دانشگاه علامه طباطبایی
دانشکده علوم تربیتی و روانشناسی
گروه سنجش و اندازه‌گیری

مقایسه‌ی کاربرد اندازه‌ی اثر در سنجش عملکرد افتراقی سوال بین دو روش رگرسیون لجستیک و منتل هنزل

استاد راهنما:
دکتر علی دلاور

استاد مشاور:
دکتر فریبرز درتاج

دانشجو:
محمد حسین ضرغامی

پایان‌نامه جهت دریافت درجه کارشناسی ارشد در رشته سنجش و اندازه‌گیری

شهریور ماه ۱۳۸۸

تقدیم به روح بلند پدر که برایم بهترین بود و با عمل و رفتار خود درس‌ها به من آموخت و تقدیم به تمام کسانی که نیروهای بیکران وجود آدمی را بیدار می‌کنند تا آدمی با گیتی پیوند بخورد و به اوج لذتی ابدی برسد.

سپاس خداوندی که اندیشه را برای پی بردن انسان به جهلش در او قرار داد

بسیار ممنون و متشکرم از استاد بزرگ و مهربانم جناب آقای **دکتر علی دلاور** که زحمت راهنمایی این پایان نامه را متقبل شدند. همچنین از استاد گرامی جناب **دکتر فریبرز درتاج** متشکرم که علاوه بر مشاورهی این پایان نامه در طول دوره مشوق بسیار خوبی برایم بودهاند.

بیشتر سختیهای این پژوهش بر دوش همسر عزیزم بود، از ایشان نیز نهایت تشکر وسپاس را دارم و امیدوارم زندگی همیشه بر وفق مرادش باشد.

نهایت سپاس و تشکر را از گسترشدهندگان شبکههای اینترنتی دارم که دسترسی به منابع معتبر را برایمان آسان کردهاند و در جهت عدالت اطلاعات گام مهمی را برداشته و معیارهای موفقیت را عوض کردهاند. سپاسگذار تمام کسانی هستم که مانند خورشید پشت ابر در عین هدایتگری ناشناخته ماندهاند.

چکیده‌ی پژوهش:

این پژوهش به مقایسه‌ی دو روش غالب در تعیین عملکرد افتراقی سوال می‌پردازد. این دو روش عبارتند از: روش رگرسیون لجستیک و روش منتل هنزل. مبنای این مقایسه، آماره‌ی اندازه‌ی اثر می‌باشد. در این مطالعه تنها به بررسی DIF یکنواخت پرداخته شده است. در ضمن پارامترهای مختلف سوال و بزرگی DIF موجود در داده‌های تجربی دستکاری شده و از طریق اطلاعات بدست آمده داده‌های لازم شبیه سازی شده‌اند. نتایج نشان می‌دهد که کاربرد اندازه‌ی اثر دقت و صحت کشف سوالات دارای DIF را در هر دو روش MH DIF و LR DIF بهبود می‌بخشد. این بهبود برای روش MH DIF در مقایسه با روش LR DIF ناچیز است.

واژگان کلیدی: عملکرد افتراقی سوال، اندازه اثر، رگرسیون لجستیک، منتل هنزل.

فهرست مطالب

فصل اول

کلیات پژوهش

مقدمه

آزمون‌های استاندارد شده[۱] و سایر اندازه گیری‌ها در حوزه علوم رفتاری ابتدا به منظور تمایز بین مهارت‌ها و سطوح مختلف توانایی مورد استفاده قرار می‌گرفتند. بعضی از این حوزه‌ها عبارت‌اند از: توانایی‌های دانشگاهی، پژوهش‌های شغلی و ویژگی های شخصیتی. عملکرد افتراقی سوالات[۲] نیز ابتدا به عنوان بخشی از روایی[۳] مدنظر بود. وظیفه ی عملکرد افتراقی سوالات در پژوهش‌های روایی تعیین درجه‌ی تفاوت توانایی‌های واقعی آزمودنی ها در رفتاری غیر سودار است. آزمون سازان و متخصصان روان‌سنجی از عملکرد افتراقی سوال به منظور تعیین امکان وجود سوگیری در یک سوال استفاده می کنند، البته محدوده‌ی کاربرد عملکردهای افتراقی سوال از شناسایی سوالات دارای سوگیری بالقوه در امتحانات مدرسه و آزمون‌های استخدامی تا کاربرد تحلیل‌های پیشرفته نظریه‌ی سوال پاسخ به منظور شناسایی پیچیدگی‌های ممکن در زمان تفسیر نتایج پیمایش‌های اجتماعی می‌باشد(راست و گلومبک[٤]، ۲۰۰۹). این پژوهش شرایطی را مورد مطالعه قرار می دهد که در آن دو روش غالب در تعیین عملکرد افتراقی سوال با یکدیگر مقایسه می‌شوند. مبنای این مقایسه، آماره‌ی "اندازه ی اثر[٥] " است. به طور قطع نمرات آزمون از منابع واریانسی غیر از توانایی‌های آزمودنی‌ها، اثر می پذیرد. در صورتی که آزمون‌ها به طور کامل آنچه که مد نظر پژوهشگر است را بسنجند، همه نمرات روا و پایا[٦] خواهند بود ولی هرگز نمی توان منابع نامربوط واریانس را به طور کامل کنترل کرد، بنابراین باید در استفاده از نتایج آزمون‌ها بسیار دقت شود تا از قضاوت‌های ناعادلانه به نفع یا ضرر گروه یا گروه‌های خاصی اجتناب شود. این قضاوت‌های ناعادلانه زمانی رخ می‌دهند که دو گروه جامعه از لحاظ توانایی‌های مورد سنجش یکسانند ولی منابع نامربوط واریانس برای آنها به طور متفاوتی توزیع شده است(کورویل[۷]، ۲۰۰٤).

---

۱. Standardized tests

۲. Differential Item Functioning (DIF)

۳. validity

٤. Rust & Golombok

٥. Effect size

٦. Valid & Reliable

۷. Courville

روش آماری که در تعیین سوگیری[۱] سوال مورد استفاده قرار می‌گیرد، تحت عنوان عملکرد افتراقی سوال (DIF) نامیده می شود. DIF مشاهده‌ی ویژگی های آماری متفاوت سوال در دو زیر جامعه است که فرض می‌شود سطح توانایی مورد سنجش در آنها برابر است (هلند و واینر[۲]، ۱۹۹۳). DIF یک فرایند دو مرحله ای است: گام اول مقایسه‌ی بروندادهای زیر گروه‌ها در یک سوال و تعیین وجود یا عدم وجود DIF در آن سوال می‌باشد. گام دوم تصمیم‌گیری در مورد این موضوع است که آیا تفاوت بین گروه‌ها(زیر جامعه[۳] ها) به اندازه‌ی کافی بزرگ است که بتوان سوال را حذف کرد و یا تغییر داد. مرحله‌ی دوم در برگیرنده‌ی آزمون‌های معناداری آماری مربوط به DIF است(فن[٤]، ۱۹۹۸).

محدودیت اساسی در استفاده از آزمون‌های معناداری آماری [٥] در هر دو مرحله‌ی تعیین DIF و برای سوالی که سوگیری ندارد، حجم بالای نمونه است که ممکن است منجر به خطای نوع اول یا مثبت غلط[٦] می‌شود(تامپسون[۷]، ۲۰۰۲). بیشتر روش‌های تعیین DIF، آزمون‌های معناداری آماری را به منظور ارزیابی DIF به کار می گیرند. کاربرد ارزش‌های احتمالی(ارزش p[۸]) و آماره ی x^2 (خی دو) در برابر حجم های متفاوت نمونه مقاوم[۹] نیستند. بنابراین چون حجم نمونه[۱۰] در مطالعات مختلف یکسان نمی باشد، مقایسه‌ی آنها با یکدیگر امکان پذیر نیست. ناتوانی در مقایسه‌ی مطالعات مختلف یک ضعف محسوب شده که مانع تعمیم یک مطالعه می شود(کیرک[۱۱]، ۲۰۰۱). استفاده از اندازه ی اثر به منظور کمی سازی[۱۲] DIF تنها

---

۱ . Bias

۲ . Holland and wainer

۳ . Subpopulation

٤ . Fan

٥ . Test of statistical significance

٦ . False positive

۷ . Thompson

۸ . P value

۹ . Robust

۱۰ . Sample size

۱۱ . Kirk

۱۲ . Quntitative

بزرگی[1] مقدارDIF را در بر نمی‌گیرد بلکه علاوه بر بزرگی، تعمیم‌پذیری[2] و تکرارپذیری[3] آن را نیز شامل می‌شود(هیوبرتی[4]، ۲۰۰۲). بنابراین نیاز به استفاده از اندازه‌ی اثر به عنوان یک آماره‌ی مکمل و به منظور کنترل خطای نوع اول به شدت احساس می‌شود(کیرک، ۱۹۹٦) علاوه بر این زمانی که در یک سوال DIF وجود دارد باید مقداری کمی از آن ارائه داد(پتنزا و دورانز[5]، ۱۹۹٤). دو محدودیت دیگر در روش شناسی[6] DIF که بر کشف DIF اثر می‌گذارند، توزیعات مختلف توانایی و توزیعات متفاوت جامعه‌ی داده‌ها است. ممکن است مفروضه‌ی نرمال بودن توزیع صفت در جامعه که توسط بسیاری از پژوهشگرن و متخصصان روان سنجی ارائه شده است، برآورده نشود. پیرسون[7](۱۸۹۵) مساله بسیار مهمی را بیان کرد. وی به بررسی ٤٤۰ مجموعه داده واقعی پرداخت تا ویژگی توزیع آنها را مشاهده کند، از این مجموع فقط در حدود ۳/۲٪ آنها دارای توزیع نرمال و نسبتا نرمال بودند(میسری[8]، ۱۹۸۹). سوینی[9](۱۹۹٦) معتقد است ناهماهنگی در توزیعات توانایی منجر به ناپایی کشف DIF می شود. این در حالی است که مفروضه‌ی هماهنگی توزیعات توانایی در گروه‌های مرجع و هدف یکی از مفروضات اصلی DIF می‌باشد.

از دهه‌ی ۱۹٦۰ که روش‌های DIF به عنوان بخشی از پژوهش‌های اندازه‌گیری مطرح شد(آنگوف[10]،۱۹۹۳)، پژوهشگران به ضعف های موجود در فرایند کشف DIF پی بردند. یکی از خطاهای بارز و معمول در روش شناسی DIF، خطای نوع اول است. این محدودیت مخصوصا در روش‌هایDIF مبتنی بر نظریه سوال-پاسخ، روش‌های مبتنی بر آماره‌ی خی دو و بالاخص آزمون‌های معناداری معمول، که براساس آنها

---

۱. Magnitude

۲. Generlizability

۳. Repeatability

٤. Huberty

٥. Potenza and Dorans

٦. Methodology

۷. Pearson

۸. Micceri

۹. Sweeney

۱۰. Angoff

معناداری تفاوت‌ها تعیین می‌شود، خود را بیشتر نشان می‌دهد. به منظور اصلاح این وضعیت باید روش‌های آماری استفاده شود که نسبت به حجم نمونه مقاوم و خدشه ناپذیر باشند.

در ٤٠ سال گذشته روش‌های متعددی برای کشف DIF گسترش‌یافته ولی تعداد کمی از این روش‌ها مورد مطالعه دقیق قرار گرفته اند. با پیشرفت روش شناسی DIF روش‌های قدیمی‌یا کنار گذاشته می‌شدند و یا از طریق روش‌های کامل‌تر و جدیدتر استنتاج می‌شدند و نواقص و محدودیت‌های آنها آشکار می‌شد. روش‌های رایج DIF عبارت‌اند از: روش‌های مبتنی بر مدل‌های IRT، روش‌های مبتنی بر آماره‌ی x^2 (مثل روش استاندارد شده[1] و روش منتل هنزل[2]) و روش رگرسیون لجستیک[3]. روش‌های مبتنی بر مدل‌های IRT، روش منتل هنزل و رگرسیون لجستیک در پژوهش‌های مربوط به DIF مناسب‌ترند(سوامیناتان و رگرز[4]، ۱۹۹۰) و از بین این روش‌ها، روش منتل هنزل و رگرسیون لجستیک عملی‌ترند، چون اولا کاربرد آنها ساده‌تر است و در عین حال برای این روش‌ها آماره‌ی "اندازه اثر[5]" تعریف می‌شود که می‌توان از طریق آن صحت کشف DIF را افزایش داد. مطالعات شبیه سازی شده[6] توانایی ترکیب "اندازه اثر" را با روش‌های منتل هنزل(روسوس و استوت[7]، ۱۹۹٦) و رگرسیون لجستیک (جودوین و گیرل[8]، ۲۰۰۱) ثابت کرده‌اند. سوامیناتان و رگرز(۱۹۹۰) رگرسیون لجستیک را به عنوان روشی در کشف DIF معرفی کرده‌اند و این روش را با روش‌های MH DIF[9] مقایسه کرده‌اند. مزور و همکاران[10] (۱۹۹۵) کاربرد تحلیل رگرسیون لجستیک را در کشف DIF به عنوان "یک روش ماندنی و پا بر جا" می‌دانند.

---

[1]. Standardized method

[2]. Manthel Haenszel

[3]. Logistic regression

[4]. Swaminathan & Rogers

[5]. Effect size

[6]. Simulation studies

[7]. Roussos & Stout

[8]. Jodoin & Gierl

[9]. Mantel Heanseal differential item functioning(MH DIF)

[10]. Mazor

DIF می‌تواند تحت دو موقعیت مورد ملاحظه قرار گیرد: موقعیت یکنواخت[۱] و موقعیت غیر یکنواخت[۲]. DIF یکنواخت زمانی اتفاق می‌افتد که تعاملی بین سطح توانایی و عضویت در گروه وجود نداشته باشد، برعکس DIF غیریکنواخت زمانی اتفاق می‌افتد که تعامل بین سطوح توانایی و عضویت گروهی وجودداشته باشد(سوامیناتان و رگرز، ۱۹۹۰). DIF رگرسیون لجستیک (LR DIF[۳]) با DIF روش منتل هنزل (MH DIF) مقایسه شده است. مقایسه نشان می‌دهد که رگرسیون لجستیک هم در DIFیکنواخت و هم در DIF غیریکنواخت بهتر عمل می‌کند(سوامیناتان و رگرز، ۱۹۹۰).

"اندازه‌ی اثر" حداقل مجذورات وزنی R^2 (WLS R^2[۴]) توسط زومبو و توماس[۵](۱۹۹۶) معرفی شده است. این "اندازه اثر" توسط جودین و گیرل[۶](۲۰۰۱) و همچنین روسوس و استوت[۷](۱۹۹۶)، به طور تجربی و از طریق داده‌های شبیه سازی شده مورد آزمایش قرار گرفت. تمرکز جودوین و گیرل بر "اندازه‌ی اثر" مربوط به LR DIF بود. آنها به عنوان راهنما، جدول طبقه‌بندی شده‌ای ارائه کردند که از طریق آن "اندازه اثر" بالا، متوسط و ضعیف(قابل اغماض[۸]) مشخص می‌شد. مقیاس گذاری این راهنما شبیه راهنمای بزرگ، متوسط ، کوچک کوهن[۹](۱۹۹۲) مربوط به اندازه می‌باشد. روسوس و استوت(۱۹۹۶) به مقایسه‌ی MH DIF و SIBTEST[۱۰] پرداختند و از "اندازه اثر" در کنار آزمون‌های آماری استفاده کردند تا به این وسیله میزان خطای نوع اول را کاهش دهند. روش‌هایMH DIF و LR DIF می‌توانند از "اندازه‌ی اثر" در کنار

---

۱. Uniform

۲. Non-Uniform

۳. Logistic regression DIF

۴. Weighted least squares

۵. Zumbo & Thomas

۶. Jodoin & Geirl

۷. Rousos & stout

۸. Negligible

۹. Cohen

۱۰. Simultaneous Item Bias TEST(SIBTEST)

آزمون‌های آماری بهره گیرند. این دو روش به صورت تجربی توسط هیدالگو و لوپزپینا[1](۲۰۰٤) با هم مقایسه شده اند.

تعاریف نظری و عملیاتی

عملکرد افتراقی سوال: DIF مشاهده‌ی ویژگی های آماری متفاوت سوال در دو زیر جامعه است که فرض می‌شود سطح توانایی مورد سنجش در آنها برابر است (هلند و واینر[2]، ۱۹۹۳). در این پژوهش DIF بر مبنای فرمول مساحت راجو[3] بدست می‌آید. در صورتی که مساحت بدست آمده بین دو ICC گروه مرجع[4] و هدف[5] بیشتر یا مساوی ٤/ باشد آن سوال به عنوان یک سوال دارای DIF شناخته می‌شود.

اندازه اثر: "اندازه اثر" عنوانی است که به مجموعه‌ای از شاخص‌ها که بزرگی اثر آزمایش را می سنجد اطلاق می‌شود. "اندازه‌ی اثر" در آمار، مقداری است که رابطه‌ی بین دو متغیر را بیان می‌کند. در آزمایشات علمی علاوه براین که ما باید از معناداری آماری باخبر باشیم؛ باید از اثرات مشاهده شده نیز مقداری کمی داشته باشیم. برای تصمیم گیری در موقعیت های عملی"اندازه‌ی اثر"، شاخص بسیار مناسبی است(گریسوم و کیم[6]، ۲۰۰۵). در این پژوهش دو اندازه‌ی اثر وجود دارد یکی متعلق به روش رگرسیون لجستیک است و دیگری متعلق به روش منتل هنزل. "اندازه‌ی اثر" حداقل مجذورات وزنی R^2 (WLS R) که توسط زومبو و توماس[7](۱۹۹٦) معرفی شده است، به عنوان اندازه اثر روش گرسیون لجستیک مد نظر است و لگاریتم نسبت شانس معرفی شده توسط زویک و اریکان(۱۹۸۹) به عنوان اندازه اثر روش منتل هنزل قلمداد می‌شود.

گروه مرجع و گروه هدف: در مطالعات عملکرد افتراقی سوال دو گروه مد نظر است گروه مرجع و گروه هدف. گروهی که سوال نسبت به انها سوگیری دارد گروه هدف است و بر عکس گروهی که سوال به نفع

---

[1]. Hidalgo & Lopezpina

[2]. Holland and wainer

[3]. Raju's Area Formula

[4]. Reference group

[5]. Focal group

[6]. Grissom & Kim

[7]. Zumbo & Thomas

آنها است گروه مرجع تشخیص داده می‌شود. در این مطالعه گروه زنان به عنوان گروه هدف و گروه مردان به عنوان گوه مرجع می‌باشند.

ارزش P: سطح معناداری که بر مبنای آن معناداری آماری مشخص می‌شود. در این تحقیق این مقادیر از طریق نرم افزار SAS بدست می‌آید.

ضرورت پژوهش و بیان مساله

در حال حاضر اکثر روش‌های کشف DIF برای تعیین وجود DIF دریک سوال بر ارزش‌هایP متکی‌اند. هر دو روش MH DIF و LR DIF "اندازه‌ی اثر"ی دارند که می‌تواند به عنوان مکمل در کنار آزمون‌های آماری معمول (مبتنی بر ارزش p) قرار بگیرند(زومبو و توماس[1]، ۱۹۹٦). "اندازه‌ی اثر" می‌تواند صحت کشف DIF را افزایش دهد چون زمانی که با حجم های مختلف نمونه مواجه‌ایم "اندازه‌ی اثر" پایاتر از آزمون‌های آماری معناداری رایج می‌باشد. آزمون‌های آماری در برابر حجم نمونه مقاوم نیستند، فرض بر این است که کاربرد "اندازه‌ی اثر" منجر به کشف دقیق تر و صحیح تر DIF موجود دریک سوال می‌شود(فینچ و سایرین[2]، ۲۰۰۱).

یک مطالعه (هیدالگو و لوپزپینا، ۲۰۰٤) موجود است که روش‌های MH DIF و LR DIF را مقایسه کرده است. نتیجه این مطالعه نشان می‌دهد زمانی که از LR DIF استفاده می‌کنیم سوالات دارای DIF نسبت به زمانی که از MH DIF استفاده می‌کنیم، بیشتر است و در عین حال روش LR DIF به شرایط خاصی در کشف DIF حساسیت ندارد. شرایط خاصی که در روش شناسی DIF و برای انواع مختلف آن (یکنواخت و غیریکنواخت) مدنظر اند عبارت‌اند از: حداقل حجم ۱۰۰۰ برای گروه هدف و مرجع و توزیع نرمال توانایی. البته بعضی از پژوهشگران حجم های کمتری را پیشنهاد می‌کنند و مناسب می‌دانند(هیدالگو و لوپزپینا، ۲۰۰٤).

زومبو و توماس(۱۹۹٦) که حداقل مجذورات وزنی[3] R^2 را به عنوان "اندازه اثر" روش LR DIF مطرح کردند، به سایر پژوهشگران پیشنهاد کردند مطالعات بیشتری برای آزمون دقت و صحت "اندازه اثر" در تعیین DIF انجام دهند. این دو پژوهشگر پایایی این دو روش را در موقعیت های مختلف بررسی نکردند و

---

[1]. Zumbo & Thomas

[2]. Finch

[3]. Weighted least squares R^2

همچنین مطالعه‌ای راجع به این که آیا "اندازه‌ی اثر" مربوط به هر روش، دقت کشف DIF را افزایش می‌دهد، انجام ندادند. بنابراین لازم است در این زمینه مطالعاتی صورت پذیرد. پژوهش حاضر در این راستا گام برمی‌دارد. در این پژوهش "کاربرد اندازه‌ی اثر در سنجش عملکرد افتراقی سوال بین دو روش رگرسیون لجستیک و منتل هنزل " مقایسه می‌شود.

اهداف پژوهش

هدف اصلی این پژوهش یک هدف بنیادی است که منجر به گسترش دانش در حوزه‌ی سنجش عملکرد افتراقی سوال می‌شود. کاربرد و استفاده از ارزش P به عنوان معیاری در تشخیص سوالات دارای DIF مخدوش و بی‌ارزش است(تامپسون و کیفر[۱]، ۲۰۰۰) بنابراین باید آماره‌ای جایگزین آن شود که مشکلات آن را نداشته باشد. در ضمن نیاز است تا قدرت اندازه‌ی اثر در کشف سوالات دارای DIF در روش‌های مختلفی که از این آماره استفاده می‌کنند مطالعه شود تا دریابیم در کدام روش در صورت استفاده نکردن این آماره نتایج نامعتبر است. به این ترتیب قضاوت‌هایی که در حوزه‌ی سنجش عملکرد افتراقی سوال بر مبنای دو روش رگرسیون لجستیک و روش منتل هنزل می‌شود، اصلاح خواهد شد.

سوالات پژوهش

در این پژوهش سه سوال عمده و اساسی مطرح است:

سوال اول: آیا کشف DIF دریک سوال هنگامی که از "اندازه اثر" مربوط به LR DIF یعنی حداقل مجذورات وزنی R^2 (WLS R^2) به عنوان مکمل آزمون‌های آماری معناداری (P Value) استفاده می‌شود، از دقت بیشتری برخوردار است؟

سوال دوم: آیا کشف DIF دریک سوال هنگامی که از "اندازه‌ی اثر" مربوط به MH DIF یعنی لگاریتم نسبت شانس[۲] (Log odds ratio) به عنوان مکمل آزمون‌های آماری معناداری (P Value) استفاده می‌شود، از دقت بیشتری برخورداراست؟

سوال سوم: کاربرد اندازه‌ی اثر در کشف سوالات دارای DIF در روش رگرسیون لجستیک موثرتر است یا در روش منتل هنزل؟

---

۱. Thompson & keiffer

۲. Log odds ratio

فصل دوم

پیشینه پژوهش

فصل دوم

مروری بر ادبیات پژوهش

واژه‌ی عملکرد افتراقی سوال(DIF) در آغاز دهه‌ی ۱۹۹۰ وارد ادبیات سنجش و اندازه گیری شد. البته قبل از این دهه، و از دهه ۱۹۶۰ واژه‌ای تحت عنوان "سوگیری سوال[۱] " مطرح بود که هدف آن گسترش کاربرد منصفانه و تخفیف اختلاف در عملکرد آزمون بین گروه‌های جامعه(به عنوان مثال سیاهان و اسپانیایی تبارها[۲]) بود(آنگوف، ۱۹۹۳).

بیشتر این مطالعات برای درک اختلاف نمرات زیر گروه‌های جامعه و مخصوصا برای اثبات این موضوع هدایت می‌شدند که تفاوت‌های مشاهده شده در نمرات، بیشتر ناشی از سوگیری سوالات می‌باشد و نه اختلاف در توانایی واقعی افراد. به طور کلی تمرکز اصلی مطالعات سوگیری سوال، کشف سوالاتی از آزمون است که ممکن است به نفع‌یا ضرریک‌یا چند گروه جانبداری کرده باشند. سوگیری سوال به صورت زیر تعریف شده است(آنگوف، ۱۹۹۳):

"یک سوال زمانی سوگیری دارد که احتمال پاسخگویی برای افرادی با توانایی‌ها و مهارت‌های‌یکسان، برابر نباشد."

شپارد و سایرین[۳](۱۹۸۱) ویژگی سوگیری را نوعی ناروایی[۴] می‌دانند که به یک گروه بیشتر از گروه دیگر ضربه وارد می‌کند." آنگوف(۱۹۹۳) معتقد است تعاریف مربوط به سوگیری نهایتا مطلبی را درباره‌ی عملکرد، ارزشیابی، اجرا ویا قضاوت های ناعادلانه ارائه می‌دهند.

آیا هر تفاوت موجود در عملکرد افراد را می‌توان به سوگیری سوالات نسبت داد؟ با این سوال مطالعات مربوط به سوگیری سوالات به چالش کشیده شد. ارتباط دادن تفاوت‌های موجود در عملکرد زیرگروه‌ها به

---

۱. Item Bias

۲. Hispanic

۳. Shepard et al

۱. Invalidity

ویژگی‌های سوال(مانند محتوی[1]، واژه بندی[2] و غیره) ویا به تفاوت در توانایی‌های واقعی آنها ذاتا امری دشوار است(آنگوف، ۱۹۹۳).

در زبان محاوره سوگیری و عادلانه نبودن خیلی به هم شبیه‌اند؛ ولی در زمینه‌ی اندازه گیری و تصمیم‌گیری، سوگیری و انصاف مفاهیم کاملا متفاوتی‌اند. سوگیری یک ویژگی آماری از نمرات آزمون می‌باشد. به همین دلیل گفته می‌شود وقتی سوگیری وجود دارد که خطاهای منظم ایجاد کند. در مقابل منصفانه بودن[3] به داوری درباره‌ی تصمیم گیری یا هر فعالیتی که براساس نتایج نمرات آزمون است، اشاره دارد. سوگیری یک ویژگی اعداد است در حالی که عادلانه بودن‌یک ویژگی تصمیمات‌یا ویژگی اشخاص است. اعداد اخلاقا و منطقا خنثی می‌باشند ولی تصمیمات و فعالیت‌های اشخاص می‌تواند عادلانه ویا ناعادلانه باشند. اعداد می‌توانند درست‌یا نادرست باشند ولی نمی‌توانند عادلانه یا ناعادلانه باشند. عادلانه یا ناعادلانه بودن‌یک آزمون به هدف و خواسته فرد از آزمون بستگی دارد نه صرفا به نمرات و اعداد که از آزمون بدست می‌آیند(آنگوف، ۱۹۹۳).

به طور کلی تفاوت‌های سوگیری و عادلانه بودن[4] در جدول مشاهده می گردد.

سوگیری	عادلانه بودن
معطوف به نمرات آزمون یا پیش‌بینی‌ها براساس آزمون می‌باشد.	معطوف به تصمیماتی است که گرفته می‌شود.
معطوف به ویژگی‌های آماری نمرات است.	داوری در مورد پیامدها است.
به صورت تجربی تعریف می‌شود.	براساس اصطلاحات فلسفی و سیاسی تعریف می‌شود.
به روش علمی تعریف می‌شود.	نمی‌تواند به روش علمی تعریف و تبیین شود.

[1]. Content

[2]. Wording

[3]. Fairness

[4]. Fairness

بطور کلی دو رویکرد برای اندازه گیری سوگیری وجود دارد(زومبو و توماس، ۱۹۹۶).

الف) قضاوت کارشناسان و ب)روش‌های آماری

باید تأکید کرد که کشف DIF دریک سؤال لزوماً به معنای سوگیر بودن آن سؤال نیست(یعنی جانبداری آن به نفع‌یک گروه و به ضرر گروه دیگر). از مفهوم جانبداری چنین به ذهن می‌آید که مقایسه‌ای هنجاری و رقابت آمیز با دیگر افراد انجام می‌شود، مثلاً در مواردی که نمره نسبتاً بالا باعث کسب نوعی پاداش و نمره نسبتاً پایین موجب بی‌بهرگی از آن می‌شود. اما همیشه چنین نیست. مثلاً در مواردی که دو گروه فرهنگی‌یا زبانی با ابزار واحدی اندازه گیری می شوند و هدف، صرفاً مقایسه‌ی زبانشناختی آن دو گروه است، حتی در مواردی که دریک سؤال DIF به وضوح مشاهد می‌شود ممکن است آن سؤال کاملاً عادلانه و بدون جانبداری باشد. این مطلب به هدف آن سؤال نیز مربوط می‌شود. برای نمونه اگر سوالی راجع به حرارت لازم برای پختن کیک باشد، چنین سؤالی حتی اگر از لحاظ آماری کاملاً به ضرر مردان سوگیر باشد، در صورتی که دریک آزمون گزینش آشپز و شیرینی‌پزی به کار رود کاملاً عادلانه خواهد بود. مثال دیگر این که دولیتل[۱] و کلیری[۲] مشاهده کردند که دانش آموزان دختر به سؤال های هندسه و استدلال ریاضی در مقایسه با دانش آموزان پسر در همان سطح توانایی ریاضی، ضعیف تر پاسخ می‌دهند(آنگوف، ۱۹۹۳). پس اگر بخواهیم از نتایج آماری به دست آمده، متابعت کنیم باید چنین سؤال هایی را از تست خارج سازیم. اما چنین کاری کاملاً بی معنی است چرا که فهم هندسه و توانایی استلال ریاضی اهداف موجه و لازم آموزش ریاضیات است و از دانش آموز انتظار می‌رود که مسائلی از این دست را حل کند. حذف چنین سوال‌هایی نه‌تنها تست را ناقص می‌کند بلکه به ضرر دانش آموزانی تمام می‌شود که نمی‌توانند به چنین سؤال هایی پاسخ دهند چرا که مشخص نمی سازد نقص دانش آموز در کجاست. دیگر این که در تحقیقات مربوط به سوگیری بسیار مشاهده شده که سؤال هایی دارای DIF شناخته می شوند، اما هیچ گونه شواهدی دال بر منشاء وجود این کارکرد افتراقی به دست نمی‌آید. باید خاطر نشان ساخت که صرف وجود تفاوت عملکرد دو گروه دریک سؤال به خودی خود دلیلی کافی بر وجود سوگیری به نفع گروهی که عملکرد قوی‌تری داشته، نیست. زیرا همان طور که می دانیم بخش اعظم تفاوت افراد و گروه‌ها در تست‌های شناختی، ناشی از ماهیت کیفیت و میزان تعلیم و تربیتی است که چه در مدرسه و چه در خارج از آن به آنها عرضه می‌شود. پس سؤال هایی را که کارکرد افتراقی دارند چگونه باید تفسیر و تعبیرکرد؟ مقدارDIF می‌تواند به این معنا باشد که سؤال مزبور در یک گروه علاوه بر سازه مورد نظر تست، سازه

---

[۱] - Doolittle

[۲] - Cleary