مقایسهٔ روش‌های تحلیل عاملی تأییدی (CFA) و نسبت درستنمایی مبتنی بر مدل پرسش-پاسخ(IRT) در ردگیری کنش افتراقی سؤالات آزمون‌های سرنوشت ساز[1]

مسعود گرامی‌پور[2]
محمدرضا فلسفی‌نژاد[3]
علی دلاور[4]
نورعلی فرخی[5]

چکیده

زمینه: اگر چه روش‌های متعددی جهت شناسایی سؤالات سودار معرفی شده است، اما تحقیقات اندکی به صورت تجربی قدرت و کارایی هر یک این روش‌ها را مورد بررسی قرار داده است. **هدف:** هدف از تحقیق حاضر مقایسه آزمون نسبت درستنمایی مبتنی بر مدل پرسش-پاسخ (IRT) و روش تحلیل عاملی تأییدی(CFA) در شناسایی کنش افتراقی سؤالاتِ آزمون‌های سرنوشت ساز بود. **روش:** برای پاسخگویی به سؤالات تحقیق از روش مطالعات شبیه‌سازی مونت کارلو استفاده شد. داده‌های مورد نیاز با استفاده از نرم افزار وین جن۲ در قالب ۱۰۰ آزمون ۳۰ سؤالی که با مدل دو پارامتری برازش داشتند، شبیه‌سازی شد. توزیع دشواری و قدرت تمیز سؤالات همه آزمون‌ها، نرمال بود. اما آزمون‌ها از حیث نوع DIF، میزان DIF و همچنین پارامترهای سؤال با هم متفاوت بودند.همچنین برای هر آزمون، پاسخهای ۱۰۰۰ آزمودنی با توزیع توانایی(θ) نرمال شبیه‌سازی شد. از روش‌های برآورد بیشینه درستنمایی حاشیه ای و روش برآورد حداقل مربعات وزنی برای تعیین نوع و میزان DIF استفاده شد. **یافته‌ها:** تجزیه و تحلیل داده‌ها در تکرارهای متوالی نشان داد که روش‌های مبتنی بر IRT در شناسایی DIF نسبت به روش‌های CFA برتری دارند. این برتری در تمامی شرایط DIF (کم، متوسط و زیاد) مشاهده شد. با این همه، تفاوت دو روش در حجمهای نمونه ۱۰۰۰ نفری، جزئی و قابل اغماض است. همچنین میان دو روش از نظر شناسایی نوع DIF تفاوتی دیده نشد. **بحث و نتیجه گیری:** نتایج پژوهش حاضر با نتایج مید و لوتنچلاجر(۲۰۰۶ و ۲۰۰۴) همخوانی دارد، اما برخلاف نتایج فلاورز و همکاران(۲۰۰۲) است. سرانجام در صورت محدودیت در استفاده از یک روش تشخیص DIF، آزمون نسبت درستنمایی مبتنی بر مدل پرسش- پاسخ توصیه شود.

واژگان کلیدی: کنش افتراقی سؤال- آزمون نسبت درستنمایی- تحلیل عاملی تأییدی- آزمون‌های سرنوشت ساز

---

۱. مقاله حاضر از رساله دکتری با عنوان " مقایسه قدرت آزمون نسبت درستنمایی مبتنی بر مدل پرسش.پاسخ(IRT) با روش‌های تحلیل عاملی تأییدی و رگرسیون لوجستیک در شناسایی کنش افتراقی سؤالات (DIF) به منظور اطمینان از عادلانه بودن سنجش در آزمون های سرنوشت ساز" استخراج شده است.

۲. دانشجوی دکترای رشته سنجش و اندازه‌گیری(نویسنده مسئول) Email: mgramipour@yahoo.com
۳. عضو هیأت علمی دانشکده روانشناسی و علوم تربیتی دانشگاه علامه طباطبائی
۴. عضو هیأت علمی دانشکده روانشناسی و علوم تربیتی دانشگاه علامه طباطبائی
۵. عضو هیأت علمی دانشکده روانشناسی و علوم تربیتی دانشگاه علامه طباطبائی

مقدمه

آزمون‌های سرنوشت ساز[1] آزمون‌هایی هستند که در مقیاس وسیع اجرا شده و بر اساس نتایج آن‌ها تصمیماتی خطیر در مورد افراد اتخاذ می‌شود. چنین تصمیماتی دارای تبعات مهم شخصی، اجتماعی و سیاسی هستند. بنابراین، اعتبار تصمیم گیری و کارایی مدل‌های سنجش در این مورد از اهمیت فوق العاده ای برخوردار بوده و از سالیان گذشته جزء مهم‌ترین چالش‌های روان سنجی باقی مانده‌اند. یکی از مهم‌ترین تهدیدها برای اعتبار[2] آزمون، سوگیری سؤال[3] یا نامتغیر بودن اندازه گیری[4] است که از آن تحت عنوان کنش افتراقی سؤال(DIF)[5] نیز یاد می‌شود.کنش افتراقی سؤال شرط لازم ولی ناکافی برای سوگیری سؤال است(رامسی[6]، ۱۹۹۳). علی رغم ریشه دار بودن موضوع و فعالیت‌های زیادی که در این مورد صورت پذیرفته است هنوز در مورد روش‌های مطلوب و بهینه تعیین DIF میان صاحب نظران اتفاق نظر وجود ندارد.

شاید مهم‌ترین اتهام وارده به سنجش و مطمئناً اولی‌ترین مسئله برای مردم، سنجش عادلانه[7] است. دعوی این که آزمون‌ها در برابر اقلیت‌های عمومی و نژادی سوگیری دارند، منجر به لوایح قانونی متعدد شده است. آزمون‌ها و حرفه آزمونگری زیر تیزبینی مردم قرار گرفته، و اکنون منتشر کنندگان و بکار برندگان آزمون‌ها باید نشان دهند که آزمون‌هایشان در برابر اقلیت‌ها، عاری از سوگیری هستند (همبلتون و همکاران، ۱۹۹۱، ص ۱۳۰، ترجمه فلسفی نژاد،۱۳۸۹، ص ۱۳۹). از طرفی عادلانه بودن برای آزمون‌های سرنوشت ساز به عنوان آزمون‌هایی که جهت تصمیم‌گیری برای آینده آزمودنی‌ها بکار می‌روند از اهمیت بیشتری برخوردار است و حساسیت بیشتری نسبت به عادلانه بودن سؤال‌های این نوع آزمون‌ها نشان داده می‌شود(پاپهام[8]، ۲۰۰۵).

1 . high stakes tests
2.Validity
3.Item Bias
4. Measurement Invariance
5. Differential Item Functioning
6. Ramsey
7.test fairness
8. Popham

اصطلاح کنش افتراقی سؤال(DIF) به عنوان ملاک تجربی برای وجود یا عدم وجود سوگیری در سؤال مورد استفاده قرار می‌گیرد. در سؤالات دارای DIF، افراد با توانایی یکسان اما متعلق به گروههای متفاوت، از شانس نابرابری برای ارائه پاسخ صحیح برخوردارند. در ادبیات DIF به این گروه‌ها، گروههای مرجع و کانونی[1] می‌گویند. گروه مرجع، گروه اکثریت و گروه کانونی گروه اقلیت یا محروم در نظر گرفته می‌شوند(سوامیناتان[2] و راجرز[3]، ۱۹۹۰). از نظر آماری اگر با کنترل توانایی در دو گروه، هنوز آزمودنی‌ها از احتمال یا بخت متفاوتی در پاسخ درست به سؤال برخوردار باشند آن سؤال دارای کنش افتراقی است. البته، صرفاً تفاوت در توانایی نشانه سوگیری سؤال نیست. چرا که در مواقعی، آزدمونی ها واقعاً با هم متفاوتند و طبیعی است که از شانس نابرابری در دادن پاسخ صحیح داشته باشند.

شناسایی و کنترل سؤالات سودار گامی مهم در سنجش علمی است. سوداری سؤال خطای نظامدار وارد نتایج سنجش می‌کند و آن نیز به نوبه خود سبب بی اعتباری نتایج آزمون می‌شود (کمیلی[4] و شپارد[5]،۱۹۹۴). لذا، روش‌های آماری مختلفی جهت تحلیل و تعیین DIF ارائه شده که می‌توان آن‌ها را در دو دسته کلی روش‌های مبتنی بر نظریه پرسش-پاسخ(IRT[6]) و روش‌های غیر مبتنی بر نظریه پرسش-پاسخ طبقه بندی کرد. روش مقایسه پارامترهای سؤال و مساحت میان منحنی‌های ویژگی سؤال(ICC[8]) و روش نسبت درستنمایی (همبلتون[9] و همکاران،۱۹۹۱) از جمله روش‌های مبتنی بر IRT است. همچنین از روش‌های تحلیل عاملی و تحلیل داده‌های طبقه ای می‌توان به عنوان روش‌های غیر IRT تعیین سوگیری سؤال یاد کرد(گومز[10] و ناواس[11]، ۲۰۰۰).

1. reference and focal groups
2. Swaminathan
3 . Rogers
4. Camilli
5 . Shepard
6. Item Response Theory
7 .Confirmatory Factor Analysis
8. Item Characteristic Curve
9. Hambleton
10. Gomez
11 . Navas

در آزمون نسبت درستنمایی (تیشن[1]،استینبرگ[2] و جرارد[3]،۱۹۸۶ ؛ تیشن، استینبرگ و وینر[4]،۱۹۸۸،۱۹۹۳) ، برازش مدل تکمیلی[5] A با برازش مدل فشرده[6] C مقایسه می‌شود. برای این کار، در مدل تکمیلی به پارامترهای سؤال در گروههای مورد مقایسه اجازه تغییر داده می‌شود اما، پارامترها در مدل فشرده محدود و مقید می‌شوند. مقدار درستنمایی در مدل فشرده در یک نوبت برآورد شده و سپس درستنمایی مدل افزایشی A محاسبه می‌شود. سپس در مورد قبول یا عدم قبول وجود DIF بر اساس مقایسه میزان درستنمایی مدل‌ها تصمیم گیری می‌شود.

در مقابل، در روش تحلیل عاملی تأییدی برای تعیین سوداری سؤالات، همسانی میان بارهای عاملی و مقادیر ثابت مدل به صورت مستقیم آزمون می‌شود. به این ترتیب که دو مدل خط پایه و مدل محدود شده تعریف می‌شود. در مدل خط پایه به همه پارامترها بجز آنهایی که برای تعیین مدل ثابت نگه داشته می‌شوند، اجازه تغییر آزادانه داده می‌شود. در حالیکه در مدل محدود شده بارهای عاملی یا مقادیر ثابت با اعمال قیدهایی در گروههای مرجع و کانونی مساوی در نظر گرفته می‌شوند. از آماره خی دو برای مقایسه دو مدل استفاده می‌شود. از آنجایی که هر کدام از مدلهای محدود شده در درون مدلهای خط پایه قرار گرفته‌اند می‌توان تغییر در مقدار خی دو را با توجه به خط پایه محاسبه کرد. برای هر مقایسه یک مقدار معنادار از نظر آماری گواهی بر کنش افتراقی سؤال است(سوربوم[7]،۱۹۷۴)

قرابت و نزدیکی مبانی فنی و آماری روش‌های فوق تحقیقات متفاوتی را بر انگیخته تا نتایج این روش‌ها را با هم مقایسه کنند. برای مثال، کیم[8] و همکاران(۱۹۹۶) در یک مطالعه شبیه‌سازی شده مونت کارلو برای آزمون‌های سرنوشت ساز کارکرد روش مبتنی بر IRT

1. Thissen
2. Steinberg
3. Gerrard
4. Wainer
5. augmented
6. compact
7. Sorbom
8 . Kim

نسبت درست نمایی را در خطای نوع اول در DIF مورد بررسی قرار دادند. مدلهای ۲ و ۳ پارامتری IRT برای شبیه‌سازی ۱۰۰ مجموعه ۵۰ سؤالی با نمونه های ۲۵۰ و ۱۰۰۰ آزمودنی مورد استفاده قرار گرفت. پارامترهای سؤال از طریق روش بیشینه درستنمایی حاشیه ای[1] برای سه مدل ۳ پارامتری، مدل ۳ پارامتری با یک پارامتر تثبیت شده و مدل ۲ پارامتری برآورد شدند. همه مقایسه های DIF با مقایسه دو به دو هر کدام از نمونه‌ها و مدلهای IRT شبیه‌سازی شدند. بنابراین برای هر نمونه و مدل IRT ۵۰ جفت گروه مرجع و کانونی وجود داشت. نتایج نشان داد که نرخ خطای نوع اول برای مدل دو پارامتری مطابق با نظریه های موجود است اما برای دو مدل دیگر نتایج کاملاً با نظریات موجود تفاوت معنا داری دارد.

همچنین نتایج مید[2] و همکاران (۲۰۰۴) نشان می‌دهد که روش CFA نسبت به روش نسبت درستنمایی IRT در شناسایی کنش افتراقی سؤال‌هایی که در پارامتر دشواری(b) و تشخیص (a) در گروههای مرجع و کانونی متفاوتند ضعیف‌تر عمل می‌کند. صاحبنظران نیز معتقدند که روش بیشینه درستنمایی CFA در نمونه های کوچک‌تر زمانی که مفروضه نرمال بودن چند متغیره رعایت شده باشد در شناسایی سوگیری مثبت سؤال قدرت بهتری نسبت به سایر روش‌ها دارد(بنتلر[3] ویان[4]،۱۹۹۹). همچنین در یک مطالعه شبیه‌سازی دیگر برای آزمون‌های سرنوشت ساز، استارک[5] و همکاران(۲۰۰۴)، یک راهبرد مشترک برای تشخیص DIF سؤال‌ها ارائه کرده‌اند که در آن هر دو روش CFA و IRT بکار گرفته می‌شوند. در این مطالعه مشخص شد که هر دو روش CFA و IRT در بسیاری از شرایط آزمایشی کارایی دارند و شبیه به یکدیگر عمل می‌کنند.

با این همه مطالعه مید و لوتنچلاجر(۲۰۰۶) نشان داد که روش CFA در شرایط ایده‌آل و حجم نمونه بزرگ و با تعداد کافی نشانگرها، به نتایج دقیقی از DIF سؤالات

1 . Marginal Maximum Likelihood
2 . Meade
3. Bentler
4 . Yaun
5 . Stark

منجر می‌شود. علاوه بر این، در مطالعه ای دیگرمید و باور[1](۲۰۰۷) در یافتند که علاوه بر حجم نمونه ،میزان بیش تعیین شدگی عوامل و میزان اشتراک نشانگرها نیز به صورت قابل ملاحظه ای روی دقت و قدرت CFA در تشخیص DIF تأثیر می‌گذارد.

تحقیقات قبلی در مقایسه روش‌های IRT و CFA در شناسایی سؤالات سودار تنها به کنترل بخشی از عوامل و متغیرهای مؤثر بر DIF پرداخته‌اند. در این میان عوامل مهمی مانند ویژگی‌های روان‌سنجی سؤالات، حجم نمونه و تعاملات احتمالی میان آن‌ها مورد توجه قرار نگرفته است. در مواردی که پژوهشگران به بررسی این عوامل پرداخته‌اند آنها به صورت مجزا مورد توجه قرار گرفته‌اند. این در حالی است که تعامل میان ویژگی‌های روان‌سنجی سؤالات و حجم نمونه از اهمیت بالایی در شناسایی کنش افتراقی سؤالات برخوردار است. به عنوان مثال اگر میزان DIF در سؤالات آزمون از حد خاصی بیشتر باشد حتی روش‌های ضعیف‌تر نیز در گروههای نمونه کوچک قادر به تشخیص این تفاوت‌ها هستند. در صورتیکه اگر میزان DIF کم باشد روش‌های قوی نیز در حجمهای نمونه متوسط از شناسایی آن باز می‌مانند. اهمیت بررسی همزمان این عوامل در آن است که به پژوهشگران امکان می‌دهد تا در شرایط خاص مناسب‌ترین و اقتصادی‌ترین روش را برای بررسیِ DIF سؤال برگزیده و تا حد زیادی به نتایج بدست آمده اعتماد کنند. و تنها در شرایطی روش‌های پر هزینه و پیچیده را انتخاب کنند که راه ارزان‌تر و ساده تری را در اختیار نداشته باشند.

انجام مطالعه ای که در آن عوامل مختلف مورد توجه قرار گیرد، علاوه بر آن که به دانش نظری موجود در مورد ماهیت و کیفیت DIF سؤالات می‌افزاید، قابلیت آن را دارد که از طریق ارائه مثالهای عینی در فرایند کشف DIF سؤالات روش شناسی مناسبی را در اختیار پژوهشگران قرار دهد. آگاهی از روش شناسی مناسب و در دسترس بودن راهبردهای دقیق علمی تعیین سوداری سؤالات از اهمیت ویژه ای در آزمون‌های سرنوشت

1 . Bauer

ساز برخوردار است. از آنجایی که بر اساس نتایج این آزمون‌ها تصمیمات خطیری در مورد افراد اتخاذ می‌شود، مراکز آزمون سازی ناگزیرند که شواهدی مبنی بر دقت نتایج و عدم سوداری این آزمون‌ها ارائه کنند. بدیهی است که در اختیار نبودن روش‌های مناسب فرایند ارائه این شواهد تجربی را با مشکل مواجه می‌کند.

علاوه بر این با توجه به شهرت و محبوبیت روش‌های مبتنی بر نظریه پرسش-پاسخ و تحلیل عاملی تأییدی، انتظار می‌رود پژوهشگران علوم انسانی از این روش‌های قدرتمند در حوزه های جدید و آزمون سازی استفاده کنند. لذا این تحقیق، روش‌های مبتنی بر IRT، روش نسبت درستنمایی[1] و از روش‌های غیر مبتنی بر IRT، روش تحلیل عاملی تأییدی (CFA) را به عنوان مجموعه ای جدید برای پژوهش انتخاب کرده و می‌کوشد تا پاسخ در خور توجهی به این سؤال که" کدامیک از روش‌های آزمون نسبت درستنمایی مبتنی بر مدل پرسش-پاسخ (IRT) و روش تحلیل عاملی تأییدی در بررسی کنش افتراقی سؤال با قدرت بیشتری می‌توانند سوگیری سؤال را در آزمون‌های سرنوشت ساز آشکار کنند؟ و عوامل مداخله گر بر نرخ آشکار سازی DIF در این روش‌ها کدامند؟"

روش پژوهش

برای مقایسه قدرت روش‌های شناسایی DIF در آزمون‌های سرنوشت ساز، از مطالعات شبیه‌سازی موسوم به مطالعات مونت کارلو[2] استفاده شد. روش داده‌های شبیه‌سازی شده مونت کارلو تولید مجموعه داده‌هایی با ویژگی‌های مورد نظر را در محیطی شبیه‌سازی شده ، تحت کنترل و تکرارهای فراوان امکان پذیر می‌کند (جن،۲۰۰۷؛ مید و لوتنچلاجر،۲۰۰۴). برای شبیه‌سازی پارامترهای دشواری و قدرت تمیز برای یک آزمون ۳۰ سؤالی سرنوشت ساز با مدل دو پارامتری IRT از نرم افزار وین جن۲[3] استفاده شد (جن،۲۰۰۷). توزیع پارامتر قدرت تمیز دارای توزیع نرمال با میانگین صفر و انحراف

1. Likelihood Ratio
2 . Monte Carlo
3 . WINGEN

استاندارد ۰/۵ و توزیع پارامتر دشواری دارای توزیع نرمال با میانگین صفر و انحراف استاندارد ۰/۷۵ بود. توزیع مقادیر توانایی نیز دارای توزیع نرمال با میانگین صفر و انحراف استاندارد ۱ بود. توزیع‌های مذکور شبیه به مطالعه پارشال و میلر[1] (۱۹۹۵) برای شبیه‌سازی داده‌های آزمون‌های سرنوشت ساز بود. در این مطالعه برای شبیه‌سازی داده‌های آزمون‌های سرنوشت ساز با بررسی ۲ روش تشخیص DIF مبتنی و غیر مبتنی بر IRT، ۳ حجم نمونه ۱۵۰ تا ۱۰۰۰ آزمودنی، ۲ نوع DIF هماهنگ یا ناهماهنگ ، ۴ مقدار متفاوت DIF ۰.۲۵ تا ۱، ۳ سطح درصد سؤالات دارای DIF ۳ سؤال (۱۰ درصد) تا ۹ سؤال (۳۰ درصد) و ۲ گونه سؤالات دو و پنج گزینه ای، ۴۳۲ شرایط مختلف آزمایشی با ۱۰۰ تکرار شبیه‌سازی شد.

یافته‌های پژوهش

تحلیل DIF با آزمون نسبت درستنمایی IRT در سؤال‌های دو ارزشی و چند ارزشی، با استفاده از مدل دو پارامتری و روش برآورد بیشینه درستنمایی حاشیه‌ای[2] (MML) از نرم‌افزارهای BILOG MG و MUTILOG انجام شد. همچنین تحلیل DIF با روش تحلیل عاملی تأییدی، با به کار گیری روش برآورد حداقل مربعات وزنی[3] و نرم افزار LISREL ۸.۷۲ انجام شد.

همانطور که در جدول ۱ ملاحظه می‌شود، یافته‌های تحقیق نشان می‌دهد که به طور کل نرخ آشکارسازی DIF در روش نسبت درستنمایی(نرخ متوسط ۹۶.۲۲ درصد) بیشتر از تحلیل عاملی تأییدی(نرخ متوسط ۹۳.۶۳ درصد) است. هنگامی که حجم نمونه ۱۰۰۰ آزمودنی است، آزمون‌های نسبت درستنمایی مبتنی بر IRT و تحلیل عاملی تأییدی می‌توانند مقادیر DIF بزرگ‌تر یا مساوی با ۰.۵ را برای پارامتر دشواری(هماهنگ) و پارامتر ضریب تمیز (ناهماهنگ) در ۱۰۰ درصد موارد آشکار کنند. در شرایطی که شدت DIF ۰.۲۵ است روش‌های نسبت درستنمایی بیش از ۹۰ درصد شرایط و تحلیل عاملی

1 . Miller
2 . Marginal Maximum Likelihood
3 . Weighted Least Squares

تأییدی بیش از ۸۲ درصد موارد DIF را تشخیص می‌دهند. نرخ تشخیص DIF در تحلیل عاملی تأییدی در حجم نمونه ۱۰۰۰ آزمودنی تقریباً شبیه آزمون نسبت درستنمایی است. با افزایش شدت DIF نرخ تشخیص DIF در دو روش تشخیص DIF افزایش می‌یابد. نرخ تشخیص DIF برای سؤال‌های دو گزینه ای در آزمون نسبت درستنمایی بیشتر از سؤال‌های پنج گزینه ای است و نرخ تشخیص DIF برای سؤال‌های پنج گزینه ای در تحلیل عاملی تأییدی بیشتر از سؤال‌های دو گزینه ای است.

هنگامی که حجم نمونه ۵۰۰ آزمودنی است، آزمون نسبت درستنمایی می‌تواند DIFهای هماهنگ و ناهماهنگ با شدت ۱ را در ۱۰۰ درصد موارد در شرایط مختلف آزمایشی آشکار کند. در این شرایط تحلیل عاملی تأییدی نمی‌تواند در تمام شرایط آزمایشی ۱۰۰ درصد موارد DIF را آشکار کند اما DIF هماهنگ را کمی بیشتر از DIF ناهماهنگ تشخیص می‌دهد. آزمون نسبت درستنمایی در شدت DIF ۰.۷۵ و هر دو نوع هماهنگ و ناهماهنگ تنها در یک مورد آزمایشی (شرایط ۳۰ درصد سؤال DIF و فرمت سؤال پنج گزینه ای) نمی‌تواند ۱۰۰ درصد موارد DIF را آشکار کند. البته نرخ تشخیص هنوز در این شرایط بالاست (۹۵ درصد موارد). با افزایش شدت DIF، نرخ تشخیص DIF در دو روش تشخیص DIF افزایش می‌یابد. نرخ تشخیص DIF برای سؤال‌های دو گزینه ای در آزمون نسبت درستنمایی بیشتر از سؤال‌های پنج گزینه ای است و نرخ تشخیص DIF برای سؤال‌های پنج گزینه ای در تحلیل عاملی تأییدی بیشتر از سؤال‌های دو گزینه ای است.

در حجم نمونه ۱۵۰ آزمودنی هنگامی که شدت DIF ۰.۲۵ و ۰.۵۰ است نمی‌توان در هیچ یک از شرایط آزمایشی ۱۰۰ درصد موارد DIF را تشخیص داد. در شدت DIF ۰.۷۵ نیز تنها یک مورد در نوع هماهنگ و ناهماهنگ وجود دارد که ۱۰۰ درصد موارد DIF آشکار می‌شود. با افزایش شدت DIF نرخ تشخیص DIF در دو روش تشخیص DIF افزایش می‌یابد. نرخ تشخیص DIF برای سؤال‌های دو گزینه ای در آزمون نسبت درستنمایی بیشتر از سؤال‌های پنج گزینه ای است و نرخ تشخیص DIF برای سؤال‌های پنج گزینه ای در تحلیل عاملی تأییدی بیشتر از سؤال‌های دو گزینه ای است. البته نرخ

تشخیص DIF ناهماهنگ به طور کلی در هر دو روش کمتر از DIF هماهنگ است.

جدول ۱- نرخ آشکار سازی سوگیری سؤال(از ۱۰۰ مرتبه) با حجم نمونه ۱۰۰۰ آزمودنی در شرایط مختلف آزمایشی با استفاده از آزمون‌های تشخیص DIF

نوع DIF	شدت DIF	درصد سؤالات DIF	تعداد گزینه های سؤال	آزمون نسبت درستنمایی IRT	تحلیل عاملی تأییدی
هماهنگ	۰.۲۵	سه سؤال (۱۰٪)	دو گزینه ای	۱۰۰	۸۸
			پنج گزینه ای	۹۲	۹۳
		شش سؤال (۲۰٪)	دو گزینه ای	۹۴	۸۷
			پنج گزینه ای	۹۱	۹۵
		نه سؤال(۳۰٪)	دو گزینه ای	۹۹	۸۹
			پنج گزینه ای	۹۳	۹۵
	۰.۵	سه سؤال (۱۰٪)	دو گزینه ای	۱۰۰	۱۰۰
			پنج گزینه ای	۱۰۰	۱۰۰
		شش سؤال (۲۰٪)	دو گزینه ای	۱۰۰	۱۰۰
			پنج گزینه ای	۱۰۰	۱۰۰
		نه سؤال(۳۰٪)	دو گزینه ای	۱۰۰	۱۰۰
			پنج گزینه ای	۱۰۰	۱۰۰
	۰.۷۵	سه سؤال (۱۰٪)	دو گزینه ای	۱۰۰	۱۰۰
			پنج گزینه ای	۱۰۰	۱۰۰
		شش سؤال (۲۰٪)	دو گزینه ای	۱۰۰	۱۰۰
			پنج گزینه ای	۱۰۰	۱۰۰
		نه سؤال (۳۰٪)	دو گزینه ای	۱۰۰	۱۰۰
			پنج گزینه ای	۱۰۰	۱۰۰
	۱	سه سؤال (۱۰٪)	دو گزینه ای	۱۰۰	۱۰۰
			پنج گزینه ای	۱۰۰	۱۰۰
		شش سؤال (۲۰٪)	دو گزینه ای	۱۰۰	۱۰۰
			پنج گزینه ای	۱۰۰	۱۰۰
		نه سؤال (۳۰٪)	دو گزینه ای	۱۰۰	۱۰۰

نوع DIF	شدت DIF	درصد سؤالات DIF	تعداد گزینه های سؤال	آزمون نسبت درستنمایی IRT	تحلیل عاملی تأییدی
			پنج گزینه ای	۱۰۰	۱۰۰
ناهماهنگ	۰.۲۵	سه سؤال(۱۰٪)	دو گزینه ای	۱۰۰	۸۸
			پنج گزینه ای	۹۲	۹۱
		شش سؤال(۲۰٪)	دو گزینه ای	۹۶	۸۳
			پنج گزینه ای	۹۱	۹۴
		نه سؤال(۳۰٪)	دو گزینه ای	۱۰۰	۸۸
			پنج گزینه ای	۹۵	۹۶
	۰.۵	سه سؤال(۱۰٪)	دو گزینه ای	۱۰۰	۱۰۰
			پنج گزینه ای	۱۰۰	۱۰۰
		شش سؤال(۲۰٪)	دو گزینه ای	۱۰۰	۱۰۰
			پنج گزینه ای	۱۰۰	۱۰۰
		نه سؤال(۳۰٪)	دو گزینه ای	۱۰۰	۱۰۰
			پنج گزینه ای	۱۰۰	۱۰۰
	۰.۷۵	سه سؤال(۱۰٪)	دو گزینه ای	۱۰۰	۱۰۰
			پنج گزینه ای	۱۰۰	۱۰۰
		شش سؤال(۲۰٪)	دو گزینه ای	۱۰۰	۱۰۰
			پنج گزینه ای	۱۰۰	۱۰۰
		نه سؤال(۳۰٪)	دو گزینه ای	۱۰۰	۱۰۰
			پنج گزینه ای	۱۰۰	۱۰۰
	۱	سه سؤال(۱۰٪)	دو گزینه ای	۱۰۰	۱۰۰
			پنج گزینه ای	۱۰۰	۱۰۰
		شش سؤال(۲۰٪)	دو گزینه ای	۱۰۰	۱۰۰
			پنج گزینه ای	۱۰۰	۱۰۰
		نه سؤال(۳۰٪)	دو گزینه ای	۱۰۰	۱۰۰
			پنج گزینه ای	۱۰۰	۱۰۰

درصد متوسط نرخ آشکارسازی DIF با استفاده از روش‌های IRT و CFA در جدول

۲ ملاحظه می‌شود. این جدول همچنین شامل میانگینهای حاشیه ای سطری و ستونی است که با عنوان میانگین کل نمایش داده شده است.

جدول ۲- درصد متوسط نرخ آشکارسازی DIF با استفاده از روش‌های نسبت درستنمایی IRT و تحلیل عاملی تأییدی

روش تشخیص DIF / عامل مداخله گر		نسبت درستنمایی IRT	تحلیل عاملی تأییدی	میانگین کل
حجم نمونه	۱۵۰	۹۱.۹۸	۸۷.۵۲	۸۹.۷۵
	۵۰۰	۹۷.۸۵	۹۵.۷۱	۹۶.۷۸
	۱۰۰۰	۹۸.۸۱	۹۷.۶۵	۹۸.۲۳
نوع DIF	هماهنگ	۹۶.۰۴	۹۴.۸۷	۹۵.۴۵۵
	ناهماهنگ	۹۶.۳۹	۹۲.۳۸	۹۴.۳۸۵
درصد سؤالات DIF در آزمون شبیه‌سازی شده	سه سؤال(۱۰٪)	۹۶.۰۲	۹۲.۷۹	۹۴.۴۰۵
	شش سؤال(۲۰٪)	۹۶.۶۰	۹۴.۳۱	۹۵.۴۵۵
	۹ سؤال(۳۰٪)	۹۶.۰۲	۹۳.۷۷	۹۴.۸۹۵
گزینه های سؤال	دو گزینه ای	۹۷.۵۳	۹۲.۰۸	۹۴.۸۰۵
	پنج گزینه ای	۹۴.۹۰	۹۵.۱۷	۹۵.۰۳۵
شدت DIF در آزمون شبیه‌سازی شده	۰.۲۵	۹۱.۶۷	۸۸.۸۹	۹۰.۲۸
	۰.۵۰	۹۵.۵۸	۹۳.۵۰	۹۴.۵۴
	۰.۷۵	۹۸.۱۴	۹۴.۴۴	۹۶.۲۹
	۱	۹۹.۴۷	۹۷.۶۷	۹۸.۵۷
میانگین کل		۹۶.۲۲	۹۶.۲۲	۹۴.۹۲

همانطور که در جدول ۲ ملاحظه می‌شود، درصد متوسط نرخ آشکارسازی در هر دو روش تشخیص DIF با حجم نمونه ۱۰۰۰ آزمودنی تفاوت چندانی ندارد. بنابراین نتایج می‌توان گفت هنگامی که حجم نمونه بالا است هر دو روش می‌توانند DIF سؤال را در آزمون‌های سرنوشت ساز به یک اندازه آشکار نمایند. درصد متوسط نرخ آشکارسازی در آزمون‌های شبیه‌سازی شده سرنوشت ساز در حجم نمونه ۱۵۰ و ۵۰۰ آزمودنی در آزمون نسبت درستنمایی مبتنی بر نظریه پرسش-پاسخ در مقایسه با آزمون‌های تحلیل عاملی

تأییدی بیشتر است. در کنش افتراقی هماهنگ، آزمون نسبت درستنمایی نسبت به تحلیل عاملی تأییدی نرخ متوسط بیشتری در آشکارسازی DIF سؤال‌های سودار دارد. در کنش افتراقی ناهماهنگ نیز، آزمون نسبت درستنمایی نسبت به تحلیل عاملی تأییدی نرخ متوسط بیشتری در آشکارسازی DIF دارد. در کنش افتراقی ناهماهنگ، آزمون نسبت درستنمایی نسبت به تحلیل عاملی تأییدی نرخ متوسط بیشتری در آشکارسازی DIF دارد. در هر دو سطح DIF سؤال ۱۰ درصد(۳ سؤال)، ۲۰ درصد(۶ سؤال) و ۳۰ درصد(۹ سؤال)، متوسط نرخ آشکار سازی DIF در آزمون نسبت درستنمایی مبتنی بر نظریه پرسش-پاسخ از آزمون‌های تحلیل عاملی تأییدی بیشتر است. در داده‌های سؤال دو وجهی، متوسط نرخ آشکار سازی DIF در آزمون نسبت درستنمایی مبتنی بر نظریه پرسش-پاسخ در مقایسه با تحلیل عاملی تأییدی بیشتر است. در داده‌های سؤال چند وجهی، متوسط نرخ آشکار سازی DIF در آزمون نسبت درستنمایی مبتنی بر نظریه پرسش-پاسخ با آزمون‌های تحلیل عاملی تأییدی تفاوتی ندارد. همچنین در تمامی سطوح شدت DIF (۰.۲۵، ۰.۵۰، ۰.۷۵ و ۱) ، متوسط نرخ آشکار سازی DIF در آزمون نسبت درستنمایی مبتنی بر نظریه پرسش-پاسخ در مقایسه با آزمون‌های تحلیل عاملی تأییدی در آزمون‌های سرنوشت ساز بیشتر است.

بحث و نتیجه گیری

تجزیه و تحلیل آزمون در گروه مرجع و گواه می‌تواند به عنوان روشی کارا برای تحلیل آزمون بکار گرفته شود. این روش مخصوصاً می‌تواند جهت وارسی اعتبار[1] مورد توجه خاص قرار گیرد(فکتو[2] و کریگ[3]،۲۰۰۱؛ وندنبرگ[4]،۲۰۰۲). تحلیل عاملی تأییدی و نظریه پرسش-پاسخ به عنوان دو روش که دارای شباهتهای زیادی هستند می‌توانند به عنوان روش‌هایی پیشرفته و کارا جهت تحلیل DIF در آزمون‌های سرنوشت ساز بکار گرفته

1 .cross validation
2 . Facteau
3 . Craig
4 . Vandenberg

شوند. این دو روش تجزیه و تحلیل آماری به صورت گسترده ای برای برآورد پارامترهای سؤال و آزمودنی نیز بکار گرفته می‌شوند. پارامترهای ضریب تمیز و دشواری در نظریه پرسش-پاسخ قابل قیاس با بارهای عاملی و عرض از مبدأ[1] در تحلیل عاملی تأییدی هستند(فلاورز[2] و همکاران،۲۰۰۲).

یافته‌های تحقیق نشان داد که به طور متوسط، آزمون نسبت درستنمایی مبتنی بر نظریه پرسش-پاسخ نسبت به تحلیل عاملی تأییدی قدرت بیشتری در تشخیص DIF در آزمون‌های سرنوشت‌ساز دارد. نتایج نشان داد که در هر دو روش تشخیص DIF، حجم نمونه، شدت DIF و گزینه‌های سؤال بر نرخ آشکارسازی DIF در آزمون‌های سرنوشت‌ساز مؤثر هستند و درصد سؤالات DIF آزمون و نوع DIF بر نرخ آشکارسازی DIF تأثیری ندارد.

نتایج تحقیق حاضر با نتایج مید و لوتنچلاجر (۲۰۰۶) در این مورد که تحلیل عاملی تأییدی به طور کلی در حجمهای نمونه بالا مؤثر است همخوانی دارد، اما برخلاف نتایج آن‌ها تحقیق حاضر نشان می‌دهد که تحلیل عاملی تأییدی در هر دو مورد DIF هماهنگ و ناهماهنگ تقریباً یکسان عمل می‌کند. البته در تحلیل عاملی تأییدی متوسط نرخ آشکاری سازی DIF هماهنگ کمی بیشتر است. همچنین نتایج تحقیق مید و لوتنچلاجر(۲۰۰۶)، تحلیل عاملی تأییدی را در مورد تشخیص DIF ناهماهنگ ناکافی می‌داند.

نتایج تحقیق حاضر با نتایج مید و لوتنچلاجر(۲۰۰۴) در این مورد که آزمون نسبت درستنمایی مبتنی بر نظریه پرسش-پاسخ هنگامی DIF موجود است قدرت بیشتری نسبت به تحلیل عاملی تأییدی دارد، همخوانی دارد. همچنین نتایج تحقیق حاضر در مورد قدرت ضعیف‌تر آزمون نسبت درستنمایی در حجم نمونه پایین با نتایج مید وهمکاران همخوانی دارد. البته دلیل آن به خاطر این است که حجم نمونه ۱۵۰ آزمودنی با استانداردهای تحلیل در نظریه سؤال ـ پاسخ فاصله زیادی دارد.

1 .intercept
2 .Flowers

همچنین نتایج تحقیق حاضر برخلاف نتایج فلاورز و همکاران(۲۰۰۲) در مورد اثر نوع DIF بر نرخ آشکار سازی DIF سؤال‌های سودار است. آن‌ها نتیجه گرفتند که هر دو روش آزمون نسبت درستنمایی مبتنی بر نظریه پرسش-پاسخ و تحلیل عاملی تأییدی در آشکار سازی DIF هماهنگ قدرت کمتری دارند. نتایج تحقیق فلاورز و همکاران با نتایج تحقیق مید وهمکاران (۲۰۰۶) نیز تفاوت دارد. این تفاوت‌ها در مورد اثر نوع DIF بر نرخ آشکارسازی در روش‌های نسبت درستنمایی و تحلیل عاملی تأییدی، لزوم تحقیق بیشتر در این مورد با کنترل سایر عوامل از قبیل توزیع ضرایب دشواری و قدرت تمیز را بیشتر می‌کند.

اگرچه در تحقیق حاضر آزمون نسبت درستنمایی مبتنی بر نظریه پرسش-پاسخ و تحلیل عاملی تأییدی به عنوان رویکردهای بررسی نامتغیر بودن اندازه گیری مورد بررسی قرار گرفتند، اما تأکید روی این نکات دارای اهمیت زیادی است: الف- این دو روش مفروضه های متفاوتی دارند، به طور مثال نظریه پرسش-پاسخ دارای این مفروضه است که رابطه توانایی پنهان و سؤالات آزمون رابطه ای غیرخطی است در حالیکه تحلیل عاملی تأییدی این رابطه را خطی فرض می‌کند، ب- این دو روش، آگاهی[1] متفاوتی از نامتغیر بودن اندازه گیری فراهم می‌کنند، ج- هیچ یک از این روش‌ها فارغ از خطا نیستند. پژوهشگران وقتی تنها از یک روش برای تشخیص DIF استفاده کنند تصویری نادرست از ویژگی‌های روان‌سنجی مقیاس اندازه گیری بدست می‌آورند. بعلاوه در بعضی مواقع نتایج آزمون‌های نامتغیر بودن اندازه گیری می‌تواند گمراه کننده باشند، مثلاً زمانی که در تحلیل عاملی تأییدی حجم نمونه پایین است یا میزان اشترک عامل‌ها کم است. همچنین زمانی که حجم نمونه پایین است، نظریه پرسش-پاسخ هیچگونه آگاهی‌ای در مورد ارتباط میان متغیرهای پنهان فراهم نمی‌کند. این ارتباط در بعضی مواقع می‌تواند بسیار مهم باشد(مید و لوتنچلاجر، ۲۰۰۴).

بنابراین بر اساس نتایج تحقیق حاضر پیشنهاد می‌شود که در صورت محدودیت در

1 .information

استفاده از تنها یک روش بررسی DIF ، روش نسبت درستنمایی مبتنی بر مدل پرسش-پاسخ بر روش تحلیل عاملی تأییدی ترجیح داده شود. در چنین شرایطی آشنایی متخصصان از نقاط ضعف و قوت روش‌های مختلف تحلیل DIF بسیار تعیین کننده است. البته مدل پرسش-پاسخ و تحلیل عاملی تأییدی با حجم‌های آزمودنی بزرگ در تحقیقات مختلف برتری خود را در روش‌های تحلیل DIF ثابت کرده‌اند(کیم و همکاران،۱۹۹۶).

عوامل مختلفی از قبیل توزیع پارامترهای سؤال و آزمودنی، حجم نمونه، اندازه DIF، تعداد سؤال‌های دارای DIF، تعداد گزینه های سؤال و شدت DIF می‌توانند بر قدرت و کارایی آزمون‌های تحلیل عاملی تأییدی و نسبت درستنمایی مبتنی بر نظریه پرسش-پاسخ تأثیر گذار باشند(استارک و همکاران،۲۰۰۴). بنابراین برای تحقیقات آینده پیشنهاد می‌شود سناریوهای دیگری از داده‌های شبیه‌سازی شده با توجه به عوامل مذکور طراحی و تجزیه و تحلیل شوند تا قدرت هر یک از این آزمون‌ها در شرایط دیگر و مداخله عوامل مختلف مشخص شود. با معرفی روش جدید شبیه‌سازی داده‌های سؤال و آزمودنی که در این مقاله پیشنهاد شده است، این امر مستلزم تلاش و صرف زمان زیادی است که می‌تواند میان پژوهشگران رشته سنجش و اندازه گیری تقسیم شود.

منابع

همبلتون، رونالد ک و سوامیناتان، اچ و راجرز، اچ. جین(۱۹۹۱). *مبانی نظریه پرسش-پاسخ* ، ترجمه محمد رضا فلسفی نژاد(۱۳۸۹). تهران: انتشارات دانشگاه علامه طباطبائی، ص ۱۳۹

Bentler, P. M., & Yaun, K.-H. (1999). Structural equation models with small samples: Test statistics. Multivariate Behavioral Research, 34, 181-197.

Camilli, G., & Shepard, L. A. (1994). *Methods for identifying biased test items.* Thousand Oaks, CA: Sage Publications.

Facteau, J. D., & Craig, S. B. (2001). Are performance appraisal ratings from different rating sources comparable? *Journal of Applied Psychology*, 86, 215–227.

Flowers, C. P., Raju, N. S.,&Oshima, T. C. (2002, April). *A comparison of measurement equivalence methods based on confirmatory factor analysis and item response theory*. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Gen,L.(2007). *WINGEN2: User Manual*. Hillsdale NJ: Erlbaum.

Gomez-Benito, J., & Navas-Ara, M. J. (2000). A comparison of $\chi 2$, RFA and IRT based procedures in the detection of DIF. *Quality and Quantity*, 34(1), 17–31.

Hambleton, R. K., Swaminathan, H., & Rogers, H .J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage Publications.

Kim, S.-H., Cohen, A. S., & Kim, H.-O. (1996).An Investigation of the Likelihood Ratio Test For Detection of Differential Item. Applied Psychological Measurement, 18, 217-228.

Meade, A. W., & Bauer D.J (2007). Power and Precision in Confirmatory Factor Analytic Tests of Measurement Invariance. *STRUCTURAL EQUATION MODELING*, 14(4), 611–635

Meade, A. W., & Lautenschlager, G. J. (2004). A Comparison of Item Response Theory and Confirmatory Factor Analytic Methodologies for Establishing Measurement Equivalence/Invariance. *Structural Equation Modeling*, 11,60-72.

Meade, A. W., & Lautenschlager, G. J. (2006). A Monte-Carlo Study of Confirmatory Factor Analytic Tests of Measurement Equivalence/Invariance. *Structural Equation Modeling*, 23,83-111.

Parshall, C. G. & Miller, T. R. (1995). Exact versus asymptotic Mantel-Haenszel DIF statistics: A comparison of performance under small-sample conditions in high stakes tests. *Journal of Educational Measurement*, 32(3), 302–316.

Popham,W.J.(2005). HIGH-STAKES TESTS: HARMFUL, PERMANENT, FIXABLE, *American* Educational Research Journal.6,p85

Ramsey, P. A. (1993). Sensitivity review: The ETS experience as a case study. In P. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 367-3881. Hillsdale, NJ: Erlbaum.

So¨rbom, D. (1974). A general method for studying differences in factor means and factor structure between groups. *British Journal of Mathematical* and Statistical Psychology, 27, 229–239.

Stark, S., Chernyshenko, O. S., & Drasgow, F.(2004). Detecting Differential Item Functioning With Confirmatory Factor Analysis and Item Response Theory: Toward a Unified Strategy. *Journal of Applied Psychology*,89,497-508

Swaminathan, H. & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement*, 27, 361– 370.

Thissen, D., Steinberg, L., & Gerrard, M. (1986). Beyond group-mean differences: The concept of item bias. *Psychological Bulletin*, 99, 118-128.

Thissen, D., Steinberg, L., & Wainer, H. (1988). Use of item response theory in the study of group differences in trace lines. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 147-169). Hillsdale NJ: Erlbaum.

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning

using the parameters of item response models. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 67-113). Hillsdale NJ: Erlbaum.

Vandenberg, R. J. (2002). Toward a further understanding of and improvement in measurement invariance methods and procedures. *Organizational Research Methods, 5,* 139–158.

Wainer, H. (1993). Model-based standardized measurement of an item's differential impact. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 123-135). Hillsdale NJ: Erlbaum.