رویش روان شناسی، سال هفتم، شماره۶، شماره پیاپی ۲۷، شهریورماه ۱۳۹۷
Rooyesh-e-Ravanshenasi, Vol.7 (6), Summer2018

افت در پارامترهای سؤال: مفاهیم، روش‌شناسی و شناسایی

Item Parameter Drift: concepts, methodology and identifying

ابوالفضل قدمی (نویسنده مسئول)
کارشناس ارشد روانسنجی دانشگاه علامه طباطبایی

عرفان حسینی سامره
کارشناس ارشد روانسنجی دانشگاه آزاد اسلامی واحد تهران مرکز

Abolfazl Ghadami*
Graduate student of Psychometric at Allameh Tabataba'i University

Erfan Hosseini Samereh
Graduate student of Psychometric at Islamic Azad University Central Tehran Branch

چکیده

افت در پارامترهای سؤال هنگامی به وجود می‌آید که در طول زمان و بنابر دلایل مختلف، سؤالات آزمون ویژگی‌های اولیه خود مانند درجه دشواری و قدرت تشخیص را از دست بدهند. ازجمله موارد ایجادکننده افت در پارامترهای سؤال فاش شدن سؤالات، تکرار بیش‌ازاندازه، تغییر در نظام آموزشی کشور، موقعیت قرار گرفتن سؤال و پارامترپردازی ضعیف اولیه هستند. وجود افت در پارامترهای سؤال سبب نقض ویژگی تغییرناپذیری می‌شود. دو نوع افت یکنواخت (تغییر پارامتر تشخیص) و غیریکنواخت (تغییر در پارامتر دشواری) وجود دارد. ازنظر مفهومی و روش‌شناسی این مفهوم موازی با مفهوم کارکرد افتراقی سؤال است با این تفاوت که به‌جای تفاوت‌های گروهی دوره‌های مختلف زمانی بررسی می‌شوند. جهت شناسایی و ردیابی افت در پارامترهای سؤال پس از رعایت مفروضه تک‌بعدی بودن با کمی تغییر از روش‌های شناسایی کارکرد افتراقی سؤال استفاده می‌شود. پس از شناسایی افت باید مشخص شود که آیا تغییراتی در محتوا، شرایط انگیزشی، مدت‌زمان پاسخگویی و مکان سؤال در مقایسه با زمانی که سؤال طراحی شده به وجود آمده

Abstract

Item Parameter Drift occurs over time for various reasons; when test items lose their initial characteristics, such as difficulty and discrimination parameters. Including cases of item parameter drift are revealed, excessive repetition, changes in the education system, and the position of items and the parameters of poor initialization. Item parameter drift causes of the invariance to be violated. There are two types of uniform (change of discrimination parameter) and non-uniformity (change of difficulty parameter) drift. Conceptually and methodologically, this concept is parallel to the concept of differential functioning, with the difference that instead of group differences the periods are examined. In order to after observing the undimensionality, with a little change, the recognition of differential item functioning methods is used. After identifying the drift, it must be determined whether changes in the content, motivation conditions, response time and place of the item are compared with when the item was

است. اگر هرکدام از این شرایط وجود داشته باشد سؤال باید از مجموعه سؤال‌ها حذف شود و یا مجدداً پارامترپردازی شود.

کلیدواژه‌ها: افت پارامترهای سؤال، تغییرناپذیری، کارکرد افتراقی سؤال، تک‌بعدی بودن

created. If any of these conditions exist, the item must be removed from the set of items or be reset to the parameter.

Key word: Item Parameter Drift, Differential Item Functioning, Invariance, Undimentionality

مقدمه

در دهه‌های اخیر تمایل به سنجش هرچه دقیق‌تر در حوزه‌های مختلف آموزش افزایش‌یافته است (آناستازی و اوربینا[1]،۱۹۹۷). هرساله آزمون‌های بسیاری ساخته و در بخش‌های مختلف استفاده می‌شوند، اما یکی از مسائلی که بسیار حائز اهمیت است آن است که ابزار سنجش و آزمون از ویژگی‌های روان‌سنجی[2] مطلوبی برقرار باشد (شریفی و نیکخو، ۱۳۹۴). از مهم‌ترین ویژگی‌های روان‌سنجی روایی آزمون[3] است. به عقیده مسیک[4] (۱۹۸۹) منظور از روایی مناسب بودن[5]، بامعنا بودن[6] و مفید بودن[7] استنباط‌هایی است که از نمره‌های حاصل از آزمون برآورد می‌شود و برای تائید این‌گونه استنباط‌ها لازم است شواهدی جمع‌آوری شود (یونسی،۱۳۹۱). ازجمله مواردی که خطر جدی برای روایی آزمون به شمار می‌رود، افت در پارامترهای سؤال[8] (IPD) است. افت در پارامترهای سؤال به دلایل گوناگون رخ می‌دهد، اما در عمل احتمالاً به علت استفاده بیش‌ازحد آزمون در طول زمان ایجاد می‌شود. همچنین این میزان افت ممکن است به علت تغییرات برنامه درسی یا پیشرفت در تدریس و یا تمرین اتفاق بیفتد (باک و دیگران،۱۹۸۸). تدریس در کلاس‌های فوق‌برنامه در زمینه موردسنجش و آزمون مقدماتی نیز می‌تواند باعث تغییر در اصل عملکرد آزمون شود. همچنین این پدیده ممکن است به علت امنیت و کنترل ضعیف در نگهداری سؤالات و فاش شدن آزمون اتفاق بیفتد. یکی دیگر از منابع انحراف می‌تواند تغییر برنامه درسی باشد (گلدشتاین[9]، ۱۹۸۳). با توجه به موارد ذکرشده

[1] Anastasi & Urbina
[2] Psychometric Properties
[3] Validity
[4] Messick
[5] Appropriateness
[6] Meaningfulness
[7] Usefulness
[8] Item Parameter Drift
[9] Goldestein

و تهدیدی که این مسئله برای روایی آزمون ایجاد می‌کند، آشنایی با افت در پارامترهای سؤال و نحوه شناسایی و ردیابی این پدیده از اهمیت ویژه‌ای برخوردار است؛ بنابراین در این مقاله جهت آشنایی افراد در حوزه سنجش و اندازه‌گیری به معرفی مفاهیم، روش‌شناسی و شناسایی سؤالاتی که بالقوه دارای افت در پارامترهای سؤال هستند پرداخته شده است اما پیش از پرداختن به موضوع افت در پارامترهای سؤال ابتدا باید با مفاهیم تک‌بعدی بودن و نامتغیر بودن آشنایی کافی وجود داشته باشد.

تک‌بعدی بودن

پیش از بررسی افت در پارامترهای سؤال بایستی مفروضه تک‌بعدی بودن[1] در مورد مجموعه مجموعه داده‌ها برقرار باشد. به عقیده امبرتسون و رایس (۲۰۰۰)، مفروضه تک‌بعدی بودن هنگامی برقرار است که یک صفت مکنون به‌تنهایی کل واریانس مشترک میان سؤالات را تبیین کند (مینائی و فلسفی نژاد،۱۳۸۹) و یا می‌توان این مفروضه را به‌صورت دیگر نیز بیان کرد: هنگامی‌که مدل، یک تتای واحد برای هر آزمودنی در مجموعه سؤال‌ها برآورد کند و عوامل دیگری که به‌جز این تتای منحصر بفرد بر پاسخ به سؤال تأثیر می‌گذارند به‌عنوان خطای تصادفی در نظر گرفته شوند (دمارس[2]،۲۰۱۰).

یکی از راه‌های بررسی مفروضه تک‌بعدی بودن نمودار ارزش ویژه[3] است. به عقیده هورن[4] هورن[4] (۱۹۶۵) اولین ارزش ویژه باید به شکل قابل‌ملاحظه‌ای از عامل‌های بعدی در نمودار بیشتر باشد تا بتوان آن را به‌عنوان یک بعد غالب در نظر گرفت (فلسفی نژاد،۱۳۸۹)؛ اما پژوهشگرانی مانند مینائی و فلسفی نژاد (۱۳۸۹) و امبرتسون و رایس (۲۰۰۰)، روش‌های اکتشافی مانند درصد واریانس تبیین شده توسط عامل اول و نسبت ارزش ویژه عامل اول به دوم را به‌عنوان شاخصی برای تک‌بعدی بودن مناسب نمی‌دانند و پیشنهاد می‌کنند روش‌های جدیدی مانند NOHARM,TESTFACT,DIMTEST,DETECT,H&R,MSP جهت بررسی تک‌بعدی بودن استفاده شود.

ویژگی نامتغیر بودن

---

[1] Unidimensionality
[2] Demars
[3] Eigenvalue
[4] Horn

ویژگی نامتغیر بودن[1] پارامترهای سؤال به علت قابل‌مقایسه کردن نمرات در آزمون‌هایی که توانایی یکسانی را می‌سنجند یکی از مهم‌ترین مسائلی است که در بسیاری از متون سنجش و اندازه‌گیری[2] به‌صورت برجسته‌ای به آن اشاره شده است (امبرتسون و رایس، ۲۰۰۰؛ وندرلیندن و گلس[3]، ۲۰۱۰). نامتغیر بودن پارامترهای سؤال را ثابت ماندن پارامترهای سؤال در برآوردهای مختلف و در طی دوره‌های زمانی متفاوت تعریف کرده‌اند که بر این اساس امکان مقایسه نمرات از فرم‌های مختلف آزمون امکان‌پذیر است (پارک، لی و زینگ[4]،۲۰۱۶). بر اساس راپ و زومبو (۲۰۰۶)، واژه‌ی نامتغیر بودن این مفهوم را می‌رساند که ارزش‌های پارامتر سؤال «درجه دشواری، قدرت تشخیص و پارامتر حدس» برای جمعیت آزمودنی‌های مجزا یا شرایط اندازه‌گیری مجزا یکسان است (راپ و زومبو[5]،۲۰۰۶). به عقیده لرد[6] (۱۹۸۰) و بیکر و کیم[7] (۲۰۰۴)، یکی از مهم‌ترین کاربردهای نامتغیر بودن را استفاده از این ویژگی در آزمون‌سازی نام برده‌اند که آزمون گیرندگان را به ساخت آزمون با استفاده از نظریه سؤال – پاسخ[8] ترغیب کرده است.

اهمیت ویژگی نامتغیر بودن در نظریه سؤال – پاسخ به‌عنوان سنگ بنای این نظریه نمی‌تواند اغراق‌آمیز باشد زیرا بدون پذیرش این مفروضه‌ی قاطع و محکم، پیچیدگی‌های مدل‌های این نظریه به‌سختی می‌تواند در زمینه تئوری و عملی تعدیل پیدا کند (فن[9]،۱۹۹۸). به عقیده راپ و زومبو (۲۰۰۶)، امبرتسون و رایس[10] (۲۰۰۰)، هرگونه تخطی از مفروضه نامتغیر بودن برآورد پارامترهای مدل را به خطر می‌اندازد و تفسیر امتیازبندی نمره شخص را دشوار می‌سازد (وی[11]،۲۰۱۳). مفروضه نامتغیر بودن در نظریه IRT باعث شده است تا مدل‌های این نظریه نسبت به دیگر نظریه‌های اندازه‌گیری برای تحلیل گران و توسعه‌دهندگان آزمون انتخاب بهتر و مطلوب‌تری فراهم آورد (وی،۲۰۱۳).

---

[1] Invariance
[2] Assessment And Measurement
[3] Van Der Linden & Glass
[4] Park, Lee & Xing
[5] Rupp & Zumbo
[6] Lord
[7] Baker, Kim
[8] Item Response Theory
[9] Fan
[10] Embretson & Reise
[11] Wei

در این زمینه بسیاری از مطالعات تغییرناپذیری پارامتر دشواری را بیشتر از نامتغیر بودن قدرت تشخیص گزارش کرده‌اند (فن ۱۹۹۸؛ مک دونالد و پائونونن[1]، ۲۰۰۲؛ آددویین[2]، ۲۰۱۰). در عمل به دو دلیل که با نام‌های DIF[3] و IPD[4] شناخته می‌شوند ویژگی نامتغیر بودن در نظریه سؤال–آزمون نقض می‌شود: ۱–تفاوت در نمرات گروه‌های مختلف جنسیتی، فرهنگی، مذهبی باوجود توانایی یکسان که کارکرد افتراقی سؤال (DIF) نام دارد.۲–افت پارامترهای سؤال در طول زمان به سبب عدم امنیت بانک سؤال و تغییر برنامه درسی که افت پارامترهای سؤال (IPD) نام دارد (راپ و زومبو، ۲۰۰۶).

افت پارامتر سؤال[5]

جهت ارزشیابی و سنجش پیشرفت تحصیلی دانش آموزان در سطح وسیع، اغلب مجموعه‌ای از سؤال‌ها که پارامترهای ویژه خود را دارند در قالب بانک سؤال حفظ می‌شوند تا در دوره‌های مختلف از این سؤال‌ها استفاده شود (ریو[6] و همکاران،۲۰۰۷). این سؤال‌ها در اجراهای متعدد باید پارامترهای خود را حفظ کنند تا بتوان نمرات افراد در دوره‌های مختلف را باهم مقایسه کرد؛ اما بنابر دلایلی این پارامترها در اجراهای مکرر تغییر می‌کند که به این تغییرات افت پارامتر سؤال (IPD) می‌گویند (وی[7]،۲۰۱۳).

در مبحث آزمون‌ها، «افت پارامتر» دلالت بر اختلاف یا تغییری در پارامترهای سؤال آزمون (درجه دشواری، قدرت تشخیص) از یک آزمون به آزمون دیگر در طی دوره‌های مختلف زمانی دارد (لی[8]،۲۰۱۸). این پدیده هنگامی اتفاق می‌افتد که یک آزمون در طی دوره‌های مختلف زمانی اجرا شود ولی پارامترهای سؤال برای آزمودنی‌ها تغییر کند (گو، ژنگ، چانگ[9]،۲۰۱۵). چیزی که در کل از این مفهوم می‌توان استنباط کرد آن است که IPD به سبب تغییر در یک یا بیشتر از یک پارامتر (دشواری و قدرت تشخیص) در طول زمان اتفاق می‌افتد (گلدشتاین،۱۹۸۳).

---

[1] Mcdonald
[2] Adedoyin
[3] Differential Item Function
[4] Item Parameter Drift
[5] Item Parameter Drift
[6] Reeve
[7] Wei
[8] Lee
[9] Guo, Zheng & Chang

در طول زمان سؤالات موجود در بانک سؤال بارها مورداستفاده قرار می‌گیرد و سؤالات استفاده‌شده «به علت آشنایی» می‌تواند در روز آزمون به‌طور نامناسبی، بر روی آزمودنی‌ها تأثیرگذار باشد (باک[1]، موراکی[2] و پفنبرگر[3]،۱۹۸۸). عدم رعایت نکات امنیتی در نگهداری و استفاده مجدد از سؤالات آزمون در طول زمان ازجمله مواردی است که به روایی آزمون لطمه می‌زند (هان[4]،۲۰۱۱).

به عقیده گلدشتاین (۱۹۹۳) مطالعه و بررسی افت پارامترهای سؤال در آزمون‌های روند[5]، روند[5]، به متخصصان آزمون ساز کمک می‌کند تا بدانند یک سؤال تا چه هنگام پارامترهای اولیه خود را حفظ می‌کند و قابلیت اجرا دارد و چه هنگام آزمون باید مورد تجدیدنظر قرار گیرد تا دوره‌های مختلف قابل‌مقایسه باشند و سبب تصمیم‌گیری غیرمنصفانه نشود (پارک، لی و زینگ،۲۰۱۶).

عوامل ایجادکننده‌ی افت پارامتر سؤال

افت پارامترهای سؤال هنگامی اتفاق می‌افتد که باوجود ثابت درنظر گرفتن پارامترهای معمول سؤال، این پارامترها بعد از چندین بار اجرا بنابر دلایل مختلف نوسان پیدا کنند (بابکوک و آلبانو[6]،۲۰۱۲). این نوسانات اغلب اوقات می‌تواند شامل تغییرات در درجه دشواری دشواری چه از نوع آسان‌تر و یا سخت‌تر شدن سؤال باشد که هرکدام بنابر شرایط و دلایل مختلفی ازجمله تغییرات سیستم آموزشی، دلایل فنی یا فرهنگی رخ دهد و عملکرد سؤال را تحت تأثیر خود قرار دهد (وینر[7] و دیگران،۲۰۱۰).

به گفته باک و دیگران (۱۹۸۸) از دیگر دلایل این تغییرات می‌توان به اثرات موقعیت‌های مختلف که در آن شرایط آزمون گرفته می‌شود اشاره کرد. کولن و برنان (۱۹۹۵) دلایلی را که در افت پارامتر سؤال می‌توانند دخیل باشند نام‌برده‌اند: نقص در سنجش اولیه و پارامترپردازی سؤال‌ها؛ جایگاه سؤال در فرم‌های مختلف؛ تغییر در طراحی برگه پاسخنامه؛ تغییر دادن جایگاه سؤال در پرسشنامه؛ تغییر در فونت و یا صفحه‌بندی برگه امتحانی؛ برگزاری امتحان در شرایط

[1] Bock
[2] Muraki
[3] Pfeiffenberger
[4] Han
[5] Trend
[6] Babcock & Albano
[7] Wainer

غیراستاندارد. همچنین نگهداری بلندمدت سؤال‌ها در بانک سؤال نیز به سبب استفاده بیش‌ازحد سؤال‌ها باعث آشنایی پاسخ‌دهندگان با سؤال‌ها می‌شود و درنتیجه درجه دشواری سؤال‌ها ممکن است کاهش پیدا کند و روایی آزمون را تحت تأثیر خود قرار دهد (وندرلیندن و گلس[1]، ۲۰۱۰).

شیوه‌های مختلفی برای جلوگیری از مواجهه با بروز افت پارامترهای سؤال وجود دارد که متخصصان آزمون ساز با کاربرد آن‌ها از بروز این‌چنین پدیده‌هایی جلوگیری می‌کنند. آن‌ها باید سؤال‌های موجود در بانک سؤالات را به مرور زمان به‌روزرسانی و تجدیدنظر قرار دهند. به این صورت که پارامترهای سؤال را در جامعه‌های کوچک مورد پارامتر پردازی مجدد قرار دهند و سؤالات نامناسب را از بانک سؤال استخراج کنند که در این زمینه روش‌های متفاوتی برای موردبررسی قرار دادن و داشتن یک بانک سؤال مناسب پیشنهادشده است (آریل[2] و وندرلیندن ۲۰۰۶، لی و وون دیویر[3]، ۲۰۱۳).

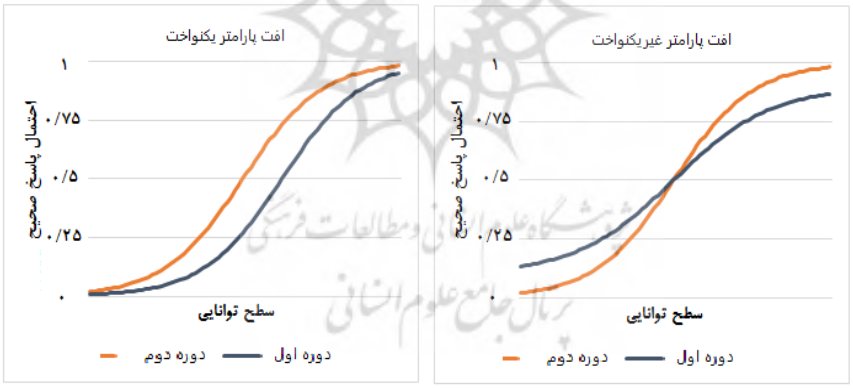
اگر هرکدام از شرایط ذکر شده وجود داشته باشد، سؤال باید به‌عنوان یک سؤال جدید در نظر گرفته شود که دارای پارامترهایی متفاوت از سؤال اصلی است و باید پارامترهای جدیدی برای این سؤال برآورد شود و جایگزین سؤال قبلی شود تا در آینده به‌عنوان سؤالی با پارامترهای جدید مورداستفاده قرار گیرد. در عمل کارشناسان آزمون توصیه به حذف یا برآورد مجدد سؤال‌هایی که دارای افت هستند می‌کنند (کولن و برنان[4]،۱۹۹۵).

اهمیت بررسی افت پارامترهای سؤال را می‌توان ازنظر اجرای آزمون‌ها به‌صورت مکرر نیز بررسی کرد، زیرا این سؤال‌ها در طول زمان بارها و بارها در خزانه سؤال‌ها مورداستفاده قرار می‌گیرند و پی بردن به افت سؤال‌ها به طراحان سؤال کمک می‌کند تا بتوانند تغییراتی که یک سؤال در یک مجموعه آزمون ایجاد می‌کند را بفهمند و در صورت نیاز آن سؤال را تغییر و حذف و یا سؤال دیگری را جایگزین آن کنند (برگستورم، استاهل و نتسکی[5]،۲۰۰۱).

افت یکنواخت و غیریکنواخت

[1] Van Der Linden & Glas
[2] Ariel
[3] Lee, & Von Davier
[4] Colin Brenan
[5] Bergstrom, Stahl & Netzky

ازآنجاکه تعاریف افت در پارامترهای سؤال و کارکرد افتراقی مفاهیمی موازی و از نظر روش‌شناسی نیز تا حدود زیادی شبیه به یکدیگر هستند بر اساس تعریف ملنبرگ[1] (۱۹۸۲) دو نوع متفاوت افت (یکنواخت[2] و غیریکنواخت[3]) را در داده‌های دو ارزشی می‌توان شناسایی کرد.۱- افت یکنواخت: هنگامی‌که بین سطح توانایی و شرکت در دوره‌های زمانی مختلف تعاملی وجود نداشته باشد این نوع از افت رخ می‌دهد که تفاوت ثابتی را در سراسر پیوستار متغیر نهفته از خود نشان می‌دهد. به‌طورکلی در نوع یکنواخت یک دوره زمانی در همه سطوح توانایی، برتری نسبی نسبت به دوره‌های دیگر دارد.۲- افت غیریکنواخت: افت غیریکنواخت برخلاف نوع یکنواخت هنگامی‌که بین سطح توانایی و دوره زمانی تعامل وجود داشته باشد اتفاق می‌افتد. سؤالاتی که افت غیریکنواخت دارند در داده‌های واقعی بیشتر اتفاق می‌افتد، برتری یک دوره ثابت نیست و ممکن است از یک دوره به دوره دیگر تغییر کند (نارایانان[4] و سوامیناتان،۱۹۹۶). جهت فهم بهتر تفاوت میان دو نوع متفاوت از افت به بیان تصویری در شکل ۱ پرداخته شده است.



شکل ۱: بیان تصویری افت یکنواخت و غیریکنواخت

[1] Mellenbergh
[2] Uniform
[3] Non Uniform
[4] Narayanan

شکل سمت راست منحنی ویژگی برای افت از نوع غیریکنواخت و در سمت چپ افت از نوع یکنواخت نشان داده‌شده است. با توجه به بیان ویژگی‌هایی که برای دو نوع متفاوت از افت داده شد در تصویر درک روشنی از این مفاهیم به دست می‌آید.

انواع مختلف افت در نظریه IRT

در نظریه IRT درجه دشواری[1]، قدرت تشخیص[2] و حدس[3]، پارامترهای سؤال را تشکیل می‌دهند. در این نظریه پارامترهای سؤال بنابر ویژگی نامتغیر بودن باید مقادیر خود را حفظ کنند. بر این اساس و بدون درنظر گرفتن پارامتر حدس سه نوع افت می‌تواند وجود داشته باشد (نوع a، b و ab). نوع b اشاره به کاهش و افزایش در درجه دشواری در طول زمان از سنجش اولیه خود دارد، نوع a اشاره به تغییر در پارامتر قدرت تشخیص و نیز نوع ab ناشی از تغییر در هردو پارامتر قدرت تشخیص و درجه دشواری دارد؛ اما معمول‌ترین افت، نوع b یا تغییر در پارامتر دشواری است (باک و دیگران،۱۹۸۸). نوع b به دلایل مختلفی از قبیل تغییر در دانش و اطلاعات آزمودنی در طول زمان، فاش‌سازی امتحان به دلیل تقلب، تغییر در برنامه درسی، تمرین و سیاست‌های آموزشی، وقایع تاریخی به وجود آمده در آن جامعه و مسائل فنی آزمون به وجود آمده است. ازجمله مواردی که موجب آسان‌تر شدن سؤال می‌شود می‌توان به تکرار بیش‌ازحد، فاش شدن سؤال به دلیل تقلب، پارامترپردازی ضعیف اولیه، وقایع تاریخی، موقعیت قرار گرفتن سؤال و تغییرات فرهنگی نام برد و دلایلی که موجب سخت‌تر شدن سؤال می‌شود را می‌توان به تغییر در شیوه اجرا، سیاست، برنامه درسی نسبت داد (ریسک[4]،۲۰۱۵).

هنگامی‌که یک سؤال در طول زمان ازنظر پاسخگویی دشوارتر می‌شود، می‌تواند به دلایل مختلفی از قبیل آموزشی، فنی و تغییرات فرهنگی باشد. سنجش ضعیف اولیه سؤال‌ها می‌تواند سبب تغییرات در پارامتر دشواری سؤال، چه از نوع آسان‌تر شدن و یا سخت‌تر شدن سؤال‌ها شود. گاهی اوقات اصطلاحات فنی در طی زمان منسوخ می‌شوند و کاربرد خود را از دست می‌دهند. علاوه بر این، تغییر مکان سؤال در آزمون (جابجایی مکان سؤال) می‌تواند باعث

[1] Difficulty
[2] Discrimination
[3] Guessing
[4] Risk

دشوارتر و یا آسان‌تر شدن سؤال در هر دو موقعیت شود؛ مانند هنگامی‌که سؤال در پیش‌آزمون در یک مکان و در آزمون اصلی در مکان دیگر قرار داشته باشد (داناهو و ایشام[1]، ۱۹۹۸).

تأثیر افت بر روی سؤال‌ها

وجود افت پارامتر سؤال می‌تواند باعث تخطی از ویژگی نامتغیر بودن پارامترهای سؤال شود و خطری جدی برای اندازه‌گیری محسوب شود. افت پارامتر سؤال می‌تواند خطاهای اندازه‌گیری را بیش‌ازاندازه بزرگ کند و بر روی روایی محتوا و سازه یک سؤال تأثیرگذار باشند (مک کوی[2]،۲۰۰۹). وجود افت همچنین بر روی بسیاری از روش‌های آزمون‌سازی مانند هم‌ارز سازی و آزمون‌های تطابقی تأثیرگذار است و نیز می‌تواند باعث ایجاد خطای هم ارزسازی شود. علاوه بر این آنچه تاکنون در این مورد گفته شد، افت پارامتر برای سؤال‌هایی که نیاز به ثبات در طول زمان دارد نیز خطری جدی محسوب می‌شود (ولز، سوبکویاک و سرلین[3]،۲۰۰۲ و ولز، همبلتون، کرک پاتریک و منج[4]،۲۰۱۴). ثبات مقیاس برای به دست آوردن گزارشی ثابت از نمرات و قابل‌مقایسه کردن در اجراهای مختلف سؤال را می‌دهد مهم و ضروری است (گوا و وانگ[5]،۲۰۰۳).

افت پارامترها نه‌تنها بر روی ثبات و درستی مقیاس، بلکه بر روی روایی نقطه‌ی برش سؤال نیز تأثیرگذار است که می‌تواند موجب استنباط غلط و نادرست در مورد عملکرد آزمودنی‌ها در این‌گونه سؤال‌ها شود (پارک، لی، زینگ،۲۰۱۶). بنا به دلایلی که ذکر شد اغلب سؤال‌های آزمون به‌ناچار به‌مرورزمان منسوخ می‌شوند و نیاز به تجدیدنظر و کنار گذاشتن از خزانه سؤال[6] می‌شوند. بروز رسانی مکرر سؤال‌های آزمون امری مطلوب و پسندیده است، به این دلیل که محتوای سؤال‌های آزمون را از تغییرات برنامه درسی و تغییرات در نظام آموزش‌وپرورش و استفاده بیش‌ازحد سؤال در آزمون حفظ می‌کند. این فرآیند شامل کنارگذاری یا جانشینی سؤال‌های انتخاب‌شده در یک آزمون یا در یک مجموعه آزمون است؛ اما قبل از کنار گذاشتن سؤالات باید از عدم سودمند بودن آن‌ها اطمینان حاصل کرد. بر اساس گلدشتاین (۱۹۸۳)، به‌منظور ارزیابی صحیح از سودمند بودن سؤال‌های آزمون، محققان باید

---

[1] Donoghue, J. R., & Isham
[2] Mccoy
[3] Wells, Subkoviak & Serlin
[4] Hambleton, Kirkpatrick & Meng
[5] Guo & Wang
[6] Item Bank

چند پرسش زیر را مدنظر قرار دهند.۱- آزمون گیرنده تا چه اندازه می‌تواند یک سؤال را در آزمون‌های بعدی تکرار کند تا آن سؤال منسوخ شود؟۲-چه زمانی یک آزمون باید مورد ارزیابی مجدد واقع شود؟۳-چه هنگام محققان می‌توانند ادعاهایی معتبر در خصوص تغییرات پاسخ آزمودنی‌ها در طول زمان داشته باشند؟

روش‌های شناسایی افت پارامتر سؤال

به عقیده داناهو و ایشام (۱۹۹۸)، جهت شناسایی افت باید به دو نکته توجه ویژه‌ای شود؛ اولاً مشخص شود که مقدار شناسایی‌شده به علت وجود و یا عدم وجود افت در پارامترهای سؤال‌های آزمون مربوط می‌شود، ثانیاً نرخ خطای نوع اول باید کنترل شود تا با اطمینان بیشتری سؤال‌های شناسایی‌شده را به علت وجود افت در پارامترهای سؤال نسبت داد. میسلوی[1] (۱۹۸۲) پنج مرحله را برای شناسایی، پیش‌بینی و به حداقل رساندن افت پارامترهای سؤال پیشنهاد کرده است. این پنج مرحله به شرح زیر است:

مرحله اول: بررسی الگوی عملکرد منحنی پاسخ سؤال‌ها[2]؛

مرحله دوم: شناسایی «عناصر درسی» که پارامترپردازی و نمره‌گذاری بر آن اساس انجام شده است؛

مرحله سوم: بررسی نمودارهای سالانه از پارامترهای برآورد شده سؤال؛

مرحله چهارم: جدا کردن سؤال‌هایی که به نظر می‌رسد دستخوش تغییر شده‌اند؛

مرحله پنجم: کنار گذاشتن سؤال‌هایی که الگوی خاصی از افت را نشان می‌دهد.

در اولین مرحله، میسلوی (۱۹۸۲) پیشنهاد می‌کند هنگامی‌که محققان برای پی بردن به وجود افت پارامتر سؤال الگوهای عملکرد آزمودنی‌ها را به‌وسیله منحنی‌های ویژگی سؤال بررسی می‌کنند دقت لازم را داشته باشند. برای مثال: سؤال‌های آزمونی که مهارت‌های مختلفی را اندازه می‌گیرد ممکن است نرخ وجود افت را بالا ببرند، زیرا آن‌ها از مفروضه تک‌بعدی بودن تخطی کرده‌اند؛ به‌عبارت‌دیگر، این سؤال‌ها ممکن است ناخواسته بیش از یک سازه را اندازه بگیرند و در این صورت بهتر است این آزمون به‌عنوان یک آزمون چندبعدی در نظر گرفته شود. درنتیجه، منحنی‌های ویژگی سؤال ممکن است درجه دشواری را در طول زمان به‌طور نامناسبی ارزیابی کنند.

---

[1] Mislevy
[2] Item Characteristic Curve

در مرحله دوم، میسلوی (۱۹۸۲) توضیح می‌دهد که با شناسایی بانک‌های سؤالی که تحت تأثیر تغییرات برنامه درسی و تأکیدی که بر مباحث قرارگرفته‌اند می‌توان به وجود آمدن افت را پیش‌بینی و از وقوع آن جلوگیری کرد. در مرحله سوم میسلوی (۱۹۸۲) برای کنترل میزان افت توصیه می‌کند که توسعه‌دهندگان آزمون یا پژوهشگران برای شناسایی آن نمودارهای ویژگی سؤال را به‌صورت سالیانه برآورد و کنترل کنند. در این مرحله پژوهشگران به‌وسیله روش‌های آماری و یا نموداری به دنبال پارامترهای سؤالی هستند که دستخوش تغییرات شدند.

مرحله چهارم به حداقل رساندن مقدار افت است. همان‌طور که میسلوی (۱۹۸۲) شرح داده است این مرحله شامل جداسازی سؤال‌های آزمونی است که افت پارامتر را از خود نشان می‌دهند.

پنجمین و آخرین مرحله، کنار گذاشتن سؤال‌هایی است که الگویی از افت را نشان می‌دهند. بر اساس میسلوی (۱۹۸۲)، چندین دلیل برای وجود این نوع از افت وجود دارد. برای مثال، اگر یک یا چند سؤال آزمون قبل از اجرای سؤال بر روی دانش آموزان دستخوش تغییراتی مانند فاش شدن سؤال‌ها قبل از اجرای آزمون و تکرار بیش‌ازاندازه سؤال‌ها شده باشد، این سؤال‌ها به‌طورمعمول آسان‌تر می‌شوند و تمایز خوبی میان دانش آموزان قائل نمی‌شود.

روش‌های آماری اندازه‌گیری افت پارامترهای سؤال

جهت بررسی میزان افت پارامترهای سؤال از روش‌هایی همانند کارکرد افتراقی استفاده می‌شود. با این تفاوت که به‌جای بررسی گروه‌های مختلف، دوره‌های مختلف زمانی مورد بررسی قرار می‌گیرد. این روش‌ها به دودسته کلی مبتنی بر نظریه کلاسیک و نظریه‌های جدید اندازه‌گیری تقسیم می‌شوند. روش‌های مبتنی بر نظریه کلاسیک نمرات مشاهده‌شده را به‌عنوان متغیری همتا برای پیش‌بینی نمرات واقعی در نظر می‌گیرند (لرد و ناویک،۱۹۶۸). از مزیت و برتری‌های روش‌های مبتنی بر نظریه کلاسیک سادگی اجرا و کار با نمونه‌های با حجم کوچک است (امبرتسون و رایس،۲۰۰۰، ترجمه شریفی و دیگران،۱۳۸۸).

برخی از روش‌های مبتنی بر نظریه کلاسیک به این شرح است:

۱- روش منتل هانسزل: این روش توسط منتل و هانسزل[1] (۱۹۵۹) به‌عنوان روشی جهت مطالعه گروه‌های همتا شده معرفی شده است. روش منتل- هانسزل، از روش‌های ناپارامتریک

[1] Mantel, & Haenszel

شناسایی افت سؤال است. در این روش سؤال‌ها به‌صورت دوارزشی[1] نمره‌گذاری می‌شوند و برای تعیین آزمودنی‌ها با سطوح توانایی یکسان در دو گروه استفاده می‌شود. از مزیت‌های این روش ارائه آزمون آماری، برآورد اندازه اثر[2] و توانایی آن در نمونه‌های با حجم کم است؛ اما این روش تنها در تشخیص کارکرد افتراقی یکنواخت مناسب است و همین موضوع یکی از محدودیت‌های روش منتل هانسزل است.

۲- روش رگرسیون لجستیک[3]: این روش که توسط سوامیناتان و راجرز[4] (۱۹۹۰)، معرفی شده است مبتنی بر مدل‌سازی آماری احتمال پاسخ صحیح به سؤال، بر اساس عضویت در گروه و ملاک است که در آن ملاک معمولاً نمره کل آزمون است. روش رگرسیون لجستیک مانند روش منتل هانسزل، آزمون معناداری آماری و اندازه اثر را ارائه می‌کند. ازجمله مزایای این روش می‌توان به توانایی بررسی کارکرد افتراقی یکنواخت و غیریکنواخت اشاره کرد. رگرسیون لجستیک به خانواده بزرگ‌تری از تحلیل‌ها تعلق دارد که مدل‌های خطی عمومی[5] نامیده می‌شوند. رگرسیون لجستیک به‌صورت گسترده‌ای برای بررسی احتمال داده‌های دووجهی به‌عنوان یک تابع لجستیک برای یک یا بیش از یک متغیر پیش‌بین استفاده می‌شود. در این روش پاسخ سؤال به‌عنوان متغیر وابسته با توزیع برنولی[6] دووجهی در نظر گرفته می‌شود (اگرستی[7]،۲۰۰۷ به نقل از گرامی پور،۱۳۹۳).

۳- روش‌های دشواری سؤال تبدیل‌شده[8]: این روش توسط آنگوف و فورد[9] (۱۹۷۳) ارائه شده است که به آن طرح دلتا[10] نیز گفته می‌شود. در این روش پارامتر دشواری سؤال در هر یک از زیرگروه‌های جامعه به‌طور جداگانه محاسبه می‌شود، سپس مقادیر دشواری سؤال به مقیاس دلتا با میانگین ۱۳ و انحراف استاندارد ۴ تبدیل می‌شوند (باقی و فرارا[11]،۱۹۸۹). همبستگی مقادیر دلتا مربوط به دو گروه محاسبه می‌شود، به‌علاوه برای تمام سؤال‌ها نمودار مقادیر دلتا

[1] Biserial Item
[2] Effect Size
[3] Logistic Regression
[4] Swaminathan, H.& Rogers
[5] General Linear Model
[6] Bernoulli
[7] Agresti
[8] Transformed Item Difficulty
[9] Angof & Ford
[10] Delta
[11] Baghi & Ferrara

مربوط به هر زوج رسم می‌شود و با داده‌ها خط مستقیمی برازش داده می‌شود. سؤال‌هایی که مقادیر دلتا در آن‌ها با فاصله زیادی از این خط قرارگرفته‌اند دارای افت هستند (انیل[1]،۱۹۹۱). این روش به علت سادگی موردتوجه قرار گرفته است، اما نسبت به عدم یکسانی قدرت تشخیص سؤال‌ها حساس است. در میان روش‌های نامبرده دو روش رگرسیون لجستیک و روش منتل هانسزل به‌عنوان رایج‌ترین روش‌های ارزیابی افت پارامترها در بسیاری از مطالعات مورد مقایسه قرارگرفته‌اند.

در چارچوب رویکرد سؤال – پاسخ نیز چندین روش برای ارزیابی افت مطرح‌شده است که این روش‌ها در دو گروه کلی مساحت میان منحنی ویژگی سؤال در دوره‌های زمانی مختلف و روش آزمون‌های آماری یکسانی پارامترهای سؤال طبقه‌بندی می‌شوند (شولتز، ویتنی و زیکار[2]،۲۰۱۳). روش آزمون‌های آماری، شامل آزمون چند متغیره و آزمون t بر روی مقادیر پارامتر دشواری می‌شود. هرچند که منحنی ویژگی سؤال‌ها بر اساس پارامترهای سؤال‌ها ترسیم می‌شوند، اما روش بررسی مساحت بین دو منحنی ویژگی سؤال، مستلزم نگاه دقیق‌تری به تفاوت بین پارامتر سؤال‌ها است. در این روش منطقه بین دو منحنی ویژگی سؤال موردتوجه قرار می‌گیرد. پس از برآورد پارامتر سؤال‌ها و قرار دادن آن‌ها روی مقیاس مشترک، منحنی سؤال در دو گروه ترسیم می‌شود. درصورتی‌که فاصله بین دو منحنی ویژگی سؤال صفر باشد و درواقع دو منحنی بر هم منطبق باشند، افت وجود ندارد. برعکس هنگامی‌که مساحت بین دو منحنی سؤال صفر نیست افت با درجات مختلفی وجود دارد. هرچقدر این مساحت بیشتر باشد، افت نیز بیشتر می‌شود. همبلتون و راجو فرمول‌هایی برای محاسبه مساحت بین منحنی‌ها ارائه کرده‌اند (همبلتون، سوامیناتان و راجرز ۱۹۹۱، ترجمه فلسفی نژاد،۱۳۸۹). راجو نارسایی رویکرد محاسبه مساحت بین دو منحنی ویژگی سؤال را شناسایی کرده و روش‌های مبتنی بر چارچوب افت در پارامتر سؤال‌ها و آزمون را برای ارزیابی آن ارائه کرده است. ویژگی‌های این را می‌توان ۱- امکان کاربرد این شاخص برای سؤال‌های دو ارزشی و چندارزشی ۲- امکان کاربرد این شاخص در مدل‌های تک‌بعدی و چندبعدی نام برد (اوشیما و موریس[3]،۲۰۰۸).

[1] O'neal
[2] Shultz, Whitney & Zickar
[3] Oshima & Morris

بحث و نتیجه‌گیری

افت در پارامترهای سؤال به دلایل مختلف رخ می‌دهد، اما در عمل احتمالاً به علت استفاده بیش‌ازحد آزمون در طول زمان و تغییرات برنامه درسی اتفاق می‌افتد (باک و همکاران،۱۹۸۸). همچنین این پدیده ممکن است به علت امنیت و کنترل ضعیف در نگهداری سؤالات و فاش شدن آزمون اتفاق بیفتد. یکی دیگر از منابع انحراف می‌تواند تغییر برنامه درسی باشد (گلدشتاین[1]، ۱۹۸۳). بنا به دلایلی که ذکر شد اغلب سؤال‌های آزمون به‌ناچار به‌مرورزمان منسوخ می‌شوند و نیاز به تجدیدنظر و کنار گذاشتن از خزانه سؤال می‌شوند. بروز رسانی مکرر سؤال‌های آزمون امری مطلوب و پسندیده است، به این دلیل که محتوای سؤال‌های آزمون را از تغییرات برنامه درسی و تغییرات در نظام آموزش‌وپرورش و استفاده بیش‌ازحد سؤال در آزمون حفظ می‌کند. این فرآیند شامل کنارگذاری یا جانشینی سؤال‌های انتخاب‌شده در یک آزمون یا در یک مجموعه آزمون است؛ اما قبل از کنار گذاشتن سؤالات باید از عدم سودمند بودن آن‌ها اطمینان حاصل کرد. بر اساس گلدشتاین (۱۹۸۳)، به‌منظور ارزیابی صحیح از سودمند بودن سؤال‌های آزمون، محققان باید چند پرسش زیر را مدنظر قرار دهند. آزمون گیرنده تا چه اندازه می‌تواند یک سؤال را در آزمون‌های بعدی تکرار کند تا آن سؤال منسوخ شود؟ چه زمانی یک آزمون باید مورد ارزیابی مجدد قرار گیرد؟ چه هنگام پژوهشگر می‌تواند ادعایی معتبر در خصوص تغییرات پاسخ آزمودنی‌ها در طول زمان داشته باشد؟ هنگامی‌که سعی در شناخت علل به وجود آمدن افت پارامترهای سؤال را داشته باشیم باید به دنبال پاسخگویی به پرسش‌های زیر بود:

آیا تأکیدی که بر محتوای سؤال در حال حاضر می‌شود، در مقایسه با زمانی که سؤال برای بانک سؤال طراحی شده تغییری کرده است؟ آیا ممکن است سؤال طوری طراحی‌شده باشد که شرایط انگیزشی برای پاسخ‌دهنده‌های مختلف متفاوت باشد؟ آیا تغییرات یا بازبینی مجددی در متن اصلی سؤال یا گزینه‌ها در بانک سؤال‌ها صورت گرفته است؟

آیا سؤال محتوای نسبتاً جدیدی را پوشش می‌دهد؟ اگر پاسخ به سؤال چهارم مثبت است، آیا فرصت یادگیری محتوای جدید اندک بوده است؟ آیا محتوای سؤال ازنظر درسی منسوخ‌شده است؟ آیا تغییری در مدت‌زمان پاسخگویی به آن سؤال به وجود آمده است؟ اگر

[1] Goldestein

پاسخ این سؤال‌ها مثبت است بنابراین پارامترهای سؤال باید دوباره برآورد شود و برآورد جدید جایگزین پارامترهای قبلی در بانک سؤال شود. باید ارزش پارامترهای سؤال که در حال حاضر وجود دارد با هنگامی‌که از ابتدا در بانک سؤال تعیین‌شده است مقایسه شود؛ اما درصورتی‌که امنیت سؤال به خطر افتاده باشد و یا سؤال بیش‌ازاندازه تکرار و استفاده‌شده باشد و یا محتوای سؤال منسوخ‌شده باشد، سؤال باید از بانک سؤال‌ها حذف شود و سؤال‌های جدید جایگزین شود (کولن و برنان،۱۹۹۵).

منابع

امبرتسون، سوزان. ای و رایس، استیون، پی (۲۰۰۰). *نظریه‌های جدید روان‌سنجی برای روان شناسان (IRT)*، ترجمه شریفی، حسن پاشا؛ فرزاد، ولی‌الله؛ حبیبی، مجتبی و ایزانلو، بلال (۱۳۸۸)، تهران: انتشارات رشد.

شریفی، حسن پاشا و نیکخو، محمدرضا (۱۳۸۳)، *راهنمای سنجش روانی*، تهران: انتشارات رشد.

گرامی پور، مسعود (۱۳۹۳). ارزیابی توان آماری تحلیل رگرسیون لجستیک در آشکارسازی کنش افتراقی سؤال‌های آزمون، *فصلنامه مطالعات اندازه‌گیری و ارزشیابی آموزشی*.۴(۸).۲۱۱-۱۸۷.

مینائی، اصغر (۱۳۹۲). سنجش مقایسه پذیری سازه و تحلیل کارکرد افتراقی سؤال‌ها (DIF) و بلوک‌های (DTF) آزمون علوم پایه هشتم تیمز ۲۰۰۷ در بین دانش آموزان ایران و آمریکا، *اندازه‌گیری تربیتی*، ۴(۱۱)، ۱۰۹-۱۴۶.

مینائی، اصغر و فلسفی نژاد، محمدرضا (۱۳۸۹). روش‌های سنجش تک‌بعدی بودن سؤال‌ها در مدل‌های دو ارزشی IRT. *فصلنامه اندازه‌گیری تربیتی*، ۳، ۷۹-۹۸.

مینائی، اصغر. (۱۳۹۱). مدل‌پردازی تشخیصی شناختی(CDM)سؤال‌های ریاضیات تیمز۲۰۰۷ در دانش‌آموزان پایه هشتم ایران با استفاده از مدل یکپارچه با پارامترپردازی مجدد(RUM) و مقایسه مهارت‌های ریاضی دانش‌آموزان دختر و پسر. *پایان‌نامه دکتری، دانشگاه علامه طباطبایی*.

همبلتون، رونالد، ک؛ سوامیناتان، اچ، راجرز، اچ، جین. (۱۹۹۱). *مبانی نظریه پرسش و پاسخ*، ترجمه محمدرضا فلسفی نژاد (۱۳۸۹). تهران: انتشارات دانشگاه علامه طباطبایی.

یونسی، جلیل (۱۳۹۱). تحلیل داده‌های آزمون تیمز پیشرفته)۲۰۰۸:(توانمندی رویکرد بیزی مدلIRT چندسطحی. پایان‌نامه دکتری، *دانشگاه علامه طباطبایی*.

Adedoyin, O. O. (2010). Investigating the invariance of person parameter estimates based on classical test and item response theories. *International Journal of Educational Sciences*, *2*(2), 107-113.

Agresti, A. (2007). *an introduction to categorical data analysis*. New York: Wiley Interscience.

Anastasi, A. Urbina. S. (1997). *Psychological testing*. Upper Saddle River. NJ: Prentice Hall.

Angoff, W. H., & Ford, S. F. (1973). Item-race interaction on a test of scholastic aptitude. Journal of Educational Measurement, 10(2), 95-105.

Ariel, A., Van Der Linden, W. J., & Veldkamp, B. P. (2006). A Strategy for Optimizing Item-Pool Management. *Journal of Educational Measurement*,*43*(2), 85-96.

Babcock, B., & Albano, A. D. (2012). Rasch scale stability in the presence of item parameter and trait drift. *Applied Psychological Measurement*, 0146621612455090.

Baghi, H., & Ferrara, S. F. (1989). A Comparison of IRT, Delta Plot, and Mantel-Haenszel Techniques for Detecting Differential Item Functioning Across Subpopulations in the Maryland Test of Citizenship Skills.

Baker F. B., Kim S. H. (2004). *Item Response Theory: Parameter Estimation Techniques, 2nd Edn*. New York, NY: Marcel Dekker.

Bergstrom, B. A., Stahl, J. A., & Netzky, B. A. (2001). Factors that influence item parameter drift. *Paper presented at the annual meeting of the American Educational Research Association, Seattle.*

Bock, D. B., Muraki, E., & Pfeiffenberger, W. (1988). Item pool maintenance in the presence of item parameter drift. *Journal of Educational Measurement*, 25(4), 275-285.

DeMars, C. (2010). Item response theory. Oxford University Press.

Donoghue, J. R., & Isham, S. P. (1998). A comparison of procedures to detect item parameter drift .*Applied Psychological Measurement*, 22(1), 33-51.

Embretson, S. E. & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates.

Fraser, C. (1988). *NOHARM: An IBM PC computer program for fitting both unidimensional and multidimensional normal ogive models of latent trait theory*. computer software]. Armidale, Australia: The University of New England.

Goldstein, H. (1983). Measuring changes in educational attainment over time: Problems and possibilities. *Journal of Educational Measurement*, 369-377.

Guo, F., & Wang, L. (2003, April). Online calibration and scale stability of a CAT program. In *annual meeting of the National Council on Measurement in Education, Chicago: IL*.

Guo, R., Zheng, Y., & Chang, H. H. (2015). A stepwise test characteristic curve method to detect item parameter drift. Journal of Educational Measurement, 52(3), 280-300.

Hambleton, R.K.,Swaminathan,H., & Rogers,H.J. (1991).*Fundementals of Item response Theory*.Newbury,Park.CA:Sage.

Han, K. T., & Guo, F. (2011). Potential impact of item parameter drift due to practice and curriculum change on item calibration in computerized adaptive testing. *GMAC Research Reports*, *RR*-11, 2.

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, 30(2), 179-185.

Jodoin, M. G., & Gierl, M. J. (2001). Evaluating type I error and power rates using an effect size measure with the logistic regression procedure for DIF detection. *Applied measurement in education*, 14(4), 329-349.

Kim, S. H., Cohen, A. S., & Park, T. H. (1995). Detection of differential item functioning in multiple groups. *Journal of Educational Measurement*, 261-276.

Lee, H. (2018). Item Parameter Drift in a Time-Varying Predictor. *Applied Measurement in Education*, 31(1), 51-67.

Lee, Y. H., & von Davier, A. A. (2013). Monitoring scale scores over time via quality control charts, model-based approaches, and time series techniques.*Psychometrika*, *78*(3), 557-575.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems. Hillsdale*, NJ: Lawrence Erlbaum Associates Publishers.

Lord,F.M.,&novick,M.R. (1968).*Statistical theories of mental test scores*. Reading,MA : Addison -wesley.

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the national cancer institute*, 22(4), 719-748.

McCoy, K. M. (2009). *The impact of item parameter drift on examinee ability measures in a computer adaptive environment*. (Unpublished doctoral dissertation). University of Illinois at Chicago, Chicago, IL.

McDonald, R. P. (1967). Nonlinear factor analysis. *Psychometric Monographs* (No. 15).

McDonald, R. P. (1982). Linear versus nonlinear models in item response theory. *Applied Psychological Measurement*, 6, 379-396.

Mellenbergh,G.j. (1989). Item Bias and Item Response Theory. *International Journal of Educational Research*, 13,127-143.

Messick, D. M., & Mackie, D. M. (1989). Intergroup relations. *Annual review of psychology*, 40(1), 45-81.

Mislevy, R. J. (1982). Five steps toward controlling item parameter drift. *Paper presented at the annual meeting of the American Educational Research Association, New York.*

Nandakumar, R. (1991). Traditional dimensionality versus essesntial dimensionality .*Journal of Educational Measurement*, 28(2), 99-117.

Narayanan,P.y., & Swaminathan, H. (1996). Identification of items that show nonuniform DIF,*Applied Psychological Measurment*, 20:257-274.

Näsström, Gunilla. (2003). Differential item functioning for items in the Swedish National test in mathematics, course B. *Paper presented at the Pre-ICME Conference in Växjö*.

O'Neal, Marcia R. (1991). A Comparison of Method for Detecting Item Bias. *Paper presented at the annual meeting of the Mid -South Educational Research Association* 20th, Lexington, *KY* ,November 12-15.

Oshima, T. C., & Morris, S. B. (2008). Raju's differential functioning of items and tests (DFIT). *Educational Measurement: Issues and Practice*, *27*(3), 43-50.

Park,s.p.,lee.Y.S.,&Xing.K. (2016).Investigating theImpact of Item Parameter drift for tem Response Theory Models with Mixture Distributions. *Distributions.Front.Psychol*.7:255.

Raju, N. S. (1988). The area between two item characteristic curves. *Psychometrika*, 53-۵۵۵ , ۲۲۲.

Raju, N. S. (1990). Determining the significance of estimated signed and unsigned areas between two item response functions. *Applied Psychological Measurement*, *14*(2), 197-207.

Reckase,M.D. (2010). Designing Item Pools to Optimize the Functioning of Computerized Adaptive Test. *Psychological Test and Assessmeny Modeling* 52,127-141.

Reeve, B. B., Hays, R. D., Bjorner, J. B., Cook, K. F., Crane, P. K., Teresi, J. A., ... & Liu, H. (2007). Psychometric evaluation and calibration of health-related quality of life item banks: plans for the Patient-Reported Outcomes Measurement Information System (PROMIS). *Medical care*, 45(5), S22-S31.

Risk, N. M. (2015). The Impact of Item Parameter Drift in Computer Adaptive Testing (CAT) (Doctoral dissertation, University of Illinois at Chicago).

Rupp, A. A., & Zumbo, B. D. (2006). Understanding parameter invariance in unidimensional IRT models. *Educational and Psychological Measurement*, 66(1), 63-84.

Shultz, K. S., Whitney, D. J., & Zickar, M. J. (2013). *Measurement theory in action: Case studies and exercises*. Routledge.

Swaminathan, H.& Rogers, H.J. (1990). Detection differential item functioning using Logistic regression procedures. *Journal of Educational Measurment*, 27,361-370.

Van der Linden, W. J., & Glas, C. A. (2010). *Elements of adaptive testing*. New York, NY: Springer.

Wainer, H., Dorans, N. J., Green, B. F., Steinberg, L., Flaugher, R., Mislevy, R. J. Thissen, D. (2010). *Computerized adaptive testing: A primer*. *Hillsdale*, NJ: Lawrence Erlbaum AssociatesPublishers.

Wei,X.E. (2013). Impacts of Item Parameter Drift on Person Ability Estimation in Multistage Testing. *Technical Report.*

Wells, C. S., Subkoviak, M. J., & Serlin, R. C. (2002). The effect of item parameter drift on examinee ability estimates. *Applied Psychological Measurement*, 26(1), 77-87.

Wells,C.S.,Hambleton,R.K.,Kirkpatrick,R.,andMeng,Y. (2014).Anexamination of two procedures for identifying consequential item parameter drift. *Appl.Meas.Educ.* 27,214–231.

Wu, A. D., Li, Z., Ng, S. L., & Zumbo, B. D. (2006). Investigating and comparing the item parameter drift in the mathematics anchor/trend items in TIMSS between Singapore and the United States. *In 32nd Annual Conference in International Association for Educational Assessment (Singapore:).*

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.*