فصلنامه علمی پژوهشی تدریس پژوهی
سال هفتم– شمارهٔ اول – بهار ۱۳۹۸
صفحات: ۱۴۹-۱۲۹
Journal of Research in Teaching
Vol 7, No 1, Spring 2019

منابع کارکرد افتراقی سؤال و کاربرد آن در آموزش

مولود علی میرزایی[۱]، علی مقدم زاده[۲*]، اصغر مینایی[۳]، بلال ایزانلو[۴]، کیوان صالحی[۵]

۱. دانشجوی دکتری سنجش آموزش، دانشگاه تهران، تهران، ایران؛ ۲. استادیار گروه روش‌ها و برنامه‌های آموزشی و درسی، دانشگاه تهران، تهران، ایران؛ ۳. استادیار گروه سنجش و اندازه‌گیری؛دانشگاه علامه طباطبائی تهران، ۴. استادیار گروه آموزشی برنامه‌ریزی درسی، دانشگاه خوارزمی، تهران، ایران؛ ۵. استادیار گروه روش‌ها و برنامه‌های آموزشی و درسی، دانشگاه تهران، تهران، ایران

دریافت مقاله: ۱۳۹۷/۰۳/۱۲
پذیرش مقاله: ۱۳۹۸/۰۱/۱۴

چکیده

هدف: ویژگی‌های روان‌سنجی مثل تغییرناپذیری اندازه‌گیری، عدم وجود کارکرد افتراقی سؤال، یا فقدان سوگیری باید برقرار باشد تا نمرات یک آزمون برای گروه‌های مختلف آزمودنی مقایسه‌پذیر شوند. تحلیل منابع کارکرد افتراقی سؤال به محققان اجازه می‌دهد، فرضیه‌هایی در ارتباط با منابع اصلی و بالقوه سوگیری و واریانس سازه مزاحم را فرمول‌بندی کنند. ازآنجایی‌که معمولاً فرضیه‌های مربوط به منابع DIF بر مبنای نظریه یا تحقیقات پیشین ساخته می‌شود، مرور نظامند مقاله‌ها مربوط به بررسی علل DIF ضروری به نظر می‌رسد. یافته‌ها نشان می‌دهد که اهمیت اصلی شناسایی منابع DIF در ساخت و تفسیر نتایج آزمون‌ها است که برای مقایسهٔ بین گروه‌ها به کار برده می‌شود. این پژوهش یک مرور نظام‌مند است که به جستجوی کلمات کلیدی در پایگاه‌های معتبر ازجمله Taylor & Francis، WILEY، Springer، SAGE می‌پردازد. از میان ۴۲ مقاله یافته شده براساس ملاک‌های ورود تعداد ۱۹ مقاله مرتبط با موضوع شناسایی شد. در استفاده از هر مقاله سعی شد نمرات مورد مقایسه در DIF، منابع شناسایی‌شده برای DIF، چگونگی شناسایی منبع DIF، راهبرد مورد استفاده برای تأثیر منبع DIF شناسایی‌شده استخراج شود.

یافته‌ها: یافته‌ها نشان می‌دهد که اهمیت اصلی شناسایی منابع DIF در ساخت و تفسیر نتایج آزمون‌ها است که برای مقایسه‌ی بین گروه‌ها به کار برده می‌شود. DIF هنگامی به وجود می‌آید که گروه‌های خاصی را مقایسه می‌نماییم. سیاست‌گذاران باید هنگام تصمیم‌گیری در مورد برنامه درسی، منابع یا آموزش بر مبنای مقایسه نمرات آزمون، مراقب روایی مقایسه‌پذیری نمرات باشند.

کلید واژه‌ها: مقایسه‌پذیری نمرات آزمون‌ها، منابع کارکرد افتراقی سؤال، ساخت آزمون، آزمون‌های آموزشی

Sources of the differential item functioning and its application in education

Molood Alimirzaie[1], Ali Moghadam zadeh[2*], Asghar Minaei[3], Balal Ezanloo[4], Keyvan Salehi[5]

1 Ph.D. Student of Educational Measurement and Evaluation, University of Tehran, Tehran, Iran; 2 Assistant Professor of Department of Curriculum Development & Instruction Methods, University of Tehran, Iran; 3. Assistant Professor of Department of Assessment & Measurement, Allameh Tabataba'i University, Tehran, Iran; 4. Assistant Professor of Department of Curriculum Planning, University of kharazmi, Tehran. Iran; 5 Assistant Professor of Department of Curriculum Development & Instruction Methods, University of Tehran, Iran;

Received Date: 2018/06/02
Accepted Date: 2019/04/03

Abstract

Purpose: Psychometric properties such as measurement inaccuracies, lack of differential action of the question, or lack of bias must be maintained so that the grades of a test can be compared to the different groups of the subject. The analysis of source of differential item functioning allows researchers to formulate hypotheses related to the main and potential sources of bias and the variance of the intruder structure. Since the hypotheses of DIF sources are usually built on the basis of previous theory or research and are discussed in the next step, a systematic review of the papers related to the study of the causes of DIF seems to be necessary. The findings show that the main importance of identifying DIF resources in constructing and interpreting test results is used for comparison between groups. This research is a systematic review that searches for keywords on valid databases such as Taylor & Francis, WILEY, Springer, SAGE. Of the 42 articles found, 19 related articles were identified based on entry criteria. In using each article, we tried to extract the scores in DIF, sources identified for DIF, how to identify the DIF source, the strategy used to impact the identified DIF source. **Findings**: The findings show that the main importance of identifying sources of DIF in constructing and interpreting test results is used for comparison between groups. DIF happens when we compare certain groups. Policy makers should be careful about the comparability of scores when deciding on curriculum, resources or training based on comparing test scores.

Key words: Comparability of test scores, Source of differential item functioning, Test construction, Educational tests

* **نویسنده مسئول:**
Email: amoghadamzadeh@ut.ac.ir

مقدمه و بیان مسئله

استفاده از آزمون‌ها به‌عنوان وسیله تصمیم‌گیری‌های آموزشی، تاریخچه‌ای طولانی داشته و مدت زمان زیادی است که آزمون‌ها وسیله‌ای برای انتخاب افراد شده است. سنجش به‌عنوان فعالیت مهم آموزشی نقش بسیار مهمی در عملکرد یادگیرندگان و آموزگاران دارد (Diamond, Persson, 2016). داده‌های حاصل از آزمون‌های سرنوشت‌ساز معمولاً برای اطلاع‌رسانی سیاست‌ها و اجرای برنامه‌های درسی و تصمیم‌گیری‌های آموزشی به کار برده می‌شود. فرض ضمنی استفاده از داده‌های آزمون این است که اندازه‌گیری‌ها در میان استان‌ها و مناطق آموزشی مقایسه‌پذیر باشند. این فرض به معنای مقایسه‌پذیری نمرات سؤال‌ها و سازه‌های اندازه‌گیری شده در سنجش‌هاست. مرکز اصلی تلاش‌ها و ایجاد تفاسیر دارای روایی بر مبنای نتایج سنجش، نمرات مقایسه‌پذیر است (Oliveri, von Davier, 2014).

در تحلیل مقایسه‌پذیری نمرات یک آزمون، دو جنبه مهم وجود دارد که شامل بررسی کارکرد افتراقی سؤال و شناسایی منابع آن است. یکی از چالش‌های DIF برای بررسی مقایسه‌پذیری سؤال‌ها آزمون، شناسایی علل بالقوه DIF است. در تحلیل مجموعه داده‌های پیچیده که شامل افراد از ملت‌ها، ایالت‌ها، اقوام و فرهنگ‌های مختلف هستند، تمرکز مطالعات در پیشینه تحقیق فراتر از شناسایی DIF است و به سمت توضیح منابع DIF حرکت کرده است (Albano, Rodriguez, 2013). شناسایی علل DIF با بازبینی محتوای سؤال‌های دارای DIF، بنیش عمیق‌تری نسبت به متغیرهای مرتبط با DIF فراهم می‌نماید و به‌این‌ترتیب اطلاعاتی به کارشناسان در خصوص جنبه‌های بالقوه‌ای از سؤال‌ها که باید در بررسی‌های مهم مورداستفاده قرار گیرد، ارائه می‌دهد.

هنگامی‌که DIF شناسایی شد، تحلیل اضافی برای بررسی منابع آن انجام می‌شود. منابع DIF موردبررسی قرار می‌گیرد تا عوامل سازه مزاحم[1] که به‌طور غیرمنتظره با سازه موردراندازه‌گیری آزمون همراه می‌شود را شناسایی نموده و تصمیم‌گیری‌های مربوط به حفظ یا حذف این سؤال‌ها را اعلام نماید. تحلیل برای بررسی منابع بالقوه DIF شامل بررسی سؤال‌ها توسط کارشناسان برای محتوا، پیچیدگی شناختی، بار فرهنگی، تفاوت‌های زبانی در نسخه‌های چندگانه زبانی آزمون و همچنین مصاحبه‌های شناختی است (Oliveri, Ercikan, Zumbo, 2013). اگرچه روش‌های شناسایی DIF برای بهبود کیفیت آزمون مفید است، اما پیشرفت اندکی در زمینه شناسایی علل و ماهیت موضوعاتی که منجر به وجود DIF در سؤال‌ها می‌شود، وجود دارد. یعنی هنگامی‌که سؤال‌ها در یک آزمون از نظر آماری دارای عملکرد افتراقی از یک گروه آزمودنی به گروه دیگر باشند، شناسایی دلایل عملکرد افتراقی سؤال‌ها، دشوار است. تحلیل برای بررسی منابع بالقوه DIF شامل بررسی سؤال‌ها توسط کارشناسان برای محتوا، پیچیدگی شناختی، بار فرهنگی، تفاوت‌های زبان در نسخه‌های چندگانه زبانی آزمون و همچنین مصاحبه‌های شناختی است. این تلاش‌ها برای شناسایی وجود سوگیری است، یعنی اینکه پاسخ‌های آزمودنی‌ها نه‌تنها توانایی در سازه موردنظر را نشان می‌دهد، بلکه به واریانسی خارج از سازه موردراندازه‌گیری (مثلاً واریانس سازه مزاحم) نیز اشاره دارد که روایی استنباط‌های مربوط به

1. construct-irrelevant

تفاوت‌های عملکرد گروهی را کاهش می‌دهد. تحلیل منابع DIF به محققان اجازه می‌دهد تا فرضیه‌های در ارتباط با منابع اصلی و بالقوه اریبی و واریانس سازه مزاحم را فرمول‌بندی کنند (Roth, Oliveri, Sandilands, Lyons-Thomas, Ercikan, 2013).

هنگامی‌که داده‌های سنجش در مقیاس بزرگ برای تصمیم‌گیری‌های سیاسی و آموزشی استفاده می‌شود، مهم است که توجه داشته باشید همه آزمودنی‌ها در هر کشور به‌طور همگن به سؤال‌ها پاسخ نداده‌اند. فرض همگونی درون‌گروهی باید به‌صورت تجربی به‌عنوان اولین گام در تحلیل مقایسه‌پذیری بین گروه‌ها (بر مبنای کشورها، جنسیت یا زبان) مورد ارزیابی قرار گیرد. زیرا هنگامی‌که داده‌ها ناهمگون هستند، ممکن است یافته‌ها فقط برای یک زیرگروه از آزمودنی‌ها نه تمام آن‌ها به‌کار برده شود. این یافته‌ها مفاهیمی برای سیاست‌گذاری در آموزش دارد زیرا سیاست‌ها اغلب برای کل یک گروه اعمال می‌شود (تفاوت‌های پیشرفت تحصیلی در دختران و پسران). اگر گروه‌های آزمودنی‌ها به‌طور مشابه به سؤال‌ها پاسخ ندهند، استنباط‌ها بر مبنای آزمون، ممکن است برای تمام آزمودنی‌ها در یک گروه قابل تعمیم نباشد. این امر منجر به نتیجه‌گیری‌هایی می‌شود که قابلیت تعمیم‌پذیری محدودی دارند و نتایجی که برای کاربرد در اصلاح آموزش برای زیرگروه‌هایی متفاوت از شرکت-کنندگان در سنجش، محدود است (Ercikan, 2008).

دغدغه شناسایی منابع DIF به تحقیق انگوف برمی‌گردد، او نوشت: سازندگان آزمون‌ها اغلب با نتایج DIF مواجه می‌شوند که نمی‌توانند آن را درک کنند و به نظر می‌رسد هیچ بررسی نمی‌تواند به توضیح اینکه چرا برخی سؤال‌ها کاملاً معقول دارای DIF هستند، کمک کند (Angoff,1993). به‌علاوه عدم شناسایی منابع DIF در استانداردها (انجمن تحقیقات آموزشی آمریکا [AERA]، انجمن روانشناسی آمریکا [APA] و شورای ملی اندازه‌گیری در آموزش‌وپرورش [NCME]،۱۹۹۵) نیز برجسته شده است. تحقیقات قبلی در مورد DIF و منابع آن نشان دادند که نسخه‌های چندزبانه یک سنجش را نمی‌توان مقایسه‌پذیر فرض نمود زیرا دارای نسبت بزرگی از DIF هستند. به‌علاوه در نسبت زیادی از سؤال‌ها دارای DIF، منابع DIF را نمی‌توان شناسایی نمود. برای مثال نسخه‌های انگلیسی و فرانسوی آزمون پیشرفت تحصیلی ملی کانادا را با استفاده از SIBTEST و رویکرد لین-هارنیش[1] (LH) بررسی شد و نشان داد که بیش از ۳۶ درصد سؤال‌ها برای آزمودنی‌های انگلیسی و فرانسوی زبان دارای کارکرد افتراقی است و منابع DIF برای ۳۰ تا ۴۰ درصد از سؤال‌ها مشخص شده یافته شد (Ercikan, Gierl, McCreith, Puhan, Koh,2004). فرآیندهای شناختی درگیر در هنگام آزمون و نوع خاص تفکر، به‌وسیله زبان و فرهنگ تحت تأثیر قرار می‌گیرد، زمینه‌های فرهنگی می‌تواند اندازه‌گیری دانش را در آزمون‌های استاندارد مختل کند، گروه‌های متمایز از لحاظ فرهنگی، الگوهای خاص تفکر و یادگیری دارند که ممکن است منابعی برای DIF باشند (Li,Cohen,Ibarra,2004, Pellegrino, Chudowsky, Glaser, 2001).

1. Linn-Harnisch

مرور تحقیقات نشان می‌دهد که تاکنون در ایران در رابطه با منابع کارکرد افتراقی سؤال، پژوهشی انجام نشده است. به‌دلیل اهمیت مقایسه گروهی نمرات در سنجش‌های در مقیاس بزرگ و لزوم هم‌ارزی سؤال‌ها برای مقایسه گروه‌های مختلف داوطلبان شرکت‌کننده در یک آزمون، شناسایی و تحلیل منابع DIF ضروری است. همچنین در ایران به دلیل وابستگی سیاست‌گذاران و تصمیم‌گیرندگان به آزمون‌های گوناگون مانند آزمون‌های ورودی آموزش عالی و آزمون‌های استخدامی، عدم‌توجه به منابع DIF می‌تواند منجر به استنباط‌هایی نادرست در مقایسه نمرات آزمون‌ها شود. لذا تهیه مقاله‌ای که به‌طور نظامند روند مقالات انجام گرفته در این زمینه را بررسی نماید، به شناسایی منابع DIF برای ساخت و آزمون فرضیه‌ها، در تحقیقات آینده کمک می‌نماید و همچنین با معرفی منابع بالقوه DIF به سازندگان آزمون، اطلاعاتی در مورد جنبه‌های از سؤال‌ها ارائه می‌دهد که باید در بررسی‌های مهم مورد استفاده قرار گیرد، بنابراین به ساخت سؤال‌هایی باانصاف بیشتر کمک خواهد کرد. اهداف ویژه مطالعه مروری حاضر عبارت است از: ۱- بررسی منابع DIF شناسایی‌شده در آزمون‌های آموزشی، ۲- تبیین روش‌های مورداستفاده برای شناسایی منابع DIF، ۳- معرفی راهبردهای بررسی تأثیر منابعی که برای DIF شناسایی‌شده است.

روش‌شناسی پژوهش

این پژوهش یک مرور نظام‌مند است که با رجوع به سایت‌های بین‌المللی انجام شد. به این منظور ابتدا Google Scholar برای جستجوی کلی کلید واژه‌های sources of differential item functioning و causes of differential item functioning مورداستفاده قرار گرفت. ملاک ابتدایی برای ورود به مطالعه ارتباط مقاله با شناسایی منابع DIF بود. از دیگر ملاک‌های ورود می‌توان به چاپ مقاله در پایگاه‌های معتبر از جمله Taylor& Francis، WILEY، Springer، SAGE اشاره نمود. جستجو، شامل مقالاتی در مجلات مربوط به سلامت، پزشکی، کیفیت زندگی و روان‌پزشکی بود که این موارد از بررسی در مطالعه خارج شدند.

از تعداد ۱۱ مقاله در پایگاه Taylor& Francis و ۸ مقاله در پایگاه Springer، ۹ مقاله در پایگاه SAGE و ۱۴ مقاله در پایگاه WILEY بر اساس ملاک‌های ورود تعداد ۱۹ مقاله مرتبط با موضوع شناسایی شد و مورد بررسی قرار گرفت. در استفاده از هر مقاله سعی شد نمرات مورد مقایسه در DIF، منابع شناسایی‌شده برای DIF، چگونگی شناسایی منبع DIF، راهبرد مورد استفاده برای تأثیر منبع DIF شناسایی‌شده استخراج شود و سپس اطلاعات با دقت دسته‌بندی و تحلیل شود. جزئیات مقالات بررسی شده در جدول۱ نشان داده شده است.

جدول (۱): جزئیات مقالات مربوط به منابع DIF

شماره	مقالات	منابع DIF	چگونگی شناسایی منبع DIF	راهبرد مورداستفاده برای بررسی تأثیر منابع DIF	نمرات مورد مقایسه	نتیجه
۱	Allalouf, Hambleton Sireci (1999)	تغییرات در دشواری کلمات یا عبارات، تغییرات در محتوا، تغییر قالب، تفاوت‌های مربوط به فرهنگ	استفاده از نظر ۵ مترجم و ۳ کارشناس زبان عبری		نمرات بخش کلامی آزمون روان‌سنجی ورودی دانشگاه در اسرائیل بین نسخه اصلی به زبان عبری و نسخه ترجمه شده به زبان روسی	نتایج این مطالعه منابع DIF یافته شده است.
۲	Bolt (2000)	قالب سؤال (چهارگزینه‌ای در مقابل بازپاسخ) و نوع سؤال (انتزاعی در مقابل حقیقی)	تحقیقات گذشته	طرح مطالعه DIF تصادفی و استفاده از SIBTEST برای آزمون اثرات علل DIF	نمرات بخش ریاضی از یک اجرای منحصربه‌فرد از آزمون سنجش مدرسه‌ای در گروه مردان وزنان	اثر کوچک اما معنادار و ثابت در میان سؤال‌ها، از قالب سؤال و اثر بزرگ و باثبات کمتری در مورد نوع سؤال. اثرات متقابل شامل اثرات سؤال و عوامل معنی‌دار نبود.
۳	Ercikan(2002)	انطباق نسخه‌های چندزبانه آزمون تیمز، برنامه درسی	تحقیقات گذشته	۱- مرور قضاوتی با مترجمان چندزبانه برای همه سؤال‌ها، ۲- روایی متقاطع DIF در گروه‌های چندگانه، ۳- بررسی توزیع DIF سؤال‌ها به‌وسیله موضوع	نمرات تیمز نسخه انگلیسی و فرانسوی در کشورهای کانادا، انگلیس، فرانسه و ایالات متحده در درس علوم و ریاضیات دانش‌آموزان ۱۳ ساله	۲۳ درصد سؤال‌های DIF ریاضی و ۱۳ درصد سؤال‌های DIF علوم در تفاوت در برنامه درسی و ۵۰ درصد سؤال‌ها به دلیل دو منبع ذکر شده نبود. ۲۷ درصد در ریاضی و ۳۷درصد سؤال‌ها در علوم دارای DIF به دلیل انطباق
۴	Cohen, Bolt (2005)	ابعاد ثانویه مزاحم، در گروه جبر، گروه هندسه و اندازه‌گیری و گروه به‌کارگیری نمادین	مدل ترکیبی IRT		نمره آزمون گمارش ریاضی (به دوره پیش محاسبات) برای دانشجویان سال اول بین زنان و مردان	آزمودنی‌ها در ابعاد مزاحم ثانویه در اندازه‌های پیشرفت تحصیلی‌شان متفاوت‌اند.
۵	Amery, Ercikan (2006)	کلاس‌های اضافه برای درس پس از مدرسه	تحقیقات گذشته، اسناد و اطلاعات موجود	مدل‌سازی DIF ترتیبی (رگرسیون لوجستیک)	نمرات ریاضی تیمز بین ایالات متحده و تایوان	کلاس‌های اضافه پس از مدرسه با کاهش در مقدار و تعداد سؤال‌های DIF مرتبط است

۶	Elosua, López-jaúregui,(2007)	نقض در ترجمه، تفاوت دستوری بین زبان‌ها، تفاوت‌های معنایی و ویژگی‌های فرهنگی	گروهی از متخصصان	تشکیل گروه دومی از متخصصان مستقل از گروه اول و بررسی هماهنگی بین نظرات دو گروه	نمرات بخش کلامی مخزن سؤال‌ها استعداد عمومی در دانش‌آموزان اسپانیایی و باسک	به‌جز ویژگی‌های فرهنگی سایر عوامل بر DIF تأثیر می‌گذارند
۷	Ardeshir, Antony (2007)	سن داوطلبان	بررسی تحقیقات گذشته و اینکه تا قبل از این مطالعه در موردبررسی سن به‌عنوان منبع DIF تحقیقی انجام نگرفته است	تحلیل آماری و تحلیل محتوا	نمرات آزمون شنیداری CAE کمبریج ۲۰۰۲ در سه گروه سنی کمتر از ۱۷ سال، ۱۸ تا ۲۲ سال و بالای ۲۳ سال	تأثیر منبع DIF معنادار نیست.
۸	Banks (2009)	کارکرد افتراقی گزینه‌های انحرافی	تحقیقات گذشته	استفاده از آزمون SIBTEST، برازش مدل لگاریتم خطی، محاسبه نسبت بخت‌ها	نمرات آزمون UE و MP در میان زنان و مردان و سفید و سیاه‌پوستان و اسپانیایی‌ها و سفیدپوستان	آزمودنی‌ها به‌طور افتراقی به گزینه انحرافی "بدون اشتباه" کشیده شدند و هیچ‌کدام از سؤال‌ها MP از هر سه مرحله گذر نکردند.
۹	Ercikan, Arim, Law, Domene, Gagnon & Lacroix (2010)	کلمات کلیدی سؤال که ممکن است کمک یا مانع برای حل سؤال باشد، وضوح و مشخصه زبان، تفاوت‌های زبانی غیرمنتظره (اشتباه خواندن یک کلمه توسط دانش‌آموزان فرانسوی‌زبان)	مرور متخصصان	پروتکل تفکر با صدای بلند	نمرات علوم و ریاضی آزمون SAIP ۲۰۰۳ یک آزمون پیشرفت تحصیلی کانادایی بین دانش‌آموزان پایه ۷ و ۸ انگلیسی و فرانسوی‌زبان	برای سایر منابع به‌جز قالب‌بندی و ارائه بصری، وضوح و هدایت زبانی، شواهد، پروتکل تفکر با صدای بلند شواهد تأییدی فراهم نمود.
۱۰	Aryadoust, Goh & Kim(2011)	محتوای سؤال‌ها ازجمله، عبارات منفی، اصطلاحات، استعاره‌ها و سؤال‌ها با طولی که ممکن است	تحقیقات گذشته	تحلیل محتوای تعقیبی[1] سؤال	مقایسه نمرات آزمون شنیداری موسسه زبان انگلیسی میشگان در گروه مردان وزنان	

1 Post hoc

		در حافظه آزمودنی‌ها با توانایی پایین باقی نماند.				
۱۱	Suh, Bolt (2011)	کارکرد افتراقی گزینه‌های انحرافی	تحقیقات گذشته	استفاده از مدل دو پارامتری آشیانه‌ای چندگروهی و سه مدل سلسله مراتبی آشیانه‌ای با محدودیت‌های متفاوت و مطالعه شبیه‌سازی برای سنجش عملکرد آماره نسبت درستنمایی برای شناسایی DDF تحت مدل دو پارامتری آشیانه‌ای	مقایسه نمرات آزمون تعیین سطح ریاضی در سطح دانشگاه در میان مردان وزنان	
۱۲	(Oliveri, Ercikan, Zumbo(2013)	عوامل مرتبط با دانش‌آموز (کشور، سن و جنس) و معلم (آموزش مهارت‌های رمزگشایی، گروه ه بندی ها با توانایی مشابه، آموزش گروه‌های کوچک)	محقق به دنبال یافتن متغیرهایی است که ممکن از بیشتر از متغیرهای مانیفست با نمرات پیشرفت تحصیلی ارتباط داشته باشند	مدل‌سازی طبقه مکنون با استفاده از سه روش آماری: تابع تشخیص توصیفی، رگرسیون لوجستیک چندجمله‌ای و تحلیل رگرسیون لوجستیک چندجمله ای چند سطحی	داده‌های آزمون خواندن پیلز ۲۰۰۶ دانش‌آموزان در دو کشور چینی زبان (تایپه و هنگ‌کنگ) و دو کشور عربی زبان (کویت و قطر)	توافق در سه روش آماری برای سه متغیر یافته شد: میانگین مقادیر قابل‌قبول، مقایسه بین هنگ‌کنگ و قطر و سن. معنی‌داری دیگر متغیرها، مقایسه دو کشور دیگر، جنسیت و آموزش مهارت‌های رمزگشایی و گروه‌های کوچک، به روش مورداستفاده برای بررسی منابع DIF طبقه مکنون وابسته است.
۱۳	Roth, Oliveri, Sandilands, Lyons-Thomas& Ercikan (2013)	تفاوت در طول نسبی دو نسخه‌ی زبانی، تفاوت در مسائل زبان‌شناختی، تفاوت در ساختار منطقی محتوا یا قالب سؤال‌ها،	تحلیل پروتکل تفکر با صدای بلند در مورد مرورگران حرفه‌ای با استفاده از روش‌های استاندارد تحلیل محتوا		نسخه انگلیسی و فرانسوی ۴۰ سؤال که در تحقیق قبلی۲۰ تا از آن‌ها دارای DIF و ۲۰ تا فاقد DIF شناسایی شدند در درس علوم آزمون SAIP آزمون پیشرفت تحصیلی ملی در کانادا	نتیجه این مقاله منابع یافته شده برای DIF است.

		تفاوت در محتوای شناختی مفهومی، تفاوت در مسائل تنوع[1]				
۱۴	Sandilands, Oliveri, Zumbo, Ercikan (2013)	ترجمه و انطباق نسخه‌های چندزبانه آزمون و بار شناختی سؤال	تحقیقات گذشته	استفاده از آزمون Poly-SIBTEST تحلیل کارکرد افتراقی دسته	نمرات آزمون خواندن پیرلز ۲۰۰۱ بین دانش‌آموزان انگلیسی و اسپانیایی‌زبان در دانش‌آموزان نه ساله در پایه چهارم	برخلاف بارهای شناختی سؤال‌ها، اثر ترجمه و انطباق تأیید نمی‌شود
۱۵	Benítez, Padilla (2014)	تفسیر سؤال‌ها به‌طور افتراقی توسط شرکت‌کنندگان در گروه‌های مختلف	تفسیرهای شرکت‌کنندگان به‌وسیله‌ی تم‌ها و زیر تم‌ها که از روایات آن‌ها به‌دست‌آمده، مقایسه می‌شوند.	ادغام نتایج به‌دست‌آمده از تحلیل کمی DIF و مصاحبه شناختی	مقایسه نمرات آزمون پیزا ۲۰۰۶ بین دانش‌آموزان ۱۵ و ۱۶ ساله از ایالات متحده و اسپانیا	سؤال‌های با DIF فراگیر حاوی عناصر مفهومی متفاوت است که دارای معانی مختلفی هستند، سؤال‌ها با DIF غیرفراگیر، به دلیل تجارب مختلف مرتبط با زمینه یا فرهنگ، متفاوت هستند. مشکلات ناشی از انطباق در اصطلاحاتی است که در دو نسخه معادل نیست.
۱۶	Chun, Stark, Kim, Chernyshenko (2016)	متغیرهای پیش‌زمینه مثل جنسیت و گروه‌های قومی و اثرات متقابل آن‌ها	تحقیقات کمی وجود دارد که به بررسی دقت MIMIC برای تشخیص DIF به دلیل ترکیب متغیرهای پیش‌زمینه و اثرات متقابل آن‌ها پرداخته باشد.	مطالعه شبیه‌سازی به‌منظور بررسی اثربخشی سه روش پیاده‌سازی MIMIC، نقطه شروع ثابت، نقطه شروع آزاد و روش جدید نقطه شروع آزاد متوالی(MIMIC) برای شناسایی کارکرد افتراقی یکنواخت و غیریکنواخت سؤال برای چندگروه	گروه‌های جنسیتی مردان وزنان و گروه‌های قومیِ اقلیت و اکثریت. گروه مردان با قومیت اکثریت به‌عنوان گروه مرجع در نظر گرفته می‌شود	روش نقطه شروع به‌طور متوالی آزاد خطای نوع اول و توان آزمون عالی فراهم می‌نماید و نتایج آن مشابه با روش نقطه شروع آزاد ایده آل است که از طرح لنگر فاقد DIF استفاده می‌کند و خیلی بهتر از روش نقطه شروع ثابت است که از همه سؤال‌ها به‌جای سؤال‌های موردمطالعه به‌عنوان لنگر استفاده می‌کند.

1 diversity

۱۷	Finch, Finch ,French (2016)	زبان مادری آزمودنی	تحقیقات گذشته	استفاده از جزءبندی بازگشتی مبتنی برمدل (MBRP) برای بررسی DIF یکنواخت	نمرات ۱۳ سؤال از آزمون پیلز ۲۰۱۱ برای ۱۱ کشور اروپایی با کنترل منابع آموزشی در دسترس آزمودنی	DIF بر مبنای زبان مادری در چندین سؤال پیلز وجود دارد اما الگوی DIF در همه کشورها یکسان نیست.
۱۸	Huang, Wilson & Wang (2016)	ترجمه آزمون، پوشش برنامه درسی متفاوت، تفاوت‌های فرهنگی	تحقیقات گذشته	استفاده از یک شخص دو زبانه برای بررسی اثر ترجمه، استفاده از یک پانل با ۱۰ نفر متخصص بررسی محتوا و ۱۵دانش‌آموز برای بررسی اثر پوشش برنامه درسی	نمرات آزمون علوم پیزا ۲۰۰۶ در میان گروه‌های : ۱- ایالات متحده و کانادا انگلیسی‌زبان، ۲- سرزمین چین و چینی‌زبان‌های هنک کنگ، ۳- ایالات متحده و چین	
۱۹	Svetina, Dai & Wang (2017)	چیرگی در ویژگی‌ها و مهارت‌های زیربنایی برای عملکرد در سؤال‌ها	مدل‌های شناختی تشخیصی ماتریس Q که شامل ویژگی‌های زیر بنایی عملکرد در سؤال‌ها است، استخراج ویژگی‌های زیر بنایی از تحقیقات گذشته و ارزیابی توسط دومتخصص، محاسبه احتمال چیرگی در ویژگی‌ها با استفاده از مدل RUM	تفاوت معنادار احتمال چیرگی در ویژگی‌ها در گروه‌های مورد مقایسه با آزمون مانوا انجام شد.	نمرات سنجش ریاضی در NAEP ۲۰۰۷ برای دانش‌آموزان بدون مساعدت (گروه مرجع)، مساعدت با زمان اضافی، مساعدت با خواندن سؤال‌ها با صدای بلند و مساعدت در گروه‌های کوچک	به‌طور متوسط تفاوت در میزان چیرگی بر یک ویژگی بین دریافت کنند و عدم دریافت‌کنندگان مساعدت در حدود ۳۳ تا ۴۴ درصد است

یافته‌های پژوهش

به دلیل اهمیت مقایسه‌های گروهی نمرات در سنجش‌های در مقیاس بزرگ و لزوم هم‌ارزی سؤال‌ها برای مقایسه گروه‌های مختلف شرکت‌کننده در یک آزمون، در طی سال‌ها با استفاده از روش‌های گوناگون به شناسایی منابع وجود DIF پرداخته شده است. ۱۹ مقاله موردبررسی در این مطالعه در امتداد سال‌های ۱۹۹۹ تا ۲۰۱۷ انجام گرفته است. در بیشتر مقالات (۱۲مقاله) به بررسی DIF در نسخه‌های زبانی مختلف یک آزمون پرداخته شده، گروه‌های مورد مقایسه دیگر برای بررسی DIF، گروه‌های قومی، جنسیتی، گروه‌های سنی و گروه‌های دریافت‌کننده مساعدت آزمون و گروه‌های بدون دریافت مساعدت آزمون است و در یک مقاله از سؤال‌هایی که در پژوهش قبلی دارای DIF شناسایی‌شده‌اند استفاده شده است.

به‌طورکلی منابع یافته شده برای DIF در ۷ دسته قرار دارند. دسته‌بندی منابع و تعداد مقالاتی که این منابع را به‌عنوان علت DIF موردبررسی قرار داده‌اند، در جدول ۲ ارائه شده است. به دلیل اینکه معمولاً مقالات چندین عامل را به‌عنوان منبع DIF شناسایی نموده‌اند، مجموع ستون تعداد مقالات از ۱۹ بیشتر شده است.

جدول (۲): دسته‌بندی منابع DIF در مقالات مرتبط با منابع و علل DIF

منبع DIF	تعداد مقالات
ویژگی‌های سؤال	۹
عوامل مرتبط با دانش‌آموز(سن داوطلب، جنس، کشور، قومیت، زبان مادری، چیرگی در مهارت‌های زیربنایی برای عملکرد در آزمون)	۵
مسائل مربوط به انطباق و ترجمه نسخه‌های چندزبانه	۴
ویژگی‌های فرهنگی	۳
برنامه درسی	۲
عوامل مرتبط با آموزش معلم	۱
کلاس‌های اضافه پس از مدرسه	۱

منابع DIF مربوط به ویژگی‌های سؤال عبارت است از بارشناختی مفهومی سؤال، تفاوت در ساختار منطقی محتوا و قالب سؤال‌ها (مثل سؤال‌های چندگزینه‌ای یا باز پاسخ)، نوع سؤال (انتزاعی یا حقیقی)، تفاوت در طول نسبی دو نسخه، تغییر در دشواری کلمات یا عبارات، کلمات کلیدی که کمک یا مانعی برای حل سؤال هستند، کارکرد افتراقی گزینه‌های انحرافی سؤال، محتوای سؤال‌ها ازجمله، عبارات منفی، اصطلاحات، استعاره‌ها و سؤال‌ها با طولی که ممکن است در حافظه آزمودنی‌ها با توانایی پایین باقی نماند، است. همچنین منابع DIF شناسایی‌شده در مسائل مربوط به ترجمه و انطباق آزمون شامل نقص در ترجمه، تفاوت دستوری بین زبان‌ها، تفاوت‌های معنایی و مسائل زبان‌شناختی است.

در مورد روش‌های شناسایی منابع DIF، تعداد ۱۴ مقاله از طریق تحقیقات گذشته منبع DIF را شناسایی نموده‌اند، که از آن جمله می‌توان استفاده از نظر ۵ مترجم و ۳ کارشناس زبان (Allalouf, Hambleton, Sireci,1999)، از نظر گروهی از متخصصان (Elosua, López-jaúregui,2007, Ercikan and et al,2010)، از تحلیل محتوای پروتکل تفکر با صدای بلند در مورد مرورگران حرفه‌ای با استفاده از روش‌های استاندارد تحلیل محتوا (Roth and et al,2013) و از تفسیر شرکت‌کنندگان در آزمون برای شناسایی منابع DIF (Benítez, Padilla,2014) استفاده شده است. همچنین از روش مدل‌های ترکیبی[1] IRT برای شناسایی ابعاد ثانویه مزاحم استفاده شده است، چندبعدی بودن آزمون یکی از علت‌های اصلی DIF است، یافته‌های این تحقیق سه طبقه مکنون را نشان می‌دهد که در توانایی اندازه‌گیری در آزمون جبر و درک مطلب، متفاوت هستند (Cohen, Bolt, 2005).

پس‌ازاینکه منابع DIF شناسایی شد، از راهبردهای آماری و یا غیر آماری برای بررسی تأثیر منابع شناسایی‌شده، بر DIF استفاده شده است. در ۱۰ مقاله از راهبردهای آماری، در ۵ مقاله راهبردهای غیر آماری، در یک مقاله، هم از تحلیل‌های آماری و هم از تحلیل محتوا استفاده شده است و طرح پژوهش یک مقاله هم طرح ترکیبی، شامل ترکیب روش‌های کیفی و کمی است. در دو مقاله هم راهبردی برای بررسی تأثیر منابع DIF معرفی نشده و هدف آن‌ها تنها ارائه منابع احتمالی برای DIF بوده است. در تمامی مقالاتی که از راهبردهای بررسی تأثیرگذاری DIF استفاده نموده‌اند، به‌جز گروه-های سنی، ویژگی‌های فرهنگی (Elosua and et al., 2007)، قالب‌بندی و ارائه بصری، وضوح و هدایت زبانی (Ercikan and et al., 2010) در تأثیر منابع DIF به تأیید رسید.

راهبردهای آماری برای بررسی تأثیر منابع DIF

در این بخش راهبردهای به کارگرفته شده در مطالعات به ترتیب زمانی قرار داده شده‌اند. برای بررسی تأثیر قالب سؤال (چهارگزینه‌ای در مقابل بازپاسخ) و نوع سؤال (انتزاعی در مقابل حقیقی) در DIF از طرح مطالعه DIF تصادفی و استفاده از SIBTEST برای آزمون اثرات علل DIF استفاده شده است. نتایج، اثر کوچک اما معنادار و ثابت در میان سؤال‌ها، از قالب سؤال و اثر بزرگ و باثبات کمتری در نوع سؤال را نشان داد و اثرات متقابل شامل اثرات سؤال و عوامل معنی‌دار نبود (Bolt, 200). یکی از راه‌هایی بررسی ویژگی‌های سؤال مضنون به DIF از طریق مطالعه DIF تصادفی است. در ساده‌ترین شکل، چنین مطالعه‌ای شامل اجرای دو نسخه از سؤال‌های یکسان در دو نمونه تصادفی از آزمودنی-هاست، در یک نسخه عوامل مضنون به DIF حاضر هستند (یعنی نسخهٔ آزمایشی) و در نسخه دیگر عوامل مضنون وجود ندارند (یعنی نسخه کنترل) (Schmitt, Holland, Dorans,1993). سهم عامل موردمطالعه در DIF با مقایسه مقدار DIF مشاهده شده در دو نسخه ارزیابی می‌شود. مطالعه DIF تصادفی ممکن است بهترین کاربرد برای مطالعه عواملی باشد که ابعاد اصلی نیستند، یا اثرشان بعید

1. Mixture

است در بیش از یک سؤال در آزمون مشاهده شود، یا تأثیر نهایی آن‌ها در عملکرد سؤال نسبتاً کوچک است. نکته مهم در مطالعات DIF تصادفی، بررسی ثبات اثر یک عامل تصادفی در میان سؤال‌ها است. یکی از راهبرد مورداستفاده برای بررسی تأثیر کلاس‌های اضافه پس از مدرسه در نمرات ریاضی تیمز بین ایالات متحده و تایوان، استفاده از مدل‌سازی DIF ترتیبی (رگرسیون لوجستیک) است. نتایج این مطالعه نشان می‌دهد که کلاس‌های اضافه پس از مدرسه با کاهش در مقدار و تعداد سؤال‌های DIF مرتبط است. معمولاً در کشورهای آسیای شرقی آزمون ورودی متمرکز ملی برای ورود به دبیرستان یا دانشگاه وجود دارد، برای اطمینان از برد رقابتی در این آزمون‌های ورودی، دانش‌آموزان برای تکمیل آموزش رسمی از آموزش خصوصی کمک می‌گیرند، به همین دلیل کلاس‌های اضافه می‌تواند عاملی برای DIF بین کشورها باشد (Amery, Ercikan, 2006).

دیگر راهبرد مورداستفاده برای بررسی تأثیر گزینه‌های انحرافی بر DIF در بانکز (۲۰۰۹) به این صورت است که ابتدا آزمون SIBTEST انجام می‌شود تا تعیین نماید آیا هر گروه مقایسه شده در احتمال پاسخ صحیح به سؤال‌های آزمون متفاوت است. دوما روش برازش مدل لگاریتم خطی مورداستفاده قرار گرفت تا تعیین کند آیا سؤال‌هایی که DIF متوسط یا زیاد دارند ارتباط معناداری با نمره کل، عضویت گروهی و گزینه‌های انحرافی نشان می‌دهند. سوم نسبت‌بخت‌ها محاسبه می‌شود برای تعیین اینکه آیا گروهی که سؤال‌های DIF بر علیه آن است دارای شانس بیشتری برای انتخاب گزینه پیچیده انحرافی نسبت به سایر گزینه‌های انحرافی در مقایسه با دیگر گروه‌ها با توانایی مشابه، است. سؤال‌هایی که از هر سه مرحله عبور کنند، به‌عنوان سؤال‌هایی که نتایج DIF آن مربوط به (کارکرد افتراقی گزینه انحرافی) DDF است شناسایی می‌شوند. نتایج نشان می‌دهد چهار سؤال دارای DIF متوسط بر علیه زنان در آزمون UE بود و یک سؤال DIF متوسط برعلیه سفیدپوستان داشت و این آزمودنی‌ها به‌طور افتراقی به سمت گزینه انحرافی "بدون اشتباه" کشیده شدند. هیچ‌کدام از سؤال‌های آزمون MP از هر سه مرحله گذر نکردند (Banks, 2009).

یک روش رایج وابسته به مدل برای تشخیص DIF در IRT، آزمون نسبت درستنمایی است که توابع درستنمایی برای ارزیابی تفاوت پارامترها در گروه‌ها، مقایسه می‌شوند (Thissen, Steinberg,Gerrard,1986; Thissen, Steinberg, Wainer,1988,1993). برای مطالعه گزینه‌های انحرافی به‌عنوان عاملی برای DIF در سؤال‌های چندگزینه‌ای تحت مدل دو پارامتری آشیانه‌ای چندگروهی، سه مدلِ سلسله مراتبی آشیانه‌ای با محدودیت‌های متفاوت در مطالعه دیگر نظر گرفته شد: ۱- یک مدل فشرده که در آن پارامترهای سؤال برای یک سؤال در تمامی گروه‌ها برابر در نظر گرفته شده، ۲- یک مدل افزوده شده که در آن تنها پارامترهای گزینه انحرافی سؤال محدود شدند که در تمام گروه‌ها مساوی باشند، ۳- مدل افزوده دوم که در آن هیچ‌کدام از پارامترهای سؤال (پارامترهای گزینه انحرافی و گزینه درست) محدود نشدند که مساوی باشند. با محاسبه آماره G^2 برای مقایسه مدل فشرده و افزوده اول (آزمون۱) می‌توان آزمود که آیا DIF وجود دارد. با محاسبه G^2

برای مقایسه مدل‌های افزوده اول و دوم (آزمون ۲) می‌تواند ارزیابی کرد که آیا DIF مشاهده شده از رد آزمون۱ به‌دلیل حضور DDF اتفاق افتاده است. این پژوهش، مطالعه‌ای شبیه‌سازی به‌منظور سنجش عملکرد آماره نسبت درستنمایی برای شناسایی DDF تحت مدل دو پارامتری آشیانه‌ای است (Suh, Bolt, 2011).

راهبردی برای اثر انطباق نسخه‌های زبانی و بارهای شناختی سؤال بر DIF، تحلیل کارکرد افتراقی دسته (DBF) با استفاده از SIBTEST است. سؤال‌هایی که تصور می‌شود منابع بالقوه DIF هستند توسط متخصصان شناسایی می‌شوند. سپس این سؤال‌ها به‌صورت دسته در می‌آیند و تحلیل DBF در مورد آن‌ها انجام می‌شود. در مطالعه‌ای که از این راهبرد استفاده نمود DBF معناداری یافته نشد بنابراین فرضیه اثر ترجمه و انطباق به‌عنوان منبع DIF در این مجموعه داده مورد تأیید قرار نگرفت. این مورد می‌تواند مثالی از سؤال‌های DIF باشد که به‌طور هماهنگ عمل می‌کنند تا در سطح دسته-بندی شده لغو شوند، پدیده‌ای که لغو DIF نامیده می‌شود. اثر بارهای شناختی سؤال بر DIF در این مطالعه تأیید شد (Sandilands and et al., 2013).

راهبرد مورداستفاده برای بررسی تأثیر ویژگی‌های معلم و دانش‌آموز در DIF به این صورت است که تحلیل ابتدا با تعیین ابعاد ساختار داده‌های آزمون و بررسی تعداد طبقه‌های مکنون انجام می‌گیرد. سپس DIF در میان طبقه‌های مکنون بررسی می‌شود. در انتها منابع DIF طبقه‌های مکنون با استفاده از دو روش بررسی می‌شود. ابتدا به‌طور نظام‌مند سؤال‌های DIF را بر مبنای جنبه‌های اصلی گروه‌بندی نموده، دوما از مدل پیش‌بینی برای بررسی عوامل مرتبط با دانش‌آموز (کشو، سن و جنس) و معلم (آموزش مهارت‌های رمزگشایی، گروه‌بندی‌ها با توانایی مشابه، آموزش گروه‌های کوچک) که می‌تواند به‌طور بالقوه زمینه‌ساز DIF باشد، استفاده می‌شود. مدل‌سازی طبقه مکنون با استفاده از سه روش آماری: تابع تشخیص توصیفی، رگرسیون لوجستیک چندجمله‌ای و تحلیل رگرسیون لوجستیک چندجمله‌ای چند سطحی انجام می‌گیرد. توافق در سه روش آماری برای سه متغیر یافته شد: میانگین مقادیر قابل‌قبول، مقایسه بین هنگ‌کنگ و قطر، و سن. معنی‌داری دیگر متغیرها، مقایسه دو کشور دیگر، جنسیت و آموزش مهارت‌های رمزگشایی و گروه‌های کوچک، به روش مورداستفاده برای بررسی منابع DIF طبقه مکنون وابسته است (Oliveri and et al., 2013).

در مطالعه‌ای دیگر با استفاده از ادغام نتایج به‌دست‌آمده از تحلیل کمی DIF و مصاحبه شناختی، تفسیرهای شرکت‌کنندگان به‌وسیله تم‌ها و زیرتم‌ها که از روایات آن‌ها به‌دست‌آمده، مقایسه شدند. نتایج این مطالعه نشان‌داد سؤال‌ها با DIF فراگیر منجر به فرایندهای پاسخ مختلف به عناصر موجود در متن سؤال می‌شود، یعنی افراد هنگامی‌که به این سؤال‌ها پاسخ می‌دهند، درباره مسائل مختلفی فکر می‌کنند. بنابراین می‌توان گفت سؤال‌ها با DIF فراگیر حاوی عناصر مفهومی متفاوت است که دارای معانی مختلفی هستند درحالی‌که سؤال‌های با DIF غیرفراگیر، به‌دلیل تجارب مختلف مرتبط

با زمینه یا فرهنگ، متفاوت هستند. نتایج همچنین نشان‌دهنده وجود مشکلات احتمالی در انطباق، ناشی از استفاده از اصطلاحاتی است که در دو نسخه معادل نیستند (Benítez, Padilla, 2014).

برای بررسی تأثیر زبان مادری آزمودنی بر DIF، از جزءبندی بازگشتی مبتنی برمدل[1] (MBRP) برای بررسی DIF یکنواخت استفاده شده است، نتایج نشان داد که DIF بر مبنای زبان مادری در چندین سؤال پیرلز وجود دارد اما الگوی DIF در همه کشورها یکسان نیست (Finch and et al., 2016).

به‌منظور بررسی اثربخشی سه روش پیاده‌سازی MIMIC، نقطه شروعِ ثابت، نقطه شروع آزاد و روش جدید نقطه شروعِ آزاد متوالی(MIMIC) مطالعه شبیه‌سازی برای شناسایی کارکرد افتراقی یکنواخت و غیریکنواخت سؤال برای چندگروه، انجام شده است. نتایج نشان می‌دهد روش نقطه شروعِ به‌طور متوالی آزاد خطای نوع اول و توان آزمون عالی فراهم می‌نماید و نتایج آن مشابه با روش نقطه شروع آزاد ایده‌آل است که از طرح لنگر فاقد DIF استفاده می‌کند و خیلی بهتر از روش نقطه شروع ثابت است که از همه سؤال‌ها به‌جای سؤال‌های موردمطالعه به‌عنوان لنگر استفاده می‌کند (Chun and et al., 2016).

در بررسی تأثیر چیرگی در ویژگی‌ها و مهارت‌های زیربنایی برای عملکرد در سؤال‌ها، مدل‌های شناختی تشخیصی ماتریس Q که شامل ویژگی‌های زیربنایی عملکرد در سؤال‌ها است تشکیل‌شده است و تفاوت معنادار احتمال چیرگی در ویژگی‌ها در گروه‌های مورد مقایسه، با آزمون مانوا انجام شد و نتایج نشان داد که به‌طور متوسط تفاوت در میزان چیرگی بر یک ویژگی بین دریافت‌کننده و عدم‌دریافت‌کنندگان مساعدت در حدود ۳۳ تا ۴۴ درصد است (Svetina and et al., 2017).

راهبردهای غیر آماری برای بررسی تأثیر منابع DIF

راهبردی که برای شناسایی انطباق نسخه‌های آزمون به‌عنوان منبع DIF به‌کارگرفته شده است، مرور قضاوتی با مترجمان چندزبانه و روایی متقاطع DIF در گروه‌های چندگانه است، دو شاهد برای حمایت از این فرض که انطباق منبعی برای DIF است به‌کار گرفته شد، اول شناسایی تفاوت‌ها در معانی، ساختار و قالب بین نسخه‌های ترجمه شده سؤال‌ها در مرور قضاوتی و دوم روایی متقاطع DIF در دو مقایسه اضافه است (Ercikan, 2002). راهبرد دیگر برای سنجش منابع یافته شده برای DIF در انطباق نسخه‌های چندزبانه آزمون، تشکیل کمیته دومی از متخصصان است، این کمیته شامل متخصصان زبانشناسی و معلمانی است که به‌طور مستقل از اولین کمیته (برای شناسایی منابع DIF) کار می‌کنند. همچنین برای بررسی تأثیر پوشش برنامه درسی به‌عنوان منبعی برای DIF، توزیع DIF سؤال‌ها به‌وسیله موضوعات درسی بررسی شد. نتایج این پژوهش نشان داد، ۲۳ درصد سؤال‌های DIF ریاضی و ۱۳ درصد سؤال‌های DIF علوم در تفاوت در برنامه درسی دارند و ۲۷ درصد در ریاضی و ۳۷درصد

1. model-based recursive partitioning

سؤال‌های در علوم دارای DIF به دلیل انطباق نسخه‌های آزمون است و ۵۰ درصد سؤال‌ها به دلیل انطباق نسخه‌های چندزبانه آزمون تیمز و برنامه درسی نبود (Elosua, López-jaúregui, 2007).

راهبرد مورداستفاده برای تأثیر سن افراد در DIF، تحلیل محتوای سؤال‌هایی است که به‌وسیله متخصصان، دارا یا فاقد DIF شناسایی‌شده است. قضاوت متخصصان به‌طور واضح منابع را برای سؤال‌های دارای DIF مشخص نمی‌کند. در این مطالعه، آزمون سوگیری علیه گروه‌های سنی ندارد (Ardeshir, Antony, 2007).

در مطالعه دیگر مشخصه‌های از سؤال‌ها که به‌وسیله مرور متخصصان به‌عنوان منبع DIF بین دانش‌آموزان از دو زبان، شناسایی‌شده است را به‌وسیله تأیید شواهد تجربی از پروتکل تفکر با صدای بلند بررسی نمودند. خواندن با صدای بلند برای درک اشتباه خواندن سؤال ضروری بود، هنگامی‌که دو نسخه زبانی دقیقاً معنای یکسانی داشتند اما دارای واژگان مستعد خطا هستند. منابع شناسایی شده در این مطالعه کلمات کلیدی سؤال که ممکن است کمک یا مانع برای حل سؤال باشد، وضوح و هدایت زبان، قالب‌بندی و ارائه بصری، تفاوت‌های زبانی غیرمنتظره (اشتباه خواندن کلمه توسط دانش‌آموزان فرانسوی‌زبان) است که به‌جز قالب‌بندی و ارائه بصری، وضوح و هدایت زبانی، پروتکل تفکر با صدای بلند برای سایر منابع، شواهد تأییدی فراهم نمود (Ercikan and et al., 2010).

در مطالعه دیگر برای بررسی تأثیر ویژگی‌های سؤال بر DIF از تحلیل محتوای تعقیبی سؤال‌ها کمک گرفته‌است، مثلاً توضیح احتمالی برای DIF غیریکنواحت مشاهده شده بر مبنای جنسیت این است که مردانی با توانایی پایین احتمالاً تمایل به ریسک داشتند و الگوی موفقیت‌های آنان با خوش‌شانسی در سؤال‌های آزمون با گزینه‌های غیر جذاب بوده‌است (Aryadoust and et al., 2011).

به‌دلیل اینکه مقایسه بین دو گروه با زبان، فرهنگ و برنامه درسی یکسان، زبان و فرهنگ مشابه اما برنامه درسی متفاوت، زبان و فرهنگ و برنامه درسی متفاوت، امکان‌پذیر شود، چهار گروه از دانش‌آموزان از ایالت متحده، کانادا، چین و هنک‌کنگ انتخاب شدند. راهبرد بررسی تأثیر منابع شناسایی‌شده شامل ترجمه آزمون، پوشش برنامه درسی متفاوت و تفاوت‌های فرهنگی، استفاده از یک شخص دو زبانه برای بررسی اثر ترجمه (به‌دلیل محرمانه بودن اطلاعات تنها یک نفر انتخاب شده) است، استفاده از یک پانل شامل ۱۰ نفر متخصص بررسی محتوا (معمولاً معلمانی که با محتوای برنامه درسی آشنایی دارند) و تعداد ۱۵دانش‌آموزان برای بررسی اثر پوشش برنامه درسی، برای بررسی اثر تفاوت فرهنگی بر DIF است. نتایج تحقیق نشان داد که جدی‌ترین عامل DIF در بین سه عامل نام برده‌شده، پوشش افتراقی برنامه‌های درسی است و آشنایی با محتوای آزمون به طور افتراقی نیز در DIF سهم داشته‌است. جدی‌ترین DIF بین دانش‌آموزان چین و ایالات متحده وجود دارد و بین دانش‌آموزان انگلیسی‌زبان کمترینDIF نشان داده‌شد (Huang and et al., 2016).

بحث و نتیجه‌گیری

یافته‌های مرور مقالات در مورد منابع DIF نشان داد، برای سازندگان آزمون‌ها یکی از اهداف اصلی در تحقیق DIF، درک بهتر علل DIF در سؤال‌های آزمون است. محدودیت مطالعات DIF این است که آن‌ها صرفاً بر مبنای مشاهدات هستند و در نتیجه تنها اجازه استنباط‌های مربوط به رابطه بین ویژگی‌های سؤال و DIF را می‌دهند (Schmitt and et al.,1993). به همین دلیل بیشتر مطالعات انجام شده به بررسی ویژگی‌های سؤال به‌عنوان منبعی برای DIF پرداخته‌اند. سازندگان آزمون لازم است در هنگام ساخت آزمون توجه کامل به ویژگی‌های سؤال نمایند خصوصاً اینکه تقریباً در تمامی مقالات موردبررسی، معناداری این عوامل تأیید شده است. راهکارهای مختلفی برای جلوگیری از تأثیر ویژگی-های سؤال به‌عنوان منابع DIF در مقالات اشاره شده است. به‌عنوان‌مثال بررسی کارکرد افتراقی گزینه انحرافی (DDF) ضروری است زیرا با ترکیب DIF و DDF تحلیل کامل‌تری از کارکرد سؤال ارائه می‌شود. ویژگی دیگر سؤال، بارهای شناختی است که منابع اساسی از DIF را نشان می‌دهد. با توجه به این منبع DIF، سازندگان آزمون بهتر است آزمون‌هایی طراحی کنند که شامل سؤال‌هایی با نسبت متعادل بارهای شناختی بالاتر نسبت به بارهای شناختی پایین‌تر باشند. محتوای سؤال‌ها همچنین می‌تواند علت احتمالی DIF باشد. برخی از عناصر در سؤال‌های داری DIF ممکن است برای آزمودنی-های با توانایی پایین دشوارتر باشد، ازجمله، عبارات منفی، اصطلاحات، استعاره‌ها و سؤال‌ها با طولی که ممکن است در حافظه آزمودنی‌ها با توانایی پایین باقی نماند. تفاوت‌های زبانی ممکن است به‌وسیله آزمودنی‌ها با توانایی بالا بهتر مدیریت شود، آزمودنی‌هایی که ممکن است با استفاده از اشارات زمینه‌ای یا زبانی، به راهبردهای استنباطی متوسل شوند. عامل حدس زدن و طول سؤال، سازه‌های مزاحم هستند و روایی استدلال آزمون را کاهش می‌دهند. طراحان آزمون باید احتمال حدس را با افزایش گزینه‌های سؤال به چهار یا حتی پنج گزینه محدود نمایند.

در مطالعات مربوط به انطباق نسخه‌های ترجمه شده و برنامه درسی، محققان اعلام می‌کنند تنها در نظر گرفتن انطباق نسخه‌های زبانی به‌عنوان منبع DIF کافی نیست، بلکه عوامل دیگری هم که می‌تواند به‌طور بالقوه DIF را توضیح دهد، بهتر است در نظر گرفت. از این جمله می‌توان به تفاوت‌های روش‌های آموزشی، تفاوت‌های فرهنگی و محدودیت در تعاریف موضوعات اشاره نمود (ارسی کان، ۲۰۰۲). یکی از محدودیت‌های مطالعات انطباق آزمون این است که تنها شامل دو زبان است. تکرار این مطالعات با استفاده از زبان‌های دیگر و انجام تحلیل‌های همزمان DIF چندگانه بین چندین زبان، راه‌حل پیشنهادی برای تحقیقات آینده است. تحقیقات بیشتر در مورد علل DIF در سؤال‌های ترجمه شده می‌تواند بر اساس ایده‌های زیر طراحی شود: ۱- تمرکز بر سؤال‌هایی که نشان‌دهنده DIF نیستند ۲- قرار دادن پرسشنامه‌ای برای متقاضیان در هر دو گروه که از آن‌ها می‌خواهد درباره پاسخ‌شان به سؤالی خاص توضیح‌دهند. تحلیل توضیحات افرادی که به سؤال دارای DIF پاسخ نادرست دادند ممکن است به درک بهتر دلیل DIF بینجامد (Amery,Ercikan, 2006).

اگر علیه دانش‌آموزانی سوگیری به دلیل فقدان موضوع سؤال در برنامه درسی‌شان ایجاد شود، احتمال دارد سوگیری به‌وسیله طراحی مناسب سؤال از بین برود. از محققان و آموزگاران خواسته می‌شود در طراحی سؤال به چند مورد مهم فکر کنند: دانش مشترک و فرآیندهای مورد انتظار از دانش‌آموز بدون در نظر گرفتن کشور مبدأ، چیست؟ چه علومی دانش‌آموزان باید بداند تا بتوانند به‌عنوان نیروی کار جهانی در آینده کار کنند؟ چه مواد آموزشی باید در برنامه‌های درسی جدید قرار داده شود تا به توسعه توانایی علمی و مهارت دانش‌آموز که موردنیاز جامعه مدرن است، کمک کند؟ تفاوت فرهنگی یکی دیگر از منابع بالقوه DIF است. تنها جنبه‌ای که مطالعات حاضر در این زمینه یافته‌اند، آشنایی متفاوت با محتوای آزمون بوده است.

منابع متعددی برای DIF وجود دارد و بسته به نوع و هدف گروه‌های موردسنجش و مقایسه، معنا و منابع DIF ممکن است متفاوت باشد. در نظر داشته باشید که یک متغیر برای منبع DIF در محتوای خاص و برای یک هدف خاص ممکن است منبع مناسبی برای DIF در سایر موارد نباشد و ویژگی‌های فرهنگی یک منبع ذاتی DIF نیست، همان‌طور که تأثیر ویژگی‌های فرهنگی به‌عنوان منبعی برای DIF تأیید نشد (Elosua, López-jaúregui, 2007). تحقیقات آینده می‌تواند بر متغیرهایی که به‌طور مستقیم مربوط به زمینه‌های آزمون است، مثل تفاوت در دانش تئوری دانش‌آموزان، مهارت‌های آزمون دادن، استفاده از ماشین‌حساب یا نگرش نسبت به ریاضیات، متمرکز شود. در سنجش‌های چندزبانه، هم‌ارزی سؤال‌ها برای مقایسه گروه‌ها بیشتر توسط نسخه‌های چندزبانه آزمون تحت چالش قرار می‌گیرد. همچنین لازم است در نظر داشته باشید که تأثیر این عوامل بر DIF بسته به اینکه آیا آزمون یک آزمون روان‌شناختی، پیشرفت تحصیلی یا آزمون کسب مجوز است، متفاوت است.

در مورد روش‌های شناسایی منابع DIF، بیشتر مقالات با استفاده از تحقیقات گذشته به شناسایی منابع پرداخته‌اند و همچنین برخی مقالات از نظر متخصصان استفاده نموده‌اند، هرچند استفاده از نظر متخصصان برای شناسایی منابع DIF بسیار مؤثر است و تعداد زیادی از مطالعات مورد بررسی از این شیوه استفاده نموده‌اند ولی پروتکل تفکر با صدای بلند منابعی از DIF را نشان می‌دهد که به‌وسیله مرور متخصصان مشخص نشده است. از طرفی سؤال‌هایی که توسط متخصصان دارای DIF شناسایی نشده در پروتکل تفکر با صدای بلند قرار نمی‌گیرد، بنابراین امکان بررسی تفاوت‌های گسترده بین گروه‌ها وجود ندارد. بااین‌حال در آزمون‌ها با تعداد زیادی از سؤال‌ها، ورود همه سؤال‌ها در پروتکل تفکر با صدای بلند برای بررسی مقایسه‌پذیری سازه عملی نیست. محدودیت پروتکل تفکر با صدای بلند، در نمونه محدود دانش‌آموزان مورداستفاده در مطالعه است. با تعداد نمونه خیلی کوچکی که به‌طورمعمول در این پروتکل استفاده می‌شود، نمی‌توان انتظار داشت که، نماینده جمعیت مربوطه باشد (Ercikan and et al., 2010).

استفاده از روش طبقه‌های مکنون برای شناسایی منابع DIF مفید است، ازآنجایی‌که چندبعدی بودن آزمون‌ها علت اصلی DIF است، درک چندبعدی بودن آزمون و اثرات این ابعاد بر DIF، توانایی

تفسیر دقیق‌تر نمرات آزمون، کنترل بیشتر بر ابعاد مزاحم مرتبط و کاهش تأثیر آن‌ها را فراهم می‌نماید (Cohen, Bolt, 2005). یکی از محدودیت‌های مطالعه DIF، فقدان اظهارات صریح در مورد ابعاد اولیه و ثانویه مربوط به سؤال‌های آزمون از سازندگان آزمون است، که اگر ابعاد آزمون مشخص بود، می‌توانست در فرضیه‌سازی DIF در سؤال‌های چندبعدی کمک کند و سپس امکان آزمون فرضیه‌ها به‌طور کامل به‌وسیله رویکرد تحلیل DIF مبتنی بر چندبعد، علاوه بر روش شناسایی DIF با SIBTEST استات[1]، وجود داشت (Roussos, Stout, 2004).

به‌طورکلی مرور مقالات نشان می‌دهد که مطالعات انجام شده یا اینکه تنها بر شناسایی منابع DIF تمرکز داشته‌اند و راهبردی برای بررسی تأثیر منابع DIF شناسایی‌شده به کار نگرفته‌اند و یا تمرکز اصلی مطالعه در راهبردی برای بررسی تأثیر منابع DIF شناسایی‌شده در مطالعات گذشته یا منابعی که توسط متخصصان یافته شده، بوده است، این مسئله از نقاط ضعف مطالعات است، انجام مطالعه‌ای که هم بر شناسایی منابع DIF و هم بر انتخاب راهبرد مناسب برای بررسی تأثیر منابع شناسایی‌شده تمرکز داشته باشد، مطالعه‌ای کامل در زمینه منابع DIF محیا خواهد نمود.

در راهبردهای کمی برای بررسی تأثیر منابع DIF مثل مدل‌سازی رگرسیونی، لازم است توجه کنیم که توفیق استفاده از این روش‌ها در انتخاب دقیق متغیرها، هم از لحاظ نظری و هم از نظر آماری است. همچنین مدل‌سازی MIMIC می‌تواند تغییرات را در پاسخ سؤال‌هایی که با سازه زیربنایی ارتباط ندارد، مشخص نماید، اما اطلاعاتی درباره اینکه چرا این تغییرات ممکن است وجود داشته باشد ارائه نمی‌دهد. اگرچه مدل‌سازی MIMIC قادر به سنجش تفاوت‌ها در مورد دشواری سؤال (DIF یکنواخت) است اما نمی‌سنجد که آیا پارامتر تشخیص سؤال‌ها در گروه‌ها یکسان است (DIF غیریکنواخت). در مورد راهبردهای غیر آماری مورداستفاده باید در نظر داشت که در شناسایی منابع DIF با استفاده از مرور قضاوتی، تفسیرها و تلاش‌های برای شناسایی منابع DIF می‌تواند وابسته به اندیشه فرد باشد. این مسئله حادتر می‌شود اگر مرورگران بدانند که کدام سؤال‌ها به‌طور افتراقی عمل می‌کنند. منابع چندگانه در بررسی منابع DIF و فرایندهای مرور قضاوتی باید در نظر گرفته شود، با تمرکز بر یک منبع انتظار نداشته باشید که منبع DIF را برای همه سؤال‌هایی که دارای DIF شناسایی‌شده‌اند، توضیح دهد.

همچنین مرور تخصصی سؤال‌ها باید به‌وسیله افرادی که در مورد یادگیری دانش‌آموزان آگاه هستند و تخصص فرهنگی یا زبانی دارند انجام شود، این روش رایج‌ترین روش برای شناسایی خواصی (مثل محتوا، قالب، زمینه و زبان) از سؤال‌های دارای DIF آزمون است. بااین‌حال حتی اگر مرور تخصصی بتواند مشخص کند که آیا برخی جنبه‌های سؤال‌های آزمون به DIF ارتباط دارد، نمی‌تواند منابع DIF را شناسایی‌کند. به‌علاوه مرور تخصصی توضیح نمی‌دهد که چگونه مشخصات سطحی سؤال ممکن

1. Stout

است منجر به کارکرد افتراقی بین گروه‌های آزمودنی شود. برای پاسخ به سؤال‌های چگونه و چرا لازمست اثر متقابل زبان سؤال‌های آزمون و فرایند تفکر آزمودنی درک شود.

در نهایت سنجش کارکرد افتراقی سؤال در روایی نمرات آزمون مسئله‌ای کلیدی است. با توجه به افزایش وابستگی سیاست‌گذاران آموزشی بر ارزیابی‌های بین‌المللی مانند آزمون‌های تیمز و پیرلز و آزمون‌های ملی مثل آزمون‌های استخدامی و آموزش عالی، عدم‌توجه به منابع DIF می‌تواند منجر به استنباط‌هایی اشتباه در مقایسه نمرات آزمون‌ها شود. سیاست‌گذاران باید هنگام تصمیم‌گیری در مورد برنامه درسی، منابع یا آموزش بر مبنای هر مقایسه مستقیم با استفاده از سؤال‌های آزمون‌ها، بسیار مراقب باشند. روایی مقایسه بین گروه‌ها همیشه باید قبل از مقایسه نمرات، بررسی شود.

پیشنهادها

۱- سازندگان آزمون لازم است در هنگام ساخت آزمون توجه کامل به ویژگی‌های سؤال نمایند خصوصاً اینکه تقریباً در تمامی مقالات موردبررسی، معناداری این عوامل تأیید شده است و بهتر است آزمون-هایی طراحی کنند که شامل سؤال‌هایی با نسبت متعادل بارهای شناختی بالاتر نسبت به بارهای شناختی پایین‌تر باشند.

۲- یکی از محدودیت‌های مطالعات انطباق آزمون این است که تنها شامل دو زبان است. تکرار این مطالعات با استفاده از زبان‌های دیگر و انجام تحلیل‌های همزمان DIF چندگانه بین چندین زبان، راه‌حل پیشنهادی برای تحقیقات آینده است.

۳- تحقیقات آینده می‌تواند بر متغیرهایی که به‌طور مستقیم مربوط به زمینه‌های آزمون است، مثل تفاوت در دانش تئوری دانش‌آموزان، مهارت‌های آزمون دادن، استفاده از ماشین‌حساب یا نگرش نسبت به ریاضیات، متمرکز شود. در سنجش‌های چندزبانه، هم‌ارزی سؤال‌ها برای مقایسه گروه‌ها بیشتر توسط نسخه‌های چندزبانه آزمون تحت چالش قرار می‌گیرد.

۴- اظهارات صریح در مورد ابعاد اولیه و ثانویه مربوط به سؤال‌های آزمون توسط سازندگان آزمون می‌تواند در فرضیه‌سازی DIF در سؤال‌های چندبعدی کمک کند و امکان آزمون فرضیه‌ها را به‌طور کامل فراهم نماید.

۵- انجام مطالعه‌ای که هم بر شناسایی منابع DIF و هم بر انتخاب راهبرد مناسب برای بررسی تأثیر منابع شناسایی‌شده تمرکز داشته باشد، مطالعه‌ای جامع در زمینه منابع DIF محیا خواهد نمود.

References:

Albano, A. D., & Rodriguez, M. C. (2013). Examining differential math performance by gender and opportunity to learn. *Educational and Psychological Measurement*, *73*(5), 836–856.

Allalouf, A., Hambleton, H. K., & Sireci, S. G. (1999). Identifying Causes of DIF in Translated Verbal Items. *Journal of Educational Measurement*. 36(2). 185-198.

Amery D. Wu. & Ercikan K. (2006). Using Multiple-Variable Matching to Identify Cultural Sources of Differential Item Functioning, *International Journal of Testing*, 6:3, 287-300, DOI: 10.1207/s15327574ijt0603_5.

Angoff, W. H. (1993). Perspective on differential item functioning methodology. In P. W. Holland., & H. Wainer. (Eds.). *Differential item functioning* (pp. 3–24). Hillsdale, NJ: Erlbaum.

Ardeshir, G., & Antony, J., K. (2007). Differential Item Functioning in Terms of Age in the Certificate in Advanced English Examination , *Language Assessment Quarterly*, 4(2), 190-222, DOI: 10.1080/15434300701375758

Aryadoust, V., Goh, C. C. M., & Kim, L. (2011). An Investigation of Differential Item Functioning in the MELAB Listening Test, *Language Assessment Quarterly*, 8(4), 361-385, DOI: 10.1080/15434303.2011.628632

Banks, K. (2009). Using DDF in a Post Hoc Analysis to Understand Sources of DIF, *Educational Assessment*, 14(2), 103-118, DOI: 10.1080/10627190903035229

Benítez, I., & Padilla, J. (2014). Analysis of Nonequivalent Assessments across Different Linguistic Groups Using a Mixed Methods Approach: Understanding the Causes of Differential Item Functioning by Cognitive Interviewing, *Journal of Mixed Methods Research*, 8(1), 52-68, DOI: 10.1177/1558689813488245.

Bolt, M. D. (2000). A SIBTEST Approach to Testing DIF Hypotheses Using Experimentally Designed Test Items, *Journal of Educational Measurement*, 37(4), 307-327.

Chun, S., Stark, S., Kim, E. S., & Chernyshenko, O. S. (2016). MIMIC Methods for Detecting DIF Among Multiple Groups: Exploring a New Sequential-Free Baseline Procedure, *Applied Psychological Measurement*, 40(7), 486-499.

Cohen, A., & Bolt, D. (2005). A Mixture Model Analysis of Differential Item Functioning. *Journal of Educational Measurement*. 42, 133 - 148. 10.1111/j.1745-3984.2005.00007.

Diamond, R., & Persson, P. (2016). The long-term consequences of teacher discretion in grading of high-stakes tests. *National Bureau of Economic Research*, 7 (12), 220-227.

Elosua, P., & López-jaúregui, A. (2007). Potential Sources of Differential Item Functioning in the Adaptation of Tests, *International Journal of Testing*, 7(1), 39-52, DOI: 10.1080/15305050709336857

Ercikan, K. (2002). Disentangling Sources of Differential Item Functioning in Multilanguage Assessments, *International Journal of Testing*, 2:3-4, 199-215, DOI: 10.1080/15305058.2002.9669493

Ercikan, K. (2008). Limitations in sample to population generalizing. In K. Ercikan & M.-W. Roth (Eds.), *Generalizing in educational research* (pp. 211–235). New York, NY: Routledge.

Ercikan, K., Arim, R., Law, D., Domene, J., Gagnon, F., & Lacroix, S. (2010). Application of Think Aloud Protocols for Examining and Confirming Sources of Differential Item Functioning Identified by Expert Reviews. *Educational Measurement: Issues and Practice*. 29. 10.1111/j.1745-3992.2010.00173.x.

Ercikan, K., Gierl, M. J., McCreith, T., Puhan, G. & Koh, K. (2004). Comparability of bilingual versions of assessments: Sources of incomparability of English and French versions of Canada's national achievement tests. *Applied Measurement in Education*, *17*(3), 301–321.

Finch, W. H., Finch, M. E. H., & French, B.F. (2016). Recursive Partitioning to Identify Potential Causes of Differential Item Functioning in Cross-National Data, *International Journal of Testing*, 16(1), 21-53, DOI: 10.1080/15305058.2015.1039644

Huang, X., Wilson, M., & Wang, L. (2016). Exploring plausible causes of differential item functioning in the PISA science assessment: language, curriculum or culture, *Educational Psychology*, 36(2), 378-390, DOI: 10.1080/01443410.2014.946890

Li, Y., Cohen, A. S., & Ibarra, R. A. (2004). Characteristics of mathematics items associated with gender DIF. *International Journal of Testing*, *4*(2), 115–136.

Oliveri, M.E. & von Davier, M. (2014). Toward Increasing Fairness in Score Scale Calibrations Employed in International Large-Scale Assessments, *International Journal of Testing*, 14(1), 1-21, DOI: 10.1080/15305058.2013.825265

Oliveri, M.E., Ercikan, K., & Zumbo, B. (2013). Analysis of Sources of Latent Class Differential Item Functioning in International Assessments, *International Journal of Testing*, 13(3), 272-293, DOI: 10.1080/15305058.2012.738266

Pellegrino, J.W., Chudowsky, N., & Glaser, R. (2001). *Knowing what students know: The science and design of educational assessment.* Washington, DC: National Academy Press.

Roth, W. M., Oliveri, M. E., Sandilands, D. D., Lyons-Thomas, J., & Ercikan, K. (2013). Investigating Linguistic Sources of Differential Item Functioning Using Expert Think-Aloud Protocols in Science Achievement Tests, *International Journal of Science Education*, 35(4), 546-576, DOI: 10.1080/09500693.2012.721572

Sandilands, D., Oliveri, M. E., Zumbo, B. D., & Ercikan, K. (2013). Investigating Sources of Differential Item Functioning in International Large-Scale Assessments Using a Confirmatory Approach, *International Journal of Testing*, 13(2), 152-174, DOI: 10.1080/15305058.2012.690140

Schmitt, A. P., Holland, P. W., & Dorans, N. J. (1993). Evaluating hypotheses about differential item functioning. In P. Holland & H. Wainer (Eds.), *Differential Item Functioning* (pp. 281-316). Hillsdale, N J: Lawrence Erlbauna.

Suh,Y., & Bolt, D. M. (2011). A Nested Logit Approach for Investigating Distractors as Causes of Differential Item Functioning, *ournal of Educational Measurement, 48(2),* 188-205.

Svetina, D., Dai, S., & Wang, X. (2017). Use of cognitive diagnostic model to study differential item functioning in accommodations, Behaviormetrika, 44, 313-349. https://doi.org/10.1007/s41237-017-0021-0.

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.