی[۵] برای پیش بینی نمرات واقعی در نظر گرفته می‌شوند (لرد[۶] و ناویک[۷]، ۱۹۶۸). روش‌هایی که در این مجموعه قرار می‌گیرند، شامل روش‌های منتل هنزل (MH[۸]) و منتل هنزل تعمیم داده شده (GMH[۹]) (پنفیلد[۱۰] و الجینا[۱۱]، ۲۰۰۳؛ سو[۱۲] و وانگ[۱۳]، ۲۰۰۵)، روش اختلاف

---

۱. Reference and focal groups

۲. Swaminathan

۳. Rogers

٤. Classical Test Theory

٥. paired variable

٦. Lord

۷. Novick

۸. Mantel-Haenszel

۹. Generalized Mantel-Haenszel

۱۰. Penfield

۱۱. Algina

۱۲. Su

۱۳. Wang

میانگین استاندارد شده[۱] (زویک[۲]، تایر[۳]و لویس[٤]، ۱۹۹۹) و روش رگرسیون لوجستیک (سوامیناتان و راجرز، ۱۹۹۰؛ زومبو[٥]، ۱۹۹۹) است.

مجموعه دوم شامل روش‌های مبتنی بر نظریه متغیر پنهان است، زیرا نمره متغیر پنهان به عنوان متغیر جفتی برای برآورد نمرات متغیر پنهان مورد استفاده قرار می‌گیرد. روش‌های این مجموعه شامل مدل‌های پارامتریک نظریه پاسخ سؤال[٦] (IRT) و تحلیل عاملی تأییدی[۷] هستند (امبرتسون[۸] و ریس، ۲۰۰۰؛ وندر لیندن[۹] و همبلتون[۱۰]، ۱۹۹۷).

در مطالعهٔ حاضر از میان روش‌های مختلف بررسی کنش افتراقی سؤال‌های آزمون، توان آماری روش رگرسیون لوجستیک بررسی شد. منظور از توان آماری در تحقیق حاضر، توان آزمون رگرسیون لوجستیک در رد فرض صفر عدم وجود کنش افتراقی (DIF) است. هدف از انتخاب روش رگرسیون لوجستیک نیز این بود که این روش به لحاظ کارکرد از سایر روش‌های کلاسیک عملکرد بهتری دارد و به لحاظ اجرایی و نرم‌افزاری نسبت به روش‌های مبتنی بر متغیر پنهان سهل الوصول‌تر است.

پیشینه

رگرسیون لوجستیک به خانواده بزرگ‌تری از تحلیل‌ها تعلق دارد که مدل‌های خطی عمومی نامیده می‌شوند. رگرسیون لوجستیک به صورت گسترده‌ای برای بررسی احتمال داده‌های دو وجهی به عنوان یک تابع لوجستیک برای یک یا بیش از یک متغیر پیش بین مورد استفاده قرار می‌گیرد. در رگرسیون لوجستیک پاسخ سؤال به عنوان متغیر وابسته با توزیع برنولی دو وجهی در نظر گرفته می‌شود. متغیرهای

---

۱. Standardized Mean Difference

۲. Zwick

۳. Thayer

٤. Lewis

٥. Zumbo

٦. Item Response Theory

۷. confirmatory factor analysis

۸. Embretson

۹. Van der Linden

۱۰. Hambleton

پیش‌بین می‌توانند پیوسته یا گسسته باشند. همچنین متغیرهای گسسته می‌توانند رتبه‌ای یا صرفاً اسمی باشند (اگرستی[۱]، ۲۰۰۷). متغیرهای پیش‌بین در بستر کنش افتراقی سؤال عبارتند از الف- متغیر عضویت گروهی V ، ب- متغیر نمره کل آزمون Z که متغیر جفتی[۲] نامیده می‌شود. این نمره معمولاً از مجموع سؤال‌های آزمون به دست می‌آید. ج- متغیر تعامل[۳]میان V و Z . متغیر گروهی معمولاً دارای دو مقدار (گروه مرجع و کانونی) است، که یکی از آنها به عنوان نشانگر[٤] در نظر گرفته می‌شود. در مطالعات بیش از دو گروه، بیش از یک نشانگر عضویت مورد نیاز است. متغیر جفتی Z متغیری است که جایگزین متغیر پنهان W_t می‌شود. در عمل متغیر Z ، نمره کل مجموع سؤال‌های آزمون است که شامل سؤال مورد مطالعه نیز می‌شود. معمولاً در مطالعات مختلف یک متغیر جفتی مورد استفاده قرار می‌گیرد، درحالی که از لحاظ نظری هیچ مانعی برای استفاده از متغیرهای جفتی چندگانه وجود ندارد. سرانجام، متغیر پیش‌بین تعاملی بسته به جایگاه نشانگرهای متغیر عضویت گروهی V می‌تواند شامل بیش از یک متغیر باشد (میلسپ[۵]، ۲۰۱۱).

به منظور تعریف مدل رگرسیون لوجستیک ، فرض کنید $P(Xj=1|Z,V)$ احتمال شرطی پاسخ درست به سؤال j به شرط متغیرهای Z و V باشد. تحت مدل رگرسیون لوجستیک، این احتمال شرطی عبارت است از

$$P(X_j=1|Z,V) = \frac{\exp\left[B0+B1Z+B2V+B3(ZV)\right]}{1+\exp\left[B0+B1Z+B2V+B3(ZV)\right]}$$

که در آن:

B_0 پارامتر عرض از مبدأ،

B_1 ضریب رگرسیون برای متغیر جفتی،

B_2 ضریب رگرسیون برای نشانگر عضویت گروهی،

و B_3 ضریب رگرسیون برای تعامل است.

---

۱. Agresti

۲. Matching variable

۳. Interaction

٤. Indicator

۵. Millsap

معادله زیر بدین معنی است که لوجیت[۱] احتمال شرطی پاسخ درست به سؤال j عبارت است از:

$$Ln[\frac{P(Xj=1|Z,V)}{1-P(Xj=1|Z,V)}]= B_0+B_1Z+B_2V+B_3 (ZV)$$

معادله فوق نشان می‌دهد که لگاریتم بخت پاسخ به سؤال، یک تابع خطی از متغیر جفتی، نشانگر عضویت گروهی و تعامل آنها است. تغییر ناپذیری شرطی مشاهده شده (OCI[۲]) برای j امین سؤال بدین معنی است که بخت پاسخ به سؤال فقط به Z بستگی دارد یا $B_2=B_3=0$. این فرضیه توسط آزمون نسبت درست‌نمایی یا آزمون والد[۳] همراه با برآورد پارامترها و خطاهای استاندارد آن، قابل آزمون است. پارامترهای این مدل با استفاده از روش بیشینه درست‌نمایی به شیوه نیوتن- رافسون[۴] محاسبه می‌شوند. روش برآورد نیوتن- رافسون با استفاده از الگوریتم‌های عددی در رگرسیون لوجستیک به برآورد پارامترها می پردازد. این الگوریتم با یک حدس آغازین برای مقدار پارامترها شروع به کار می‌کند، تا جایی که تابع بخت را بیشینه نماید. تقریب‌های متوالی این الگوریتم، برآوردهای بیشینه درست‌نمایی پارامترها را محاسبه می‌کند. برای انجام این کار در رگرسیون لوجستیک، الگوریتم نمره‌گذاری فیشر[۵] به کار می‌رود که به آن الگوریتم نیوتن-رافسون می‌گویند (اگرستی، ۲۰۰۷).

رویکرد نسبت درست‌نمایی برای ارزیابی OCI با برازش سه مدل آغاز می‌شود. مدل اول فقط متغیر جفتی Z را به عنوان متغیر پیش‌بینی کننده استفاده می‌کند و متغیرهای عضویت گروهی V و تعامل آنها ZV را از مدل حذف می‌کند. این مدل محدود شده فرضیه OCI را بدین ترتیب ارائه می‌کند که احتمال پاسخ به سؤال فقط به متغیر جفتی Z بستگی دارد. مدل دوم و سوم تخلف از OCI را نشان می‌دهند. در مدل دوم، نشانگر عضویت گروهی به مدل اول اضافه می‌شود. در این مدل با کنترل متغیر Z، گروه‌های مختلف به طور منظم احتمال متفاوتی برای پاسخ به سؤال دارند. این مدل شبیه به سوگیری یک جهتی[۶] (DIF هماهنگ) در مدل متغیر پنهان است.

---

۱. logit

۲. Observed Conditional Invariance

۳. Wald

٤. Newton-Raphson

٥. Fisher

٦. unidirectional

مدل سوم متغیر تعامل Z و V را به مدل دوم اضافه می‌کند. مدل سوم احتمال پاسخ به سؤال را در گروه‌های مرجع و کانونی با وجود رابطه میان V و Z محاسبه می‌کند. در این مدل، سوگیری دو جهتی[۱] (DIF ناهماهنگ) مورد بررسی قرار می‌گیرد. مدل سوم، مدل کامل[۲] است. مدل دوم با قرار دادن $B_3=0$ از مدل سوم به دست می آید و در عوض مدل اول نیز با قرار دادن $B_2=0$ از مدل دوم به دست می‌آید.

مدل کامل سوم، مدلی اشباع شده[۳] است و نمی‌تواند به صورت مستقیم توسط آزمون نسبت درست‌نمایی بررسی شود، اما می‌توان توسط آزمون نسبت درست‌نمایی، مدل اول و دوم را با مدل سوم مقایسه کرد. درست‌نمایی نمونه‌ای تحت مدل سوم برای j امین سؤال عبارت است از

$$L_3= \Pi^{M}_{m=1} \Pi^{K}_{k=1} [P(X_j=1|Z,V)]^{fjmk} [1- P(X_j=1|Z,V)]^{Fmk -fjmk}$$

معادلات درست‌نمایی برای مدل‌های اول و دوم به عنوان موارد ویژه‌ای از درست‌نمایی معادله فوق به دست می‌آید. به منظور بررسی سوگیری سؤال، مدل اول با مدل سوم توسط لگاریتم نسبت درست‌نمایی زیر مقایسه می‌شود:

$$G_1^2= -2\ln[\frac{L1}{L3}]= 2\ln(L_3) - 2\ln(L_1)$$

در معادله فوق G_1^2 مقداری خی دو نسبت درست‌نمایی برای آزمون مدل اول در مقابل مدل اشباع شده سوم است. با توجه به اینکه در مدل اول دو پارامتر از مدل سوم حذف می‌شود، آماره مدل اول دارای df=2 است. در نمونه‌های بزرگ، G_1^2 دارای توزیع خی دو تحت فرض صفر $B_3 = 0$ $B_2 =$ است. آماره آزمون تخلف از OCI را ارزیابی می‌کند، این آماره می‌تواند سوگیری یک جهتی (DIF هماهنگ) و دو جهتی(DIF ناهماهنگ) را ارزیابی کند (سوامیناتان و راجرز، ۱۹۹۰).

بعضی از محققان، آزمون دارای df=2 را به دو آزمون دارای df=1 تقسیم می‌کنند که هر کدام روی شکل خاصی از سوگیری تمرکز می‌کند (کمیلی و کانگدون[٤]، ۱۹۹۹؛ زومبو، ۱۹۹۹). زومبو به رگرسیون لوجستیک به عنوان یک رگرسیون سلسله مراتبی نگاه می‌کند که شامل سه گام است: ۱- متغیر پیش بین جفتی وارد معادله می‌شود، ۲- سپس نشانگر عضویت گروهی وارد معادله می‌شود، ۳- و در آخر، متغیر تعامل وارد

---

۱. bidirectional

۲. full model

۳. Saturated

٤. Congdon

معادله رگرسیون لوجستیک می‌شود. در آزمونی با df=2، مدل مرحلهٔ سه با مدل مرحلهٔ یک مقایسه می‌شود. این آزمون، دو آزمون دارای df=1 را ترکیب می‌کند. آزمون اول، مدل مرحلهٔ دو را با مدل مرحلهٔ یک مقایسه می‌کند. این آزمون OCI یک جهتی (DIF هماهنگ) را بررسی می‌کند. رد فرض صفر به معنی وجود سوگیری یک جهتی (DIF هماهنگ) است. آزمون دوم، مدل مرحلهٔ سه را با مدل مرحلهٔ دو مقایسه می‌کند. این مقایسه، OCI یا سوگیری دو جهتی (DIF ناهماهنگ) را ارزیابی می‌کند. در شیوه استاندارد، مقایسه میان مدل مرحلهٔ دو و سه در وهله اول انجام می‌شود. اگر متغیر تعامل مورد نیاز بود، آزمون مرحلهٔ دو برای اثر اصلی متغیر عضویت گروهی اختیاری است. اگر متغیر تعامل V و Z مورد نیاز نبود، مدل مرحلهٔ دو و یک با یکدیگر مقایسه می‌شود.

بسته به اینکه متغیر تعامل در معادله مورد نیاز باشد، رویکرد رگرسیون لوجستیک انتخاب‌های متفاوتی برای برآوردهای اندازه اثر[۱] فراهم می‌کند. اگر متغیر تعامل موجود نباشد ولی نشانگر عضویت گروهی مورد نیاز باشد، ضریب رگرسیون B_2 برای متغیر عضویت گروهی به عنوان لگاریتم نسبت بخت برای پاسخ به سؤال مورد مطالعه تلقی می‌شود. نسبت بخت برای تفاوت‌های گروهی در توزیع متغیر جفتی Z تطبیق داده می‌شود.

زمانی که مدل کامل مورد نیاز است، تفسیر برآورد ضریب رگرسیون B_2 به سادگی نسبت بخت نیست. یک رویکرد متفاوت برای مستندسازی اندازه اثر سوگیری، استفاده از چندین اندازه R^2 است (کلوزر[۲] و میزر[۳]، ۱۹۹۸؛ زومبو، ۱۹۹۹). مقصود کلی این مقادیر، بررسی واریانس افزایشی تبیین شده توسط متغیرهای گروهی و تعاملی در ارتباط با مدل یک است. اما مشکل این مقادیر، عدم وجود نقطه برش واضح برای قضاوت در مورد معنی‌داری وجود سوگیری در سؤال است. اگرستی (۲۰۰۷) معتقد است که مقدار R^2 در رگرسیون لوجستیک هنوز مانند همتای خود در رگرسیون خطی مفید نیست.

---

۱. Effect size

۲. Clauser

۳. Mazor

توان آماری آزمون رگرسیون لوجستیک برای بررسی کنش افتراقی (DIF) ابتدا در دنیا توسط سوامیناتان و راجرز (۱۹۹۰) مطرح شده است. هیدلاگو[۱] و لوپز- پینا[۲] (۲۰۰٤) با استفاده از روش رگرسیون لوجستیک به طور متوسط توانستند ۷۸/٤۲ درصد سؤال‌های کنش افتراقی (DIF) را به درستی تشخیص دهند. همچنین جودین[۳] و گیرل[٤] (۲۰۰۱) در مطالعه خود توانستند کنش افتراقی (DIF) هماهنگ را به‌طور متوسط در ۷۵/۳۰ درصد تکرارها در شرایط شبیه‌سازی شده آشکار کنند.

یافته‌های نارایانان[٥] و سوامیناتان (۱۹۹٦) نشان می‌دهد که درصد سؤال‌های کنش افتراقی (DIF) با عملکرد رگرسیون لوجستیک رابطه معنی‌داری دارد. نتایج آنها نشان می‌دهد که با افزایش درصد سؤال‌های سودار در آزمون، توان آماری آزمون کاهش یافته و نرخ خطای نوع اول افزایش می یابد.

روش رگرسیون لوجستیک برای بررسی کنش افتراقی (DIF) آزمون در چندین مطالعه با روش‌های داده‌های طبقه‌ای نظیر آزمون منتل – هنسل (MH[٦]) مورد مقایسه قرار گرفته است (هیدالگو و لوپز- پینا، ۲۰۰٤). مطالعات آنها برتری نسبی روش رگرسیون لوجستیک را بر سایر روش‌های داده‌های طبقه‌ای نشان می دهد.

در ایران روش رگرسیون لوجستیک پیش‌تر برای بررسی سوگیری‌های جنسیتی در آزمون‌های زبان انگلیسی به کار گرفته شده است (رضایی و شعبانی، ۱۳۸۹)، همچنین گرامی‌پور و فلسفی‌نژاد (۱۳۹۳) تحلیل رگرسیون لوجستیک را برای شناسایی سؤال‌های سودار آزمون‌های روان‌شناسی و علوم تربیتی معرفی کرده‌اند. اما اکثر تحقیقاتی که انجام شده است توان آماری این روش را در تعامل با متغیرهای دیگر به صورت جامع مطالعه نکرده‌اند.

در این راستا نکته قابل تأمل در مورد روش‌های بررسی کنش افتراقی سؤال از جمله رگرسیون لوجستیک این است که توان آماری این روش‌ها در آشکارسازی کنش افتراقی (DIF) همیشه کامل نیست. برخی از منابع، توان آماری این روش‌ها در

---

۱. Hidalgo

۲. López-Pina

۳. Jodoin

٤. Gierl

٥. Narayanan

٦. Mantel-Haenszel

رد کردن فرض صفر عدم وجود کنش افتراقی (DIF) را عموماً تحت تأثیر حجم نمونه، نوع و شدت کنش افتراقی (DIF) و درصد سؤال‌های دارای کنش افتراقی (DIF) دانسته‌اند. نتایج این مطالعات به صورت مجزا در مورد رگرسیون لوجستیک غالباً به این نتیجه رسیده‌اند که حجم نمونه و شدت کنش افتراقی (DIF) رابطهٔ مستقیم با آشکارسازی کنش افتراقی (DIF) دارند و درصد سؤال‌های کنش افتراقی (DIF) با آشکارسازی کنش افتراقی (DIF) رابطهٔ معکوس دارد و آشکارسازی کنش افتراقی (DIF) ناهماهنگ، دشوارتر از کنش افتراقی (DIF) هماهنگ است (سوامیناتان و راجرز، ۱۹۹۰، ۱۹۹۳؛ زومبو، ۱۹۹۹؛ هیررا[۱]، ۲۰۰۵ و ایلوسا[۲] و ولس[۳]، ۲۰۱۳). برخی مطالعات نشان داده‌اند که رگرسیون لوجستیک در آشکارسازی کنش افتراقی (DIF) تحت تأثیر حجم نمونه و نوع کنش افتراقی (DIF) هماهنگ یا ناهماهنگ (ضرایب دشواری و تمیز سؤال) مورد بررسی است. به طور مثال هیررا (۲۰۰۵) نشان داد که اندازه گروه مرجع به صورت معنی‌داری بر میزان خطای نوع اول تأثیر ندارد، اما نسبت اندازه نمونه میان گروه مرجع و کانونی بر سؤال‌هایی با ضریب تمیز پایین تأثیر معنی‌داری دارد. همچنین تعامل معنی‌داری میان این دو عامل (اندازه و نسبت حجم نمونه) برای سؤال‌هایی با ضریب دشواری بالا وجود دارد. نتایج هیررا نشان داد که با افزایش حجم نمونه، توان آماری آزمون رگرسیون لوجستیک در آشکارسازی کنش افتراقی (DIF) افزایش می‌یابد. اما با افزایش تفاوت نسبت حجم نمونه در گروه‌های مرجع و کانونی، میزان آشکارسازی به نسبت کاهش می‌یابد. همچنین هیررا (۲۰۰۵) دریافت که در حجم نمونه‌های مختلف توان آماری آزمون رگرسیون لوجستیک در آشکارسازی کنش افتراقی (DIF) پایین است. سانتانا[۴] (۲۰۰۹) دریافت که حجم نمونه و درصد سؤال‌های کنش افتراقی (DIF) آزمون بر توان آماری آزمون رگرسیون لوجستیک تأثیر دارند. نتایج او نشان داد که با افزایش حجم کل نمونه، توان آماری آزمون رگرسیون لوجستیک افزایش می‌یابد و افزایش درصد سؤال‌های کنش افتراقی (DIF) با آشکارسازی کنش افتراقی (DIF) رابطهٔ معکوس دارد. ایلوسا و ولس (۲۰۱۳) در مطالعه‌ای نشان دادند که توان آماری آزمون

۱. Herrera

۲. Elosua

۳. Wells

٤. Santana

رگرسیون لوجستیک به صورت خطی و مستقیم با افزایش شدت کنش افتراقی (DIF) افزایش می‌یابد، در حالی که نرخ خطای نوع اول نیز به همان اندازه افزایش می یابد.

در عین حال با وجود این مطالعات در خصوص توان آماری آزمون رگرسیون لوجستیک در آشکارسازی کنش افتراقی (DIF)، چگونگی تأثیر این عوامل با وجود یکدیگر در یک مطالعه واحد و نحوه تعامل آنها در تأثیرگذاری بر آشکارسازی کنش افتراقی (DIF) تاکنون مورد بررسی قرار نگرفته و هنوز نامشخص است. زیرا برخی نتایج با یکدیگر ناهمخوان هستند و هیچ‌کدام از مطالعات قبلی همه عوامل مذکور را در یک مطالعه با هم مورد بررسی قرار نداده‌اند و به طور قطع نمی‌توان گفت که در شرایط مختلف همچنان رگرسیون لوجستیک در آشکارسازی کنش افتراقی (DIF) توان آماری قابل قبولی دارد. بنابراین سؤال‌های اصلی تحقیق حاضر عبارتند از:

– توان آماری آزمون رگرسیون لوجستیک در آشکارسازی کنش افتراقی سؤال‌های آزمون چگونه است؟

– نحوهٔ مداخلهٔ عواملی شامل حجم نمونه، نوع و شدت کنش افتراقی (DIF) و درصد سؤال‌های دارای کنش افتراقی (DIF) در آشکارسازی آن چگونه‌اند؟

روش

به منظور تعیین توان آماری رگرسیون لوجستیک در شناسایی کنش افتراقی سؤال‌های آزمون از مطالعات شبیه‌سازی شده موسوم به مطالعات مونت کارلو[۱] استفاده شد. چنین روشی به محقق اجازه می دهد که پارامترهای برآورد شده آزمودنی و سؤال را با مقادیر واقعی آن که در داده‌های واقعی غیرقابل مشاهده هستند، مقایسه کند. همچنین مطالعات شبیه‌سازی شده به محقق اجازه می‌دهد که نتایج نظری را در عمل تأیید کند. سرانجام این روش سریع‌تر، ارزان‌تر و آسان‌تر از جمع‌آوری اطلاعات از آزمودنی‌های واقعی است و به محقق اجازه می‌دهد مدل‌ها و روش‌های جدید روان‌سنجی را به سرعت و با هزینه بسیار کمی مورد بررسی قرار دهند (هارول[۲] و همکاران، ۱۹۹۶؛ اسپنس[۳]، ۱۹۹۳).

در شبیه‌سازی داده‌های مدل پرسش– پاسخ سه مرحله اساسی طی شد:

---

۱. Monte Carlo

۲. Harwell

۳. Spence

۱- تعیین شکل مدل پاسخ سؤال- پاسخ‌های سؤال غالباً به صورت تک بعدی، به لحاظ شرطی مستقل و به صورت مدل‌های لوجستیک تولید شد. انتخاب‌های پیش رو غالباً مدل‌های لوجستیک یک، دو و سه پارامتری هستند که در آنها احتمال پاسخ صحیح تابعی از توانایی آزمودنی و ویژگی‌های عملیاتی سؤال هستند. در تحقیق حاضر مدل دو پارامتری انتخاب شد.

۲- تعیین پارامترهای مدل پاسخ سؤال- مدل‌های پاسخ سؤال با وجود ویژگی‌های عملیاتی متغیر برای تعیین مدل سؤال‌های آزمون بسیار انعطاف پذیر هستند. بنابراین ویژگی‌های اندازه‌گیری آزمونی شبیه‌سازی شده تا قبل از اینکه پارامترهای سؤال مشخص نشده است، ثابت نیست. در تحقیق حاضر پارامترها از تحقیقات قبلی که توسط پارشال[۱] و میلر[۲] (۱۹۹۵) روی آزمودنی‌های واقعی انجام شده است، انتخاب شدند. آنها ابتدا پارامترهای سؤال را بر اساس مطالعات قبلی اجرای آزمون بر آزمودنی‌های واقعی برآورد کردند، سپس از آنها برای شبیه‌سازی داده‌ها استفاده کردند.

۳- تعیین شکل توزیع توانایی آزمودنی‌ها- توانایی آزمودنی‌ها معمولاً از توزیع جامعه‌ای مشخص به صورت تصادفی انتخاب می‌شود که اکثر اوقات توزیع نرمال در آن استاندارد است (هان[۳] و همبلتون، ۲۰۰۷).

به منظور تعیین توان آماری رگرسیون لوجستیک در شناسایی کنش افتراقی (DIF) آزمونی سی سؤالی با مدل دو پارامتری نظریه پاسخ سؤال[٤] (IRT) در مقیاس بزرگ یعنی با تعداد زیادی از آزمودنی‌ها شبیه‌سازی شد. معیار شناسایی کنش افتراقی (DIF) بر اساس سطح معنی داری آماری یعنی توان آماری آزمون در رد فرض صفر (با نرخ خطای نوع اول ۰/۰۵) بود. در این تحقیق تنها کنش افتراقی (DIF) مثبت درست[٥] مورد بررسی قرار گرفت. این نوع کنش افتراقی (DIF) زمانی بررسی می‌شود که وجود کنش افتراقی (DIF) شبیه‌سازی شده و با فرض وجود کنش افتراقی (DIF)، توان آماری روش‌های شناسایی کنش افتراقی (DIF) مورد بررسی قرار

---

۱. Parshall

۲. Miller

۳. Han

٤. Item Response Theory

٥. True Positive

می‌گیرد. سرانجام عوامل مداخله‌گر که تأثیر آن بر عملکرد رگرسیون لوجستیک در تشخیص کنش افتراقی (DIF) مورد بررسی قرار گرفت عبارتند از:

– حجم نمونه– سه نمونه با حجم‌های ۱۵۰، ۵۰۰ و ۱۰۰۰ آزمودنی که مختص آزمون‌های سرنوشت ساز است، شبیه‌سازی شد.

– **کنش افتراقی** (DIF) هماهنگ و ناهماهنگ: بسته به اینکه گروه مرجع نسبت به گروه کانونی در کدام‌یک از پارامترهای دشواری و تمیز سؤال با یکدیگر متفاوت باشند، روش‌های مختلف تشخیص کنش افتراقی (DIF) ممکن است در آشکار ساختن کنش افتراقی (DIF) این گونه سؤال‌ها متفاوت عمل کنند. برای بررسی تأثیر این عوامل به صورت شبیه‌سازی شده مقادیر ثابت ۰/۲۵ ، ۰/۵۰ ، ۰/۷۵ و ۱ به عنوان شدت کنش افتراقی (DIF) به پارامتر دشواری یا تمیز در گروه مرجع اضافه شد. اگر تفاوت گروه مرجع با گروه کانونی در مقدار پارامتر دشواری سؤال باشد، سؤال دارای کنش افتراقی هماهنگ است و در صورتی که اختلاف در پارامتر قدرت تمیز باشد، کنش افتراقی ناهماهنگ را به وجود می آورد (فلاورز[۱]، اوشیما[۲] و راجو، ۱۹۹۹).

– درصد سؤال‌های دارای کنش افتراقی (DIF)– سه سطح کنش افتراقی (DIF) سؤال ۱۰ درصد (۳ سؤال)، ۲۰ درصد (٦ سؤال) و ۳۰ درصد (۹ سؤال) در تحقیق حاضر شبیه‌سازی شدند. این سطوح در مطالعات قبلی مونت کارلو برای تشخیص کنش افتراقی (DIF) نیز به کار گرفته شده است (پارشال و میلر، ۱۹۹۵).

به منظور شبیه‌سازی پارامترهای دشواری و قدرت تمیز برای یک آزمون سی سؤالی با مدل دو پارامتری IRT از نرم افزار WINGEN 2 استفاده شد (هان و همبلتون، ۲۰۰۷). توزیع‌های مورد استفاده در تحقیق حاضر شبیه به مطالعه پارشال و میلر (۱۹۹۵) بود. آنها ابتدا بر اساس مطالعات قبلی اجرای آزمون واقعی روی آزمودنی‌های زنده برآوردی از پارامترهای سؤال‌های آزمون به دست آورده و سپس آن را برای شبیه‌سازی داده‌ها مورد استفاده قرار دادند.

شیوه تولید داده‌ها در نرم افزار WINGEN 2 به این صورت بود که برای یک تکرار ابتدا تعداد آزمودنی‌ها مشخص شد. با توجه به فرضیه‌های تحقیق حاضر، تعداد نمونه در یکی از حجم‌های ۱۵۰، ۵۰۰ و یا ۱۰۰۰ آزمودنی می توانست تعیین شود.

---

۱. Flowers

۲. Oshima

سپس توزیع توانایی آزمودنی‌ها مشخص شد. این توزیع می‌توانست در یکی از اشکال توزیع نرمال، هماهنگ[۱] یا بتا[۲] باشد. در تحقیق حاضر توزیع توانایی آزمودنی‌ها نرمال فرض شد. همچنین توزیع مقادیر توانایی نرمال با میانگین صفر و انحراف استاندارد یک مشخص شد. همچنین بر اساس هدف مطالعه می‌توان در این نرم افزار داده‌ها را تک بعدی[۳] یا چند بعدی[٤] تولید کرد. در تحقیق حاضر داده‌های آزمون قبل از اعمال کنش افتراقی (DIF) به صورت تک بعدی تولید شد. سپس ویژگی‌های سؤال شامل تعداد سؤال‌های آزمون، تعداد گزینه‌های سؤال و مدل IRT مورد نظر باید تعیین شد. تعداد سؤال‌های آزمون در تحقیق حاضر شامل سی سؤال بود. داده‌های تولید شده برای سؤال‌های دو گزینه‌ای قبل از اعمال کنش افتراقی (DIF) با مدل دو پارامتری IRT برازش داشت. سپس توزیع، میانگین و انحراف استاندارد پارامترهای دشواری و تمیز سؤال‌های آزمون مشخص شد. توزیع پارامتر قدرت تمیز سؤال‌های آزمون می‌توانست در یکی از اشکال توزیع هماهنگ، نرمال یا لگاریتم نرمال[۵] و توزیع پارامتر دشواری سؤال‌های آزمون می توانست در یکی از اشکال توزیع هماهنگ، نرمال یا بتا تعیین شود. در تحقیق حاضر توزیع پارامتر قدرت تمیز دارای توزیع نرمال با میانگین صفر و انحراف استاندارد ۰/۵ و توزیع پارامتر دشواری دارای توزیع نرمال با میانگین صفر و انحراف استاندارد ۰/۷۵ تعیین شد. برای تعیین توان آماری رگرسیون لوجستیک با توجه به عوامل مداخله گر مختلف، سه حجم نمونه متفاوت، دو نوع کنش افتراقی (DIF) هماهنگ یا ناهماهنگ ، چهار مقدار با شدت متفاوت کنش افتراقی (DIF) و سه سطح درصد سؤال‌های دارای کنش افتراقی (DIF)، ۷۲ شرایط مختلف آزمایشی با صد تکرار شبیه‌سازی شد. بنابراین در تحقیق حاضر ۷۲۰۰ مجموعه داده پاسخ سؤال شبیه‌سازی شد.

نتایج

۱. uniform

۲. Beta

۳. unidimesional

٤. multidimesional

۵. lognormal

برای آشکارسازی کنش افتراقی (DIF)، تحلیل رگرسیون لوجستیک با استفاده از نرم افزار SPSS انجام شد. نتایج مطالعهٔ کنش افتراقی (DIF) برای یک سؤال آزمون سی سؤالی در یک تکرار تحلیل رگرسیون لوجستیک برای یک مجموعه داده با کنش افتراقی (DIF) هماهنگ ، شدت کنش افتراقی (DIF) یک درصد سؤال‌های کنش افتراقی (DIF) ده درصد (سه سؤال) و تعداد دو گزینه برای هر سؤال، در جدول شمارهٔ (۱) ملاحظه می‌شود.

جدول (۱) نتایج تحلیل رگرسیون لوجستیک برای داده‌هایی با کنش افتراقی (DIF) هماهنگ ، شدت کنش افتراقی (DIF) یک درصد سؤال‌های ده درصد (سه سؤال) و تعداد دو گزینه برای هر سؤال

آماره / منبع	مقدار برآورد	خطای استاندارد برآورد	مقدار والد	درجه آزادی	سطح معنی‌داری	توان نمایی برآورد[1]
گروه	۰/٤٥۰	۰/۱۳۱	۱۱/۷٦۸	۱	۰/۰۰۱	۱/٥٦۸
نمره کل آزمون	۰/۱۷٥	۰/۲٥	٤۷/٤۷۸	۱	۰/۰۰۰	۱/۱۹۱
مقدارثابت	-۳/٤۲۲	۰/٤۲۷	٦٤/۳۰۸	۱	۰/۰۰۰	۰/۰۳۳

برای تحلیل فوق، نتایج تحلیل رگرسیون لوجستیک نشان می‌دهد که در سطح معنی‌داری ۰/۰۰۱ گروه مرجع احتمال بیشتری برای دادن پاسخ درست به سؤال یک آزمون دارد (P<۰/۰۱، b=۰/٤٥). همچنین مقدار ثابت معادله رگرسیون لوجستیک (P<۰/۰۱، α= -۳/٤۲) و نمره کل آزمون (P<۰/۰۱، b= ۰/۱۷٥) در سطح ۰/۰۰۱ معنی‌دار است. مقدار توان نمایی برآورد برای متغیر گروه که شامل گروه مرجع و گروه کانونی است، نشان می‌دهد که بخت گروه مرجع برای دادن پاسخ درست به سؤال یک آزمون بیش از یک و نیم برابر بخت گروه کانونی است که این شواهد گواهی بر وجود کنش افتراقی (DIF) در سؤال یک آزمون است. نمره کل آزمون از جمع نمرات سی سؤال آزمون تشکیل شده است و در معادله رگرسیون لوجستیک تأثیر آن بر احتمال پاسخ به سؤال کنترل می‌شود. مقدار ثابت نیز در معادله رگرسیون لوجستیک نشان‌دهنده برآورد خام احتمال پاسخ به سؤال است، در صورتی که هیچ یک از متغیرهای گروهی و نمره کل آزمون قدرت پیش‌بینی احتمال پاسخ به سؤال را نداشته باشند. این مدل برای بررسی کنش افتراقی (DIF) هماهنگ مدل کامل فرض می‌شود. بنابراین معادله رگرسیون لوجستیک به شرح زیر خواهد بود:

۱. Exp(B)

۰/۱۷۵ +گروه ۰/٤٥۰+ ۳/٤۲۲- =[(پاسخ درست= سؤال۱ آزمون)P] لوجیت: مدل کامل نمره کل آزمون

معادله فوق بدین معنی است که لوجیت احتمال پاسخ درست به سؤال یک آزمون عبارت است از: تأثیر گروه بر لگاریتم بخت پاسخ درست به سؤال یک آزمون در حالی که تأثیر نمره کل آزمون بر احتمال پاسخ درست به سؤال کنترل شده است.

جدول شمارهٔ (۲) نشان دهنده درصد پیش‌بینی پاسخ‌ها توسط مدل رگرسیون لوجستیک است. هر چقدر درصد پیش بینی پاسخ‌ها برای سؤال یک آزمون بیشتر باشد، مدل رگرسیون لوجستیک توان آماری بیشتری دارد. جدول شمارهٔ (۲) نشان می‌دهد که مدل رگرسیون لوجستیک توانسته است حدود ٦۲ درصد پاسخ‌ها برای سؤال یک آزمون را پیش‌بینی کند.

جدول (۲) طبقه‌بندی پیش‌بینی‌های درست رگرسیون لوجستیک

مشاهده شده		پیش بینی شده		
		سؤال ۱		
		پاسخ غلط	پاسخ درست	درصد درست
سؤال ۱	پاسخ غلط	۳٦٤	۱۷۰	٦۸/۲
	پاسخ درست	۲۱۳	۲٥۳	٥٤/۳
درصد کل پیش‌بینی درست		٦۱/۷		

همچنین جدول خلاصه مدل شمارهٔ (۳) مقدار واریانس پیش‌بینی شده مدل توسط متغیر گروهی و نمره کل آزمون را نشان می‌دهد.

جدول (۳) خلاصه مدل رگرسیون لوجستیک برای پیش‌بینی پاسخ سؤال ۱ آزمون

مقدار ۲- برابر لگاریتم بخت	ضریب تبیین کاکس و اسنل[1]	ضریب تبیین ناگلکرک[2]
۱۳۱٤/٦۳۲	۰/۰٦٥	۰/۰۸۷

نتایج جدول شمارهٔ (۳) نشان می‌دهد که حدود هفت تا نُه درصد احتمال پاسخ برای سؤال یک آزمون توسط متغیر گروهی و نمره کل آزمون پیش‌بینی می‌شود. تفاوت ضرایب تبیین ملاحظه شده در جدول شمارهٔ (۳) در این است که ضریب تبیین

۱. Cox & Snell R Square

۲. Nagelkerke R Square

کاکس و اسنل ضریب تبیین نمونه و ضریب ناگلکرک مقدار تبیین‌شده آن برای جامعه است. البته این مقادیر ضریب تبیین نباید مانند مقادیر مجذور همبستگی چندگانه برای رگرسیون چندگانه تفسیر شوند، زیرا این مقادیر معمولاً بسیار کوچک هستند. مقدار (۲-) برابر لگاریتم بخت نشان‌دهنده میزان برازش مدل است. هر چقدر این مقدار به صورت نسبی کمتر باشد، مدل رگرسیون لوجستیک برای پیش‌بینی احتمال پاسخ سؤال یک آزمون برازش بهتری دارد. مقدار (۲-) برابر لگاریتم بخت نقش کلیدی در مورد قضاوت در مورد وجود کنش افتراقی (DIF) سؤال مورد بررسی دارد. این مقدار که در مورد مدل کامل است باید از مقدار (۲-) برابر لگاریتم بخت مدل کاهش‌یافته[1] کم شود. مقدار تفاوت بین مدل کامل و مدل کاهش‌یافته با یک درجه آزادی از توزیع خی دو تبعیت می‌کند و مقدار معنی‌دار آن نشان دهنده کنش افتراقی (DIF) هماهنگ است. معادله محاسبه شده برای مدل کاهش‌یافته در مدل حاضر عبارت است از:

لوجیت :مدل کاهش یافته [P(پاسخ درست= سؤال۱ آزمون)] = ۰/۱۷۹ + نمره کل آزمون ۲/۱۸-

همان‌طور که ملاحظه می‌شود در مدل کاهش یافته متغیر گروهی از معادله رگرسیون لوجستیک حذف شده و محاسبه دوباره انجام شده است. مقدار (۲-) برابر لگاریتم بخت برای معادله کاهش یافته معادل با ۱۳۲٦/٤٥۷ است، که اگر مقدار (۲-) برابر لگاریتم بخت مدل کامل از آن کم شود مقدار ۱۱/۸۲۵ به دست می‌آید، این مقدار با یک درجهٔ آزادی نشان‌دهندهٔ کنش افتراقی (DIF) هماهنگ سؤال مورد بررسی در سطح معنی‌داری ۰/۰۰۱ است. بنابراین مقایسهٔکنش افتراقی (DIF) هماهنگ شبیه‌سازی شده در این تکرار با موفقیت توسط رگرسیون لوجستیک آشکار شد. نتایج آشکارسازی کنش افتراقی (DIF) مثبت درست در هر صد مجموعه داده پاسخ سؤال شبیه‌سازی شده در جدول شمارهٔ (٤) ملاحظه می‌شود.

جدول (٤) درصد آشکارسازی کنش افتراقی (DIF) (از هر صد تکرار) در شرایط مختلف شبیه‌سازی شده

نوع کنش افتراقی (DIF)	شدت کنش افتراقی (DIF)	درصد سؤال‌های کنش افتراقی (DIF)	حجم نمونه		
			۱۵۰	۵۰۰	۱۰۰۰
هماهنگ	۰/۲۵	سه سؤال(۱۰٪)	۷۷	۸۸	۹۷

۱. reduced model

نوع کنش افتراقی (DIF)	شدت کنش افتراقی (DIF)	درصد سؤال‌های کنش افتراقی (DIF)	حجم نمونه		
			۱۵۰	۵۰۰	۱۰۰۰
		شش سؤال(۲۰٪)	۸۳	۸۶	۹۵
		نه سؤال(۳۰٪)	۷۵	۸۹	۹۷
	۰/۵	سه سؤال(۱۰٪)	۸٤	۹۳	۱۰۰
		شش سؤال(۲۰٪)	۸۲	۹٤	۱۰۰
		نه سؤال(۳۰٪)	۸۷	۹۰	۹۸
	۰/۷۵	سه سؤال(۱۰٪)	۸۵	۹٤	۱۰۰
		شش سؤال(۲۰٪)	۸۷	۹۲	۱۰۰
		نه سؤال(۳۰٪)	۷۸	۹۱	۱۰۰
	۱	سه سؤال(۱۰٪)	۹۱	۱۰۰	۱۰۰
		شش سؤال(۲۰٪)	۸۹	۱۰۰	۱۰۰
		نه سؤال(۳۰٪)	۹۰	۹۸	۱۰۰
ناهماهنگ	۰/۲۵	سه سؤال(۱۰٪)	۷۵	۸٤	۹۱
		شش سؤال(۲۰٪)	۷۹	۷۱	۹۳
		نه سؤال(۳۰٪)	۷۳	۸۷	۹٤
	۰/۵	سه سؤال(۱۰٪)	۸۰	۸۶	۹۷
		شش سؤال(۲۰٪)	۷۷	۹۲	۹۹
		نه سؤال(۳۰٪)	۸۵	۹۰	۹۸
	۰/۷۵	سه سؤال(۱۰٪)	۸۱	۹۱	۱۰۰
		شش سؤال(۲۰٪)	۸۳	۸۸	۱۰۰
		نه سؤال(۳۰٪)	٦۹	۸۹	۱۰۰
	۱	سه سؤال(۱۰٪)	۸۹	۹٦	۱۰۰
		شش سؤال(۲۰٪)	۸٦	۹۵	۱۰۰
		نه سؤال(۳۰٪)	۷۷	۸۸	۱۰۰

همان‌طور که در جدول شمارهٔ (٤) ملاحظه می‌شود، در همه شرایط آزمایشی، حجم نمونه به صورت خطی باعث افزایش نرخ آشکارسازی کنش افتراقی (DIF) می‌شود. در حجم‌های نمونه مختلف نتایج به شرح زیر است.

حجم نمونه هزار آزمودنی: با این حجم نمونه، رگرسیون لوجستیک در مقادیر DIF با شدت ۰/۷۵ و ۱ می‌تواند صد درصد انواع کنش افتراقی (DIF) هماهنگ و ناهماهنگ را مشخص کند. هنگامی که کنش افتراقی (DIF) هماهنگ و شدت کنش

افتراقی (DIF) ۰/۵ است، به غیر از یک مورد (سی درصد سؤال‌ها کنش افتراقی (DIF) هستند) رگرسیون لوجستیک می‌تواند صد درصد شرایط کنش افتراقی (DIF) سؤال را تشخیص دهد. در چنین شرایطی با در نظر گرفتن نوع کنش افتراقی (DIF) ناهماهنگ ، رگرسیون لوجستیک در هیچ یک از شرایط نمی‌تواند صد درصد موارد کنش افتراقی (DIF) را آشکار کند، اما در عین حال هنوز نرخ آشکارسازی بیشتر از ۹۶ درصد است. این نتیجه نشان می‌دهد که رگرسیون لوجستیک در حجم نمونه هزار آزمودنی کنش افتراقی (DIF) هماهنگ را بهتر از کنش افتراقی (DIF) ناهماهنگ تشخیص می‌دهد. در شرایطی که شدت کنش افتراقی (DIF) ۰/۲۵ است، رگرسیون لوجستیک برای هر دو حالت کنش افتراقی (DIF) هماهنگ و ناهماهنگ بیش از نود درصد موارد کنش افتراقی (DIF) را تشخیص می‌دهد. این در حالی است که در چنین شرایطی همچنان رگرسیون لوجستیک کنش افتراقی (DIF) هماهنگ را بیشتر از کنش افتراقی (DIF) ناهماهنگ آشکار می کند. همچنین درصد سؤال‌های کنش افتراقی (DIF) تأثیر قابل مشاهده‌ای نداشت.

حجم نمونه پانصد آزمودنی: با این حجم نمونه همچنان رگرسیون لوجستیک کنش افتراقی (DIF) هماهنگ را بیشتر از کنش افتراقی (DIF) ناهماهنگ تشخیص می‌دهد. با افزایش شدت کنش افتراقی (DIF) نرخ تشخیص کنش افتراقی (DIF) نیز توسط رگرسیون لوجستیک افزایش می‌یابد. همچنین درصد سؤال‌های کنش افتراقی (DIF) تأثیر قابل مشاهده‌ای نداشت.

حجم نمونه ۱۵۰ آزمودنی: با این حجم نمونه نیز رگرسیون لوجستیک تا حدودی کنش افتراقی (DIF) هماهنگ را بیشتر از کنش افتراقی (DIF) ناهماهنگ تشخیص می‌دهد. همچنین در شدت کنش افتراقی (DIF) ۰/۲۵ و ۰/۵۰ نمی‌توان در هیچ یک از شرایط آزمایشی صد درصد موارد کنش افتراقی (DIF) را تشخیص داد، بنابراین در چنین شرایطی محقق باید با احتیاط نتایج را تفسیر کند. همچنان درصد سؤال‌های کنش افتراقی (DIF) تأثیر قابل مشاهده‌ای در نرخ آشکارسازی کنش افتراقی (DIF) نداشت.

بنابراین نتایج می توان اذعان داشت که وقتی همه عوامل مداخله‌گر کنترل شوند، به غیر از درصد سؤال‌های کنش افتراقی (DIF)، همچنان سایر عوامل (حجم نمونه، نوع کنش افتراقی (DIF)، شدت کنش افتراقی (DIF) بر نرخ آشکارسازی کنش افتراقی (DIF) تأثیر دارند.

بحث و نتیجه‌گیری

تغییرناپذیری اندازه‌گیری اهمیت زیادی دارد زیرا یکی از شرایط تعمیم‌پذیری علمی است، اما اثبات وجود آن نیازمند شواهد محکمی است (انگلهارد[۱]، هنسچه[۲]و راتلج[۳]، ۱۹۹۰). در این‌باره روش‌های متعددی برای آشکارسازی تغییرناپذیری اندازه‌گیری معرفی شده است که هر کدام دارای شرایطی هستند. برخی از پارامترهای انتخاب روش بررسی کنش افتراقی (DIF) عبارتند از: الف- قابلیت آشکارسازی کنش افتراقی (DIF) هماهنگ یا ناهماهنگ، ب- قابلیت کار با داده‌های دو وجهی ، چند وجهی یا ترکیبی، و ج- قابلیت کار با حجم‌های متفاوت نمونه است. هر یک از روش‌های بررسی کنش افتراقی (DIF) دارای مزایا و محدودیت‌هایی هستند، مثلاً روش‌های پیچیده‌تر بررسی کنش افتراقی (DIF) نیازمند حجم‌های نمونه بزرگ‌تر هستند. به طور مثال الدر[٤]، مک نامارا[٥] و کانگدون (۲۰۰۳) با نمونه‌ای ۱۳۹ نفری از آزمودنی‌ها دریافتند که تحلیل کنش افتراقی (DIF) مبتنی بر نظریهٔ سؤال پاسخ (IRT) با یک مدل سه پارامتری دارای توان آماری کمی است. کانولی[٦] (۲۰۰۳) در مطالعه خود از روش نظریهٔ سؤال پاسخ (IRT) برای تحلیل کنش افتراقی (DIF) استفاده نکرد، زیرا نمی‌خواست محدودیت‌های غیرضروری به مدل نظریهٔ سؤال پاسخ (IRT) اعمال کند. به جای آن از رگرسیون لوجستیک استفاده کرد که برای برآورد دقیق به حجم‌های نمونه کوچک تر نیاز دارد. اما هنوز مانند سایر روش‌های نمره مشاهده شده بررسی کنش افتراقی (DIF) یک محدودیت وجود دارد. این محدودیت همان استفاده از نمره

---

۱. Englehard

۲. Hansche

۳. Rutledge

٤. Elder

٥. Mc Namara

٦. Conoley

کل به عنوان ملاک همتای درونی است، که به عنوان پیش‌بینی کننده بدون سوگیری[۱] از توانایی آزمودنی در نظر گرفته می‌شود.

با توجه به محدودیت‌های سایر روش‌های بررسی کنش افتراقی (DIF)، روش رگرسیون لوجستیک در تحقیق حاضر مورد مطالعه قرار گرفت و کارکرد این روش با استفاده از روش‌شناسی مطالعات شبیه‌سازی داده‌ها مطالعه شد. یافته‌های تحقیق نشان می‌دهد که تحلیل رگرسیون لوجستیک می‌تواند به‌طور متوسط ۹۰/۱۸ درصد کنش افتراقی (DIF) شبیه‌سازی شده را آشکار کند. هیدلاگو و لوپز- پینا (۲۰۰٤) با استفاده از روش رگرسیون لوجستیک به‌طور متوسط توانستند ۷۸/٤۲ درصد سؤال‌های کنش افتراقی (DIF) را به درستی تشخیص دهند. نتایج این مطالعات با نتایج تحقیق حاضر از نظر نرخ متوسط آشکارسازی کنش افتراقی (DIF) در شرایط شبیه‌سازی شده تقریباً همخوانی دارد. همچنین جودین و گیرل (۲۰۰۱) در مطالعه خود توانستند کنش افتراقی (DIF) هماهنگ را به‌طور متوسط در ۷۵/۳۰ درصد تکرارها در شرایط شبیه‌سازی شده آشکار کنند. نتیجه تحقیق حاضر در همخوانی با پژوهش جودین و گیرل (۲۰۰۱) نشان می‌دهد که با وجود کنترل عواملی شامل حجم نمونه، نوع کنش افتراقی (DIF)، و شدت آن و درصد سؤال‌های دارای کنش افتراقی (DIF) هماهنگ همچنان تقریباً به یک اندازه آشکار می‌شود.

همچنین نتایج کاربرد تحلیل رگرسیون لوجستیک در تشخیص کنش افتراقی (DIF) نشان داد که حجم نمونه، نوع کنش افتراقی (DIF) و شدت کنش افتراقی (DIF) بر نرخ آشکارسازی کنش افتراقی (DIF) مؤثر هستند، اما درصد سؤال‌های کنش افتراقی (DIF) آزمون بر نرخ آشکارسازی کنش افتراقی (DIF) تأثیری ندارد. یافته‌های میزور و همکاران (۱۹۹۲) و سانتانا (۲۰۰۹) در مورد تأثیر حجم نمونه با نتایج تحقیق حاضر همخوانی دارند. اما نتایج سانتانا (۲۰۰۹) در مورد درصد سؤال‌های کنش افتراقی (DIF) آزمون با نتایج تحقیق حاضر همخوانی ندارد. این تفاوت ممکن است به دلیل کنترل سایر متغیرها (حجم نمونه، نوع کنش افتراقی (DIF)، شدت کنش افتراقی (DIF)) و مطالعه هم‌زمان این متغیرها در یک مطالعه واحد در تحقیق حاضر باشد. میزور و همکاران (۱۹۹۲) معتقدند که عواملی نظیر مقدار کنش افتراقی (DIF) و دشواری سؤال بر کارکرد روش رگرسیون لوجستیک در

۱. unbiased

تشخیص کنش افتراقی (DIF) سؤال مؤثر است. همچنین مطالعات قبلی نشان می‌دهند که روش رگرسیون لوجستیک به حجم نمونه حساس است و در نمونه‌های کوچک برآورد دقیق پارامترها میسر نیست (اگرستی، ۲۰۰۷). نتایج تحقیق حاضر چنین گزاره‌ای را کاملاً مورد حمایت قرار می دهد.

یافته‌های نارایانان و سوامیناتان (۱۹۹٦) نشان می دهد که درصد سؤال‌های کنش افتراقی (DIF) با عملکرد رگرسیون لوجستیک رابطه معنی‌داری دارد. نتایج آنها نشان می‌دهد که با افزایش درصد سؤال‌های دارای سوگیری در آزمون، توان آماری آزمون کاهش و میزان خطای نوع اول افزایش می‌یابد. اما یافته‌های تحقیق حاضر رابطه خطی معنی داری میان درصد سؤال‌های کنش افتراقی (DIF) با افزایش توان آماری تحلیل رگرسیون لوجستیک نشان نمی‌دهد. این تفاوت می تواند احتمالاً ناشی از این باشد که در تحقیق حاضر علاوه بر درصد سؤال‌های دارای سوگیری در آزمون، حجم نمونه و نوع کنش افتراقی (DIF) نیز کنترل شده است. همچنین نتایج تحقیق حاضر نشان می‌دهد که توان آماری آزمون رگرسیون لوجستیک در آشکارسازی کنش افتراقی (DIF) ناهماهنگ کمتر است. این یافته با نتایج هیررا (۲۰۰۵) همخوانی دارد، اما برخلاف این یافته آنان است که رگرسیون لوجستیک به طور کلی در حجم‌های نمونه مختلف توان آماری کمی دارد. راجرز و سوامیناتان (۱۹۹۳) معتقدند که این امر به این دلیل است که آزمون رگرسیون لوجستیک برای آشکارسازی کنش افتراقی (DIF) ناهماهنگ یک پارامتر اضافه به مدل می‌افزاید که باعث از دست رفتن یک درجه آزادی شده و توان آماری آزمون کاهش می یابد.

به طور کلی بر اساس یافته‌های تحقیق حاضر می‌توان گفت که با کنترل عواملی نظیر حجم نمونه، نوع کنش افتراقی (DIF)، تعداد سؤال‌های کنش افتراقی (DIF) و شدت کنش افتراقی (DIF)، برخی از گمانه زنی‌های قبلی در مورد توان آماری آزمون رگرسیون لوجستیک در آشکارسازی کنش افتراقی سؤال‌های آزمون ناهمخوان است، اما هنوز این روش به طور متوسط دارای کارکردی بسیار مطلوب (بیش از نود درصد) در آشکارسازی کنش افتراقی (DIF) است.

بر اساس نتایج تحقیق حاضر پیشنهاد می شود که از روش رگرسیون لوجستیک بیشتر برای تشخیص کنش افتراقی (DIF) هماهنگ و تا حد امکان در حجم‌های نمونه بسیار بزرگ استفاده شود. در ضمن این قضیه باید مورد تأکید قرار گیرد که آشکارسازی کنش افتراقی (DIF) فرایندی آماری است که در آن سؤال‌هایی که

سوگیری دارند با روش‌شناسی خاصی مشخص شده است و برای آن اندازه اثر[۱] محاسبه می‌شود. البته فرایند قضاوت در مورد سوگیری سؤال در این نقطه به پایان نمی‌رسد، بلکه مطالعات کیفی و اکتشافی بعد از مرحله تحلیل آماری کنش افتراقی (DIF) شروع می شود. الدر[۲] و همکاران (۲۰۰۳) در مطالعه‌ای بعد از حذف سؤال‌های مظنون به سوگیری دریافتند که پاکسازی آزمون از این گونه سؤال‌ها به سود آزمون است، اما این نکته نیز قابل توجه است که حذف سؤال‌های دارای سوگیری بر پایایی آزمون تأثیر نمی‌گذارد، ولی بر اعتبار محتوایی و ساختاری آزمون تأثیر می گذارد. به‌ویژه اگر سؤال‌های زیادی از آزمون حذف شوند باعث کمرنگ شدن[۳] سازه مورد بررسی و ویران شدن جدول مشخصات آزمون می‌شود.

قضاوت نهایی در مورد سوگیری سؤال بر اساس نظر گروهی از متخصصان سازه نظری مورد بررسی یا موضوع درسی و طراحی سؤال صورت می‌گیرد. کنش افتراقی (DIF) معمولاً به خاطر اثر چند بعدی بودن[٤] به‌وجود می آید که به آن مؤلفه مزاحم ثانوی گفته می‌شود. بنابراین اگر تدوین آزمونی یک بعدی[٥] مد نظر است، باید استفاده از سؤال‌های دارای سوگیری پرهیز می‌شود. (شیلی[٦] و استات[۷]، ۱۹۹۳). سرانجام باید متذکر شد که مهم‌ترین نگرانی بعد از آشکارسازی سؤال‌های کنش افتراقی (DIF)، تصمیم‌گیری در مورد سؤال‌های دارای سوگیری است. در این شرایط ممکن است سؤال‌های دارای سوگیری از آزمون حذف نشوند، اصلاح شوند یا شکل‌های متفاوتی از آزمون برای گروه‌های مرجع و کانونی اجرا شود. بنابراین تصمیم‌گیری و قضاوت کیفی صحیح در مورد سؤال‌های آزمون بعد از آشکارسازی کمی کنش افتراقی (DIF) می‌تواند عاقبت درستی برای سؤال‌های آزمون رقم بزند.

---

۱. effect size

۲. Elder

۳. underrepresenting

٤. multidimesionality

٥. unidimesional

٦. Shealy

۷. Stout

منابع

- رضایی، عباسعلی و شعبانی، عنایت‌الله (۱۳۸۹). تحلیل کارکرد افتراقی جنسیتی آزمون سنجش توانش عمومی زبان دانشگاه تهران. *مجلهٔ پژوهش‌های زبان خارجی، شمارهٔ ٥٦.*
- گرامی‌پور، مسعود و فلسفی‌نژاد، محمدرضا (۱۳۹۲). *روش‌های آماری بررسی کنش افتراقی سؤال(DIF) در آزمون‌های سرنوشت‌ساز*. تهران: انتشارات جهاد دانشگاهی واحد تربیت معلم.

- Agresti, A. (2007). *an introduction to categorical data analysis*. New York: Wiley Interscience.
- Byrne, B. M. & Stewart, S. M. (2006). The MACS approach to testing for multigroup invariance of a second-order structure: A walk through the process. *Structural Equation Modeling, 13*: 287-321.
- Camilli, G. & Congdon, P. (1999). Application of a method of estimating DIF for polytomous test items. *Journal of Educational and Behavioral Statistics* 24: 323–341.
- Camilli, G. & Shepard, L. A. (1994). *Methods for identifying biased test items*. Thousand Oaks, CA: Sage Publications.
- Clauser, B. & Mazor, K. (1998). Using statistical procedures to identify differentially functioning test items. *Educational Measurement: Issues and Practice, 17*(1): 31–44.
- Conoley, C. A. (2003). *Differential item functioning in the Peabody Picture Vocabulary Test – Third Edition: Partial correlation versus Expert judgment*. Unpublished doctoral dissertation, Texas A&M University, TX
- Drasgow, F. (1984). Scrutinizing psychological tests: Measurement equivalence and equivalent relations with external variables are central issues. *Psychological Bulletin, 95*: 135-135.
- Elder, C.; Mc Namara, T. & Congdon, P. (2003). Rasch techniques for detecting bias in performance tests: An example comparing the performance of native and non-native speakers on a test of academic English. *Journal of Applied Measurement*, 4: 181–197.
- Elosua, P. & Wells, C. S. (2013). Detecting DIF in Polytomous Items Using MACS, IRT and Ordinal Logistic Regression. *Psicológica*, 34: 327-342.
- Embretson, S. E. & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates.

- Englehard, G.; Hansche, L. & Rutledge, K. E. (1990). Accuracy of bias review judges in identifying differential item functioning on teacher certification tests. *Applied Measurement in Education*, 3**:** 347–360.
- Flowers, C. P.; Oshima, T. C. & Raju, N. S. (1999). A description and demonstration of the polytomous-DFIT framework. *Applied Psychological Measurement, 23*: 309–326.
- Han, Kyung T. & Hambleton, Ronald K. (2007). *User's Manual for WinGen: Windows Software that Generates IRT Model Parameters and Item Responses.* Center for Educational Assessment Research. Amherst, MA: University of Massachusetts, Center for Educational Assessment.
- Harwell, M.; Stone, C. A.; Hsu, T. C & Kirisci, L. (1996). Monte Carlo studies in item response theory. *Applied Psychological Measurement*, 20: 101-125.
- Herrera A. N. (2005). *Sample size effect and rate of sample sizes to detect differential item functioning, Doctoral thesis*, university of Barcelona, Barcelona (Spain).
- Hidalgo, M. D. & López-Pina, J. P. (2004). Differential item functioning detection and effect size: A comparison between logistic regression and Mantel Haenszel procedures. *Educational and Psychological Measurement*, 64: 903–915.
- 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: 329–349.
- Lord, F. M. & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.
- Millsap, R. E (2011). *Statistical Approaches to Measurement Invariance*. New York: NY, Routledge
- Narayanan, P. & Swaminathan, H. (1996). Identification of items that show nonuniform DIF. *Applied Psychological Measurement*, 20: 257-274.
- Parshall, C. G. & Miller, T. R. (1995). Exact versus asymptotic Mantel-Haenszel DIF statistics. *Journal of Educational Measurement*, 32 (3): 302–316.
- Penfield, R. D. & Algina, J. (2003). Applying the Liu–Agresti estimator of the cumulative common odds ratio to DIF detection in polytomous items. *Journal of Educational Measurement*, 40: 353–370.
- Raju, N. S.; Laffitte, L. J. & Byrne, B. M. (2002). Measurement equivalence: A comparison of methods based on Confirmatory Factor Analysis and item response theory. *Journal of Applied Psychology*, 87: 517–529.

- Reise, S. P.; Widaman, K. F. & Pugh, R. H. (1993). Confirmatory Factor Analysis and item response theory: Two approaches for exploring measurement invariance. *Psychological Bulletin*, 114: 552-566.
- Rogers, H. J. & Swaminathan, H. (1993). A comparison of the logistic regression and Mantel-Haenszel procedures for detecting differential item functioning. *Applied Psychological Measurement*. 17: 105–116.
- Santana, A. C. (2009). *Effect of the ratio of sample sizes to detect differential items dunctioning through logistic regression procedure*, Master thesis, National University of Colombia, Bogotá (Colombia).
- Shealy, R. T.; Stout, W. F. (1993). A model based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DTF as well as item bias/DIF. *Psychometrika*, 58: 197–239.
- Spence, I. (1993). Monte Carlo simulation studies. *Applied Psychological Measurement*, 7: 405-425
- Swaminathan, H. & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement*, 27: 361– 370.
- Su, Y. -H. & Wang, W. C. (2005). Efficiency of the Mantel, generalized Mantel-Haenszel, and logistic discriminant function analysis methods in detecting differential item functioning for polytomous items. *Applied Measurement in Education*, 18: 313–350.
- Vandenberg, R. J. (2002). Toward a further understanding of and improvement in measurement invariance methods and procedures. *Organizational Research Methods*, 5*:* 139–158.
- Van der Linden, W. J. & Hambleton, R. K. (1997). *Handbook of modern item response theory*. New York: Springer-Verlag.
- Zumbo, B. D. (1999). *A Handbook on the theory and methods of differential item functioning (DIF): Logistic regression modeling as a unitary framework for binary and Likert-type (ordinal) item scores*. Ottawa, ON: Directorate of Human Resources Research and Evaluation, Department of National Defense.
- Zwick, R.; Thayer, D. T. & Lewis, C. (1999). An empirical Bayes approach to Mantel–Haenszel DIF analysis. *Journal of Educational Measurement*, 36: 1–28.