

روشهای آماری بررسی کنش افتراقی

سؤال (DIF) در آزمون های سرنوشت ساز

دکتر مسعود گرامی پور^۱، دکتر محمد رضا فلسفی نژاد^۲

^۱ عضو هیأت علمی دانشکده روانشناسی و علوم تربیتی دانشگاه خوارزمی
^۲ عضو هیأت علمی دانشکده روانشناسی و علوم تربیتی دانشگاه علامه طباطبائی

سرشناسه	: گرامی‌پور، مسعود، ۱۳۵۹ -
عنوان و نام پدیدآور	: روش‌های آماری بررسی کنش افتراقی سوال (Dif) در آزمون‌های سرنوشت‌ساز/ مسعود گرامی‌پور، محمدرضا فلسفی‌نژاد.
مشخصات نشر	: تهران: جهاد دانشگاهی، واحد خوارزمی، ۱۳۹۲.
مشخصات ظاهری	: ۱۰۴ ص: جدول، نمودار.
شابک	: ۴۰۰۰۰ ریال 3-8-90599-978-964
وضعیت فهرست‌نویسی	: فیا
یادداشت	: کتابنامه.
موضوع	: آزمون‌ها -- اعتبارسنجی -- روش‌های آماری
شناسه افزوده	: فلسفی‌نژاد، محمدرضا، ۱۳۴۶ -
شناسه افزوده	: جهاد دانشگاهی. واحد خوارزمی
رده بندی کنگره	: ۱۳۹۲ LB۳۰۶۰/۷۸۳۴
رده بندی دیویی	: ۳۷۱/۲۶۱۳
شماره کتابشناسی ملی	: ۳۱۲۵۱۵۰



انتشارات جهاد دانشگاهی واحد خوارزمی

شابک : ۳ - ۸ - ۹۰۵۹۹ - ۹۷۸-۹۶۴

ISBN : 978- 964 - 90599 - 8 - 3

واحد خوارزمی

نام کتاب : روش‌های آماری بررسی کنش افتراقی سؤال (DIF) در آزمون‌های سرنوشت (ساز)

تألیف : دکتر مسعود گرامی‌پور / دکتر محمد رضا فلسفی‌نژاد

نوبت چاپ اول : بهار ۱۳۹۲

لیتوگرافی : پویش

چاپ : تربیت معلم

شمارگان : ۱۰۰۰

قیمت : ۴۰۰۰۰ ریال

آدرس ناشر : تهران - خیابان انقلاب اسلامی - خیابان شهید مفتاح - پلاک ۴۹ - دانشگاه خوارزمی جهاد دانشگاهی واحد خوارزمی تلفن و دورنگار : ۸۸۳۰ ۲۵۲۴ - ۸۸۳۰۹۲۹۲ مرکز پخش : ۷۷۵۰۳۸۲۴

این اثر مشمول قانون حمایت مؤلفان و مصنفان و هنرمندان مصوب ۱۳۴۸ است. هر کس تمام یا قسمتی از این اثر را بدون اجازه مؤلف (ناشر) نشر یا پخش یا عرضه کند مورد پیگرد قانونی قرار خواهد گرفت.

۷	پیشگفتار مؤلفان
۹	۱- مقدمه
۱۶	۲- استانداردهای مصفانه بودن آزمون
۱۹	۳- سوگیری سؤال و رویکردهای کلی بررسی کنش افتراقی سؤال (DIF)
۲۲	۴- آزمونهای سرنوشت ساز
۲۵	۵- سوگیری در اندازه گیری
۲۸	۶- جامعه های چند گانه
۲۹	۷- مفهوم آماری سوگیری یا کنش افتراقی سؤال (DIF)
۳۱	۸- تک بعدی بودن و تغییر ناپذیری
۳۳	۹- روش های بررسی کنش افتراقی سؤال (DIF)
۳۳	۹-۱- تحلیل عاملی تأییدی
۳۶	۹-۱-۱- اعمال محدودیت های تغییر ناپذیری
۳۷	۹-۱-۲- آزمون های تغییر ناپذیری
۳۹	۹-۲- نظریه پاسخ سؤال
۴۱	۹-۲-۱- مدل دو پارامتری
۴۳	۹-۲-۳- آزمون های نسبت درستی برای بررسی کنش افتراقی سؤال (DIF)
۴۵	۹-۳- رگرسیون لجستیک

- ۱۰- رابطه میان روش‌های بررسی کنش افتراقی سؤال مبتنی بر IRT، CFA و رگرسیون لجستیک ۴۹
- ۱۱- پیشینه مطالعاتی روش‌های بررسی کنش افتراقی سؤال (DIF) ۵۱.....
- ۱۲- روش‌های مقایسه قدرت آزمون‌های بررسی کنش افتراقی سؤال (DIF) ۵۵.....
- ۱۲-۱- شیوه شبیه سازی داده‌های پاسخ سؤال ۵۷.....
- ۱۴- روش‌های عملی تجزیه و تحلیل کنش افتراقی سؤال (DIF) ۵۸.....
- ۱۵- نتیجه گیری ۸۶.....
- فهرست منابع ۹۲.....

پیشگفتار مؤلفان

امروزه آزمونهای سرنوشت ساز از قبیل آزمون های سراسری و دریافت مدارک حرفه ای بصورت گسترده ای در سطح کشور اجرا می شوند. نتایج این آزمون ها مشخص می کنند که چه افرادی صلاحیت ادامه تحصیل و یا انجام مشاغل خاصی را دارند. مشاغل حساسی از قبیل جراحی که مستقیماً با جان انسانها ارتباط دارند و یا مدارک تحصیلی که افراد را مجاز به انجام مشاغلی از قبیل مهندسی، معلمی و کارمندی می کنند و یا حتی مشاغل سیاسی که سرنوشت کشور را رزم می زنند، نیازمند مدارکی هستند که وابسته به نتایج آزمون های سرنوشت ساز است. بنابراین اعتبار نمراتی که از این آزمونها حاصل می شوند، از اهمیت خاصی برخوردار است. یکی از منابع مهمی که اعتبار آزمونهای سرنوشت ساز را خدشه دار می کند کنش افتراقی سؤال (DIF) است. DIF نشان می دهد که آیا سؤالهای آزمون به نفع گروه خاصی از آزمودنی ها سوگیری دارند. اگر سؤالهای آزمون دارای چنین خاصیتی نباشند آزمون مورد نظر دارای تغییر ناپذیری اندازه گیری است.

روشهای مختلفی جهت بررسی DIF معرفی شده اند اما در کتاب حاضر سه روش پیشرفته بررسی DIF شامل روشهای آماری مبتنی بر تحلیل عاملی تأییدی، نسبت درستیابی مبتنی بر نظریه پاسخ سؤال و رگرسیون لجستیک معرفی خواهند شد. اسناد بیشتر به این روشها در مقالات پژوهشی، پیچیدگی محاسباتی و دقت بیشتر این روشها در آشکارسازی DIF دلیل انتخاب آنها در این کتاب بوده اند. در کتاب حاضر ابتدا این سه روش معرفی می شوند، سپس با استفاده از داده های شبیه سازی شده و بصورت عملی نحوه بکارگیری این سه روش در تحلیل DIF توضیح داده خواهند شد. سرانجام نقاط ضعف و قوت هر یک از این روشها مورد بحث و نتیجه گیری قرار می گیرند و پیشنهادهای عملی برای استفاده از این روشها در شرایط پژوهشی مختلف ارائه خواهند شد.

مطالعه کتاب حاضر به متخصصان روش شناسی تحقیق در رشته های علوم انسانی و دانشجویان مقاطع تحصیلات تکمیلی رشته های سنجش و اندازه گیری، سنجش آموزش، تحقیقات آموزشی، پژوهشگری علوم اجتماعی و دانشجویان کلیه مقاطع رشته های آمار پیشنهاد می شود.

آزمون‌های سرنوشت‌ساز^۱ آزمون‌هایی هستند که در مقیاس وسیع اجرا شده و بر اساس نتایج آن‌ها تصمیم‌هایی خطیر در مورد افراد اتخاذ می‌شود. چنین تصمیماتی دارای تبعات مهم شخصی، اجتماعی و سیاسی هستند. لذا، اعتبار تصمیم‌گیری و کارایی مدل‌های سنجش در این مورد از اهمیت فوق‌العاده‌ای برخوردار بوده و در سالیان گذشته جزء مهمترین چالش‌های روان‌سنجی باقی مانده‌اند. یکی از مهمترین تهدیدها برای اعتبار^۲ آزمون، سوگیری سؤال^۳ یا فقدان تغییر ناپذیری اندازه‌گیری^۴ است که از آن تحت عنوان کنش افتراقی سؤال (DIF)^۵ یاد می‌شود. کنش افتراقی سؤال شرط لازم ولی ناکافی برای سوگیری سؤال است (رامسی^۶، ۱۹۹۳). تغییر ناپذیری^۷ اندازه‌گیری ویژگی‌ای است که در آن سؤال‌های آزمون در شرایط مختلف ویژگی‌های یکسانی را اندازه‌گیری می‌کنند (هورن^۸ و مک آردل^۹، ۱۹۹۲). این شرایط شامل گروه‌ها، مراکز اجرایی و روش‌های مختلف آزمون (مداد-کاغذی یا رایانه‌ای) می‌شوند. یک مقیاس زمانی دارای تغییر ناپذیری اندازه‌گیری است که در آن بیش از یک گروه نمرات یکسانی در یک سازه پنهان داشته باشند و نمرات خام مشاهده شده و مورد انتظار آنان در گروه‌های مختلف در سطح سؤال و آزمون یکسان باشند (دراستگو و کانفر^{۱۰}، ۱۹۸۵).

روانشناسان اهمیت تغییر ناپذیری اندازه‌گیری را به عنوان پیش‌نیازی برای مقایسه‌های گروهی گوشزد کرده‌اند (به طور مثال دراستگو، ۱۹۸۴؛ راجو^{۱۱} و همکاران، ۲۰۰۲؛ ریس^{۱۲} و همکاران، ۱۹۹۳؛ وندربرجک^{۱۳}، ۲۰۰۲). هر چقدر که آزمون کارکرد متفاوتی برای گروه‌های مختلف داشته باشد، مقایسه‌های گروهی نیز غیر قابل اعتماد خواهد بود. بنابراین فقدان تغییر ناپذیری اندازه‌گیری

¹ high stakes tests

² validity

³ item Bias

⁴ measurement invariance

⁵ Differential Item Functioning

⁶ Ramsey

⁷ equivalence

⁸ Horn

⁹ McArdle

¹⁰ Kanfer

¹¹ Raju

¹² Reise

¹³ Vandenberg

و کنش افتراقی سؤال به عنوان تهدیدی جدی برای اعتبارآزمون تلقی می‌شود، به طوریکه استانداردهای انجمن روانشناسی آمریکا (APA)^۱ به بررسی DIF سؤال‌های آزمون برای ارزیابی منصفانه بودن آن تأکید دارد (برن^۲ و استوارت^۳، ۲۰۰۶).

روش‌های مختلفی برای بررسی DIF سؤال‌های آزمون معرفی و به کار گرفته شده‌اند، اما علی‌رغم 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 (میلپس، ۲۰۱۱).

شاید مهمترین اتهام وارده به سنجش و مطمئناً مهمترین مسئله برای مردم، سنجش منصفانه^۴ است. ادعای این که آزمون‌ها در برابر اقلیت‌ها و نژادهای خاص سوگیری دارند، منجر به لوایح قانونی متعدد شده است. آزمون‌ها و آزمون‌سازها مورد توجه مردم قرار گرفته‌اند، بنابراین آزمون‌سازها باید نشان دهند که آزمون‌هایشان در برابر اقلیت‌ها، سوگیری ندارند (همبلتون^۵ و سوامیناتان^۶، ۱۹۹۱).

از طرفی عادلانه بودن برای آزمون‌های سرنوشت‌ساز به عنوان آزمون‌هایی که جهت تصمیم‌گیری برای آینده آزمودنی‌ها به کار می‌روند از اهمیت بیشتری برخوردار است و حساسیت بیشتری نسبت به عادلانه بودن سؤال‌های این نوع آزمون‌هاشان داده می‌شود (پاپهام^۷، ۲۰۰۵).

اصطلاح کنش افتراقی سؤال (DIF) به عنوان گواهی تجربی برای وجود یا عدم وجود سوگیری در سؤال مورد استفاده قرار می‌گیرد. یک سؤال موقعی دارای DIF است که افرادی با توانایی یکسان اما با گروهی متفاوت، احتمال یا بختی متفاوت در دادن پاسخ درست به سؤال داشته باشند. این گروه‌ها در ادبیات DIF اصطلاحاً گروه‌های مرجع و کانونی^۸ نامیده می‌شوند. معمولاً گروه مرجع، گروه اکثریت و گروه کانونی گروه اقلیت یا محروم در نظر گرفته می‌شود (سوامیناتان و

^۱ American Psychological Association

^۲ Byrne

^۳ Stewart

^۴ Test fairness

^۵ Hambleton

^۶ Swaminathan

^۷ Popham

^۸ Reference and focal groups

راجرز^۱، 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 واقعاً در توانایی مورد نظر تفاوت داشته باشند. تعیین سؤال‌های سودار گامی مهم در سنجش معتبر است. یک سؤال سودار دارای خطای نظام‌دار است که باعث نامعتبر شدن نتایج آزمون می‌شود (کمیلی^۲ و شپارد^۳، ۱۹۹۴). روش‌های آماری مختلفی جهت تعیین DIF وجود دارد که می‌توان آن‌ها را به دو دسته کلی روش‌های مبتنی بر نظریه پاسخ سؤال (IRT) و روش‌های غیر مبتنی بر نظریه پاسخ سؤال تقسیم کرد. روش‌های موجود در گروه اول شامل روش‌هایی برای مقایسه پارامترهای سؤال و مساحت میان منحنی‌های ویژگی سؤال (ICC) است (همبلتون^۴ و سامینان^۵، ۱۹۸۵). روش‌هایی که شامل گروه دوم می‌شوند مبتنی بر نظریه اندازه‌گیری کلاسیک (CTT) و روش‌های تحلیل عاملی و روش‌های مبتنی بر تحلیل داده‌های طبقه‌ای هستند (گومز^۶ و ناواس^۷، ۲۰۰۰).

تیشن^۸، استینبرگ^۹ و جرارد^{۱۰} (۱۹۸۶) و تیشن، استینبرگ و وینر^{۱۱} (۱۹۹۳) آزمون نسبت درستنمایی را مطرح کرده‌اند که مبتنی بر نظریه IRT است. در این روش، برازش مدل تکمیلی^{۱۲} A که در آن پارامترهای IRT سؤال مورد مطالعه اجازه دارند در گروه‌های مورد مقایسه تغییر کنند با برازش مدل فشرده^{۱۳} C که در آن پارامترهای سؤال در گروه‌ها محدود می‌شوند، مقایسه

^۱ Rogers

^۲ Camilli

^۳ Shepard

^۴ Item Response Theory

^۵ Item Characteristic Curve

^۶ Hambleton

^۷ Classical Test Theory

^۸ Gomez

^۹ Navas

^{۱۰} Thissen

^{۱۱} Steinberg

^{۱۲} Gerrard

^{۱۳} Wainer

^{۱۴} Augmented

^{۱۵} Compact

می‌شود. مقدار درستمایی L در مدل فشرده در یک نوبت برآورد محاسبه می‌شود. در محاسبات بعدی، درستمایی مدل افزایشی A محاسبه می‌شود که در آن از ترکیب پارامترهای ضریب تمیز (a)، ضریب دشواری (b) و حدس (c) برای یک سؤال در گروه کانونی صرفنظر می‌شود. در این روش سؤال مورد مطالعه در هر یک از گروه‌ها متفاوت در نظر گرفته می‌شود. فرض صفر (مدل C: H_0 : عدم DIF) به نفع فرض خلاف (مدل A: H_a) با استفاده از آزمون نسبت درستمایی رد می‌شود.

روش بعدی مورد مطالعه در کتاب حاضر که بر اساس نظریه پاسخ سؤال نیست روش تحلیل عاملی تأییدی است. در این روش همسانی میان بارهای عاملی و مقادیر ثابت مدل به صورت مستقیم آزمون می‌شوند. ابتدا باید خط پایه مدل تعیین شود، جایی که همه پارامترها می‌توانند آزادانه تغییر کنند، به جز آن‌هایی که برای تعیین مدل ثابت نگه داشته می‌شوند. مرحله بعد تعیین مجموعه‌ای از مدل‌های محدود شده است که در آن بارهای عاملی یا مقادیر ثابت توسط اعمال محدودیت‌هایی در گروه‌های مرجع و کانونی مساوی در نظر گرفته می‌شوند و سپس مدل‌های محدود شده با مدل‌های خط پایه با استفاده از آماره‌های مقایسه می‌شوند. به دلیل اینکه هر کدام از مدل‌های محدود شده در درون مدل‌های خط پایه قرار گرفته‌اند می‌توان تغییر در مقدار خبی دو را با توجه به خط پایه محاسبه کرد. برای هر مقایسه یک مقدار معنی‌دار از نظر آماری گواهی بر کنش افتراقی سؤال است (سوربوم^۱، ۱۹۷۴).

سوامیناتان^۲ و راجرز^۳ (۱۹۹۰) روش کنش افتراقی سؤال بر اساس رگرسیون لجستیک را معرفی کرده‌اند. رگرسیون لجستیک به خانواده بزرگتری از تحلیل‌ها تعلق دارد که مدل‌های خطی عمومی نامیده می‌شوند (اگرستی^۴، ۲۰۰۷). در این رویکرد مجموع نمرات یا نمرات صحیح به عنوان متغیری پیوسته و مستقل به عنوان توانایی در نظر گرفته می‌شود. در رگرسیون لجستیک پاسخ سؤال به عنوان متغیر وابسته با توزیع برنولی دو وجهی در نظر گرفته می‌شود و یک متغیر گروهی دیگر به عنوان متغیری که قصد بررسی پاسخ‌ها در آن را داریم به عنوان گروه‌های مرجع و کانونی در نظر گرفته می‌شوند که معنی‌داری اثر آن در این مدل به عنوان گواهی برای کنش

^۱ Sorbom

^۲ Swaminathan

^۳ Rogers

^۴ Agresti

افتراقی سؤال در نظر گرفته می‌شود. این در حالی است که اثر متغیر توانایی و تعامل آن با متغیر گروهی کنترل شده است. معنی‌داری تعامل متغیر گروهی و توانایی نیز خود گواهی بر نوعی سوگیری سؤال است که کنش افتراقی ناهماهنگ نامیده می‌شود.

همانطور که ملاحظه می‌شود تعاریف مفهومی از سوگیری بسیار کلی است و برای مدل‌های مختلف مورد استفاده قرار می‌گیرد. از دهه‌های گذشته روان‌سنج‌هایی که درباره تغییر ناپذیری اندازه‌گیری تحقیق می‌کنند، روش‌های آماری متفاوتی را برای اندازه‌گیری کنش افتراقی سؤال (DIF) پیشنهاد کرده‌اند. این روش‌ها معمولاً مفروضات آماری متفاوتی را در مورد شکل رابطه میان متغیرهای پنهان و مشاهده شده و شکل توزیع متغیر پنهان در جامعه در بر دارند و راهبردهای مدل‌یابی متفاوتی را برای تعیین معنی‌داری کنش افتراقی سؤال به کار می‌برند. بنابراین تعیین اینکه کدامیک از روش‌ها نتایج بهتری را به دست می‌دهد بسیار مشکل است، به علاوه معلوم نیست در چه مواقعی استفاده از روش‌های متفاوت، نتایج متفاوتی را در بر دارد. به طوریکه متغیرهایی از قبیل حجم نمونه، DIF هماهنگ، DIF ناهماهنگ، درصد سؤالهای دارای DIF و تعداد گزینه‌های سؤال به عنوان عوامل اندازه‌گیر عنوان شده‌اند (کرین^۱ و همکاران، ۲۰۰۶).

روش‌های تشخیص DIF که در این کتاب مد نظر است به صورت گسترده‌ای مورد استفاده قرار گرفته و به نتایج آن‌ها در مطالعات روان‌سنجی اکتفا می‌شود، اما به نظر می‌رسد که مطالعه‌ای در زمینه مقایسه همزمان روش رگرسیون لجستیک با روش‌های CFA و IRT برای تعیین قدرت نسبی آن در شناسایی DIF انجام نشده است و علی‌رغم اهمیت مطالعه DIF این ویژگی مهم روان‌سنجی کمتر مورد مطالعه قرار گرفته است و کمتر به صورت عملی در آزمون‌های سرنوشت‌ساز انجام و نتایج آن گزارش شده است. تحقیقات انجام شده غالباً نتایج ناهمسانی را گزارش کرده‌اند و هنوز به طور قطع برتری این روش‌ها در مقابل یکدیگر به اثبات نرسیده است. کتاب حاضر از روش‌های مبتنی بر IRT، روش نسبت درستنمایی^۲ و از روش‌های غیر مبتنی بر IRT، روش‌های رگرسیون لجستیک دو وجهی^۳ و روش تحلیل عاملی تأییدی (CFA)^۴ را به عنوان روشهای پیشرفته در این حوزه انتخاب کرده است و قصد دارد با استفاده از داده‌های

^۱ Crane

^۲ Likelihood Ratio

^۳ Binary Logistic Regression

^۴ Confirmatory Factor Analysis

شبیه سازی شده که به مطالعات مونت کارلو^۱ معروف هستند، هر کدام از این روش ها را در تشخیص کنش افتراقی سؤال بصورت کاربردی توضیح دهد.

همچنین بیشتر مطالعات انجام شده در حوزه DIF صرفاً به صورت نظری و در قالب پژوهش های بنیادی مسأله را بررسی نموده اند. به نظر می رسد یکی از عوامل اصلی فقدان مطالعات لازم در شرایط واقعی، اجرای آزمون ها، جدید بودن، نامأنوس بودن و همچنین پیچیدگی نسبی روش های شبیه سازی و نرم افزارهای انجام مطالعات DIF است. در مورد آزمون های سرنوشت ساز مطالعات شبیه سازی شده بستری را فراهم می کنند تا محقق بتواند مجموعه بزرگی از داده های پاسخ سؤال را در تکرارهای مختلف ایجاد کند و با شبیه سازی سؤالهای سودار و دستکاری پارامترهای مختلف کارایی روش های آماري مختلف را در آشکارسازی سؤالهای سودار در آزمون های سرنوشت ساز مورد بررسی قرار دهد (هان و همبلتون ۲۰۰۷).

تغییر ناپذیری اندازه گیری بحوبی می تواند توسط قیاس با اندازه گیری فیزیکی معرفی شود. به طور کلی مدتهاست که مثالهای فیزیکی در نظریه های اندازه گیری نقش مهمی ایفا می کنند (لرد و ناولیک^۳، ۱۹۶۸). تغییر ناپذیری اندازه گیری بر این اساس است که ابزار اندازه گیری باید در شرایط مختلف یکسان عمل کند. برای مثال یک ترازو باید اشیاء مختلفی که وزن های متفاوتی دارند، از هم متمایز کند. نگرانی در این مثال این است که ابزار اندازه گیری، اشیایی با وزن یکسان ولی شکلی متفاوت را یکسان وزن نکند. در این مثال ابزار اندازه گیری در گزارش وزن اشیاء در مورد شکل آن ها سوگیری دارد یعنی اینکه در گزارش وزن اشیاء شکل آن ها را نیز دخیل می کند (میلسپ، ۲۰۱۱).

در اندازه گیری های روانی و تربیتی، آزمون یا پرسشنامه ای که برای اندازه گیری توانایی یا ویژگی خاصی ساخته شده باید بتواند تفاوت های افراد را نشان دهد. برای مثال مقیاسی که برای اندازه گیری افسردگی ساخته شده است باید تفاوت های افرادی که از نظر افسردگی متفاوتند را نمایان کند. از طرف دیگر، ابزار اندازه گیری نباید افرادی که از نظر متغیر مورد بررسی تفاوتی ندارند را متفاوت نشان دهد. مثلاً مقیاس افسردگی نباید زنان و مردانی که از نظر افسردگی تفاوتی

^۱ Monte Carlo Studies

^۲ Han & Hambleton

^۳ Novick

ندارند را متفاوت نشان دهد، در غیر این صورت اصطلاحاً می‌گوییم مقیاس افسردگی در مورد جنسیت سوگیری دارد (هورن^۱ و مک آردل^۲، ۱۹۹۲).

سوگیری در اندازه‌گیری نوعی خطای نظام‌دار است، این نوع خطا تکرار پذیر است در حالیکه خطاهای تصادفی مربوط به پایایی ابزار اندازه‌گیری می‌شود. آزمونی که پایایی دارد ممکن است سودار باشد یا نباشد. آزمونی که پایایی ندارد ممکن است در اندازه‌گیری‌های تکراری سوگیری نداشته باشد، یا یک آزمون سودار می‌تواند کاملاً در اندازه‌گیری‌های متفاوت پایا باشد. با وجود چنین پیش‌فرض‌هایی از کجا می‌توان فهمید که یک ابزار اندازه‌گیری سودار است؟ در مورد وزن ممکن است نتایج وزن کردن یک شیء را با ابزار اندازه‌گیری دیگری که معمولاً گرانتر و وقتگیرتر است، مقایسه کرد. در اندازه‌گیری‌های روانی و تربیتی چنین کاری معمولاً کاربردی نیست. زیرا اندازه‌گیری مکرر افراد با یک ابزار اندازه‌گیری موجب خستگی آن‌ها می‌شود یا اینکه بخاطر عواملی مثل حافظه باعث تغییر رفتار آن‌ها می‌شود. به علاوه، در اکثر حوزه‌های اندازه‌گیری روانی و تربیتی مقیاس‌های "استاندارد طلایی" که پایایی بالایی دارند و فارغ از سوگیری هستند، وجود ندارند. بنابراین اسکارسازی سوگیری در اندازه‌گیری‌های روانی و تربیتی بسیار دشوار است.

به علاوه، وجود تفاوت‌های گروهی معنادار در متغیرهای روانی و تربیتی لزوماً به معنای وجود سوگیری در ابزار اندازه‌گیری نیست. ممکن است تفاوت‌های گروهی واقعی در جامعه وجود داشته باشد. در بسیاری از حوزه‌های اندازه‌گیری روانی و تربیتی، نمی‌توان فهمید که تفاوت‌های گروهی نشان دهنده تفاوت‌های واقعی میان گروه‌ها است، مگر اینکه داده‌هایی از تحقیقات گذشته در مورد ابزار مورد نظر یا ابزاری شبیه به آن در دست باشد (میلز^۳، ۲۰۱۱).

¹ Horn

² Mc Ardle

³ Millsap

۲- استانداردهای منصفانه بودن^۱ آزمون

قبل از پرداختن به موضوع استانداردهای منصفانه بودن آزمون باید بعضی از اصطلاحات تعریف شوند. این اصطلاحات شامل منصفانه بودن، سوگیری^۲ آماری، تأثیر نابرابر^۳، کنش افتراقی سؤال (DIF^۴) و سوگیری آزمون می شوند. اکثر این اصطلاحات فنی هستند، اما منصفانه بودن استفاده گسترده تری دارد. منصفانه بودن به جنبه هایی از آزمون اشاره دارد که در آن نمرات تفسیر می شوند و برای تصمیم گیری در مورد آزمودنی مورد استفاده قرار می گیرند. منصفانه بودن با اعتبار آزمون رابطه نزدیکی دارد. بطور کلی ارزیابی عادلانه بودن آزمون علاوه بر داده ها و شواهد تجربی نیازمند توجیهات قانونی، اخلاقی، سیاسی، فلسفی و اقتصادی است.

منصفانه بودن آزمون نیازمند تعاریف و تفاسیر متفاوت در شرایط اجتماعی و سیاسی مختلف است. با توجه به نقش گروهها در تعریف عدالت آزمون، تعاریف مختلفی از منصفانه بودن آزمون برای دو گروه می توان ارائه داد. هالند^۵ و تایلر^۶ (۱۹۸۸) معتقدند که باید عملکرد دو گروه در یک سؤال آزمون یا نمره کل آزمون و یا قدرت پیش بینی یک معیار مورد مقایسه قرار گیرد. گروه کانونی^۸ که گاهی گروه تحت پوشش^۹ نامیده می شود، در درجه اول اهمیت قرار دارد. این گروه با گروه دیگری که معمولاً گروه مرجع^{۱۰} نامیده می شود.

اولین گام در بررسی منصفانه بودن آزمون، تعریف گروه ها است. سپس باید تفاوت های گروه مرجع و کانونی در دستیابی به یک فرصت یا شغل و یا نمره یک سؤال در آزمون مورد مقایسه قرار گیرد. تفاوت بدست آمده "تأثیر نابرابر"^{۱۱} نامیده می شود.

¹ fairness

² - bias

³ - disparate impact

⁴ - Differential Item Functioning

⁵ - validity

⁶ - Holland

⁷ - Thayer

⁸ - focal

⁹ - protected

¹⁰ - reference

¹¹ - Disparate impact