فصلنامه مطالعات اندازه‌گیری و ارزشیابی آموزشی
سال هفتم، شمارهٔ ۱۹، پاییز ۱۳۹۶، صص ۶۳-۳۵

Journal of Educational Measurement & Evaluation Studies
Vol. 7, No. 19, Autumn 2017

کنش افتراقی سؤال (DIF) و آزمون (DTF) مرتبط با جنسیت در آزمون‌های کنکور سراسری سازمان سنجش آموزش کشور[1]

مسعود گرامی‌پور*
اعظم رمضانی صدر**
احمد رضایی***
لیلا نوروزی****
فرانک مختاریان*****

تاریخ دریافت مقاله: ۹۵/۰۷/۱۲
تاریخ پذیرش مقاله: ۹۶/۰۳/۲۷

چکیده

عملکرد آزمودنی با کنترل آماری توانایی ایشان در گروه های مختلف جنسیتی ممکن است متفاوت باشد. در این صورت، وقوع کنش افتراقی سؤال (DIF) / آزمون (DTF) مرتبط با جنسیت و به دنبال آن سوگیری سؤال/آزمون محتمل است. هدف اصلی از اجرای این پژوهش، مطالعه DTF/DIF و سوگیری جنسیتی در آزمون‌های کنکور سراسری ایران است. آزمون‌های تخصصی یک دفترچه آزمون در پنج گروه آزمایشی شرکت‌کننده در کنکورهای سراسری سال های ۱۳۸۷ تا ۱۳۹۰ به صورت خوشه‌ای تک مرحله‌ای انتخاب شدند. سپس برای مطالعه DIF از تحلیل رگرسیون لوجستیک و برای مطالعه DTF از رویکرد مبتنی بر IRT استفاده شد. نتایج نشان داد که به‌طور متوسط، حدود ١٤ درصد از پرسش‌های آزمون های مورد مطالعه دارای DIF جنسیتی با اندازه اثر ناچیز (EF<۰/۰۰۰۱) هستند و حداکثر ۲ درصد آنها بر اساس نقطه نظرات گروه‌های خبرگان نسبت به جنس مؤنث یا مذکر سوگیری دارند. همچنین یافته‌های تحلیل DTF نشان داد که به جز آزمون خلاقیت‌های نمایشی گروه هنر که نسبت به زنان دارای سوگیری کمی است، سایر آزمون ها فاقد DTF هستند.

واژگان کلیدی: کنش افتراقی سؤال/آزمون، DTF/DIF مرتبط با جنسیت، سوگیری سؤال/آزمون، آزمون‌های ورودی دانشگاه، کنکور سراسری

---

[1]. این پژوهش با حمایت سازمان سنجش آموزش کشور اجرا شده است.

* استادیار دانشکده روان‌شناسی و علوم تربیتی دانشگاه خوارزمی (نویسنده مسئول: mgramipour@yahoo.com)
** کارشناسی ارشد تحقیقات آموزشی، دانشگاه خوارزمی
*** دانشجوی دکتری رشته سنجش و اندازه‌گیری دانشگاه علامه طباطبایی
**** کارشناسی ارشد تحقیقات آموزشی، دانشگاه خوارزمی
***** کارشناس پژوهشی سازمان سنجش آموزش کشور

مقدمه

آزمون‌های بزرگ‌مقیاسی مانند کنکور سراسری که سازمان سنجش آموزش کشور در ایران برگزار می‌کند، نقش کلیدی در آینده جوانان دارند. دختران و پسران بعد از گذراندن دوره متوسطه برای ورود به دانشگاه و رشته مورد علاقه خود، باید با یکدیگر در کنکور سراسری رقابت کنند. رقابت در چنین آزمونی مشکل بوده و شاید بتوان گفت مهم‌ترین آزمون در ایران است که به‌صورت سالیانه برگزار می‌شود. والدین دانش‌آموزان نیز زمان و هزینه بسیاری متحمل می‌شوند تا فرزندان آنها در کنکور سراسری پذیرفته شوند. یکی از مهم‌ترین متغیرهایی که در چنین آزمونی باید مد نظر قرار گیرد، جنسیت است. به‌طوری که این اطمینان حاصل شود که آزمون برای دختران و پسران به یک اندازه منصفانه است. بنابراین مطالعه روایی[1] آزمون بر اساس جنسیت بسیار اهمیت دارد. در دهه‌های گذشته، متخصصان اندازه‌گیری و پژوهشگران درباره کارکرد متفاوت آزمون‌های روانی در گروه‌های مختلف (به‌طور مثال زنان و مردان) به‌طور فزاینده‌ای نگران شده‌اند. به همین دلیل، پژوهش‌های زیادی درباره سوگیری سؤال در آزمون‌های روانی و تربیتی انجام شده است (امبرتسون[2]، ۲۰۰۰). در چنین پژوهش‌هایی، پرسش این‌چنین مطرح می‌شود: آیا نمره‌های زنان و مردان در جامعه‌های مختلف در یک مقیاس اندازه‌گیری قابل مقایسه است؟ زمانی گفته می‌شود سؤالی دارای کنش افتراقی (DIF[3]) است که ویژگی‌های روان‌سنجی آن در گروه مورد نظر (گروه کانونی[4]) در مقایسه با گروه مورد مقایسه (مرجع[5]) متغیر باشد (دراسگو[6]، ۱۹۸٤، ۱۹۸۷). کارکرد افتراقی آزمون (DTF[7]) نیز زمانی اتفاق می‌افتد که توابع نمره‌های کل آزمون در گروه‌ها (مثلاً جنسیت) متفاوت باشد. در آزمون‌های دارای DTF نمره‌های واقعی مورد انتظار آزمون در دو گروه آزمودنی‌ها یکسان نیست. بنابراین می‌توان گفت که کنش افتراقی در دو سطح

1. Validity
2. Embretson
3. Differential Item Functioning
4. Focal group
5. Reference
6. Drasgow
7. Differential Test Functioning

اتفاق می‌افتد: ۱- سطح سؤال (DIF)؛ ۲- مجموع سؤال‌های آزمون (DTF) (رادنر[1]، جتسون[2] و نایت[3]، ۱۹۸۰).

در پیشرفت‌های اخیر حوزه DTF دو نوع آزمون کلی[4] آن شامل sDTF و uDTF که مبتنی بر نظریه پاسخ سؤال (IRT[5]) است، توسعه یافته است. در این روش‌ها، کنش افتراقی آزمون با رویه‌های انتگرال‌گیری سطوح میان منحنی‌های کل آزمون در دو گروه آزمودنی (مثلاً جنسیت مذکر و مؤنث) محاسبه می‌شود. در این رویکرد sDTF با بررسی اختلاف دو گروه مرجع و کانونی، مقداری علامت‌دار مبنی بر ارجحیت گروه کانونی (مثبت) یا گروه مرجع (منفی) فراهم می‌کند، در حالی که uDTF متوسط این اختلاف را بدون علامت مثبت یا منفی نشان می‌دهد. مقادیر منفی sDTF بر این دلالت دارد که نمره گروه مرجع به‌طور متوسط کمتر از گروه کانونی است، در حالی که مقادیر مثبت آن دلالت بر توفق نمره‌های گروه کانونی دارد. مقدار uDTF بر قدر مطلق انحراف‌های مقادیر پرسش دلالت دارد که روی کل آزمون متراکم شده و نشانگر متوسط مساحت میان دو منحنی آزمون است. دامنه نمره‌های uDTF بین ۰ تا نمره آزمون است و اگر شکل تابعی منحنی‌های آزمون دقیقا شکل هم باشد، مساحت میان منحنی‌ها صفر است. هنگامی‌که DTF کلی به‌واسطه sDTF و uDTF بررسی می‌شود، سه سناریوی پایانی ممکن است مشاهده شود: ۱- هنگامی‌که sDTF و uDTF هر دو کوچک است، در این حالت، مقدار DTF در کل دامنه توانایی بسیار کوچک یا صفر است؛ ۲- هنگامی که sDTF کوچک و uDTF بزرگ است، منحنی‌های آزمون در یک یا چند نقطه یکدیگر را قطع می‌کنند تا یک نمره کل متعادل را بسازند اما در این حالت، سوگیری غیر قابل اغماضی در نقاط به‌خصوصی از سطوح توانایی وجود دارد؛ ۳- هنگامی‌که sDTF و uDTF هر دو بزرگ است، DTF کلی درمجموع نمره‌های آزمون وجود

---

1. Rudner
2. Getson
3. Knight
4. Omnibus test
5. Item Response Theory

دارد و سوگیری و تفاوت منحنی‌های آزمون نظام‌دار است (چالمرز، کانسل و فلورا[1]، ۲۰۱۵).

در مورد ارتباط میان DIF و DTF گمانه‌زنی‌هایی وجود دارد مبنی بر اینکه DTF ناشی از تجمع پرسش‌های DIF آزمون است. راجو، وندرلیندن و فلییر[2] (۱۹۹۵) دو نوع DIF جبرانی و غیر جبرانی را مفهوم‌سازی کرده‌اند که به نظر می‌رسد مجموع پرسش‌های DIF نوع جبرانی آن به DTF در کل آزمون منجر شود. بنابراین حتی اگر تک‌تک پرسش‌های آزمون دارای DIF معنی‌داری نباشند، ممکن است مجموع آن مقدار DTF معنی‌داری را نشان دهد. البته وجود پرسش‌های DIF غیر جبرانی در آزمون ممکن است نوع جبرانی آن را خنثی کرده و آشکارسازی DTF را تحت تأثیر قرار دهد (فلورا، کارن، هوسانگ و ادواردز[3]، ۲۰۰۸). از نقطه‌نظر آماری نیز از آنجایی که کل آزمون نسبت به تک‌تک پرسش‌های آن آگاهی بیشتری فراهم می‌کند، DTF می‌تواند بسیار آسیب‌زننده باشد. از نقطه‌نظر غیر آماری و اجرایی نیز از آنجایی که همه تصمیم‌های سرنوشت‌ساز برای آزمودنی‌ها بر اساس نمره کل آزمون گرفته می‌شود، DTF نقش مخرب‌تری در تصمیم‌گیری‌های نهایی خواهد داشت (راسل[4]،۲۰۰۵؛ پای و پارک[5]، ۲۰۰۶). در چنین وضعیتی نمره‌های نهایی آزمون، اندازه‌گیری یکسانی در جامعه آزمودنی‌ها ندارند. در این صورت، آزمون دارای «تغییرناپذیری اندازه‌گیری[6]» نیست و تفاوت میانگین‌ها در گروه‌های مختلف یا همبستگی نمره‌های آزمون با نمره یک آزمون ملاک، گمراه‌کننده خواهد بود (راجو، وندر لیندن و فلیر، ۱۹۹۵). علی‌رغم DTF/DIF که ویژگی منفی برای آزمون است، تغییرناپذیری اندازه‌گیری، ویژگی مثبتی برای آزمون‌های روانی و تربیتی است. این ویژگی آزمون با روایی سازه‌ای آزمون در ارتباط است. علاوه بر این صرف وجود DTF/DIF آزمون به این معنی نیست که آزمون، سوگیری دارد بلکه مراحل تحلیل محتوای پرسش‌ها و آزمون‌های دارای DIF و DTF توسط متخصصان محتوایی و موضوع درسی به تشخیص نهایی پرسش‌ها و آزمون‌های سودار منجر می‌شود. بنابراین وجود DTF/DIF در آزمون، شرط لازم برای

---

1. Chalmers, Counsell & Flora
2. Raju, Van der Linden & Fleer
3. Flora, Curran, Hussong & Edwards
4. Russell
5. Pae & Park
6. Measurement invariance

سوگیری بوده ولی کافی نیست (گرامی‌پور و فلسفی‌نژاد، ۱۳۹۲). در این رابطه، پژوهش حاضر نیز به منظور مطالعه این‌گونه روایی در آزمون‌های کنکور سراسری سازمان سنجش آموزش کشور، اجرا می‌شود و بر آن است که پرسش‌ها و آزمون‌هایی را که نسبت به جنسیت دارای سوگیری هستند، مشخص کند.

در مورد روایی آزمون‌های کنکور سراسری با استفاده از روش‌های مبتنی بر نظریه‌های کلاسیک و پیشرفته اندازه‌گیری در داخل کشور مطالعات متعددی انجام گرفته است (معلمی اوره، ۱۳۸۷؛ آبکار، ۱۳۹۱؛ حبیبی، ۱۳۹۲؛ نژادنجف، ۱۳۹۳؛ ترکاشوند، ۱۳۹٤ و میری، ۱۳۹٤)، اما درباره روایی آزمون‌های کنکور سراسری ایران از منظر DIF جنسیتی، چند مطالعه بیشتر انجام نگرفته است (براتی و احمدی[1]، ۲۰۱۰؛ امیریان، علوی و فیدالگو[2]، ۲۰۱٤؛ فلاحی‌سرشت، ۱۳۹٤؛ آسیابی، ۱۳۹۱). مطالعه براتی و احمدی (۲۰۱۰) در مورد DIF مرتبط با جنسیت در آزمون زبان تخصصی انگلیسی کنکور سراسری انجام گرفته است. بررسی آنها نشان داد که حدود ٤۳ درصد پرسش‌های آزمون زبان انگلیسی دارای DIF مرتبط با جنسیت هستند. به‌علاوه، آنها نشان دادند که فرمت پرسش‌های آزمون و محتوای آن با DIF جنسیتی در ارتباط است. در پژوهش امیریان، علوی و فیدالگو (۲۰۱٤) درباره DIF مرتبط با جنسیت اجرا شده و در آن آزمون زبان انگلیسی دانشگاه تهران، مطالعه شده است. بررسی آنها نشان داد که حدود ۲۸ درصد پرسش‌های آزمون زبان انگلیسی دانشگاه تهران دارای DIF مرتبط با جنسیت هستند، در حالی که اندازه‌های اثر مربوط به DIF بسیار کوچک بودند. مرحله تحلیل محتوا در این مطالعه نشان داد که پرسش‌هایی که دارای مضامین علوم انسانی هستند، نسبت به مردان و پرسش‌های دارای مضامین علوم سخت نسبت به زنان سوگیری دارند. فلاحی‌سرشت (۱۳۹٤) و آسیابی (۱۳۹۱) نیز DIF مرتبط با جنسیت را به ترتیب در کنکور دکتری و کارشناسی ارشد بررسی کرده‌اند. این در حالی است که تنها یک مطالعه DTF داخلی توسط مینایی (۱۳۹۲) در مورد آزمون علوم پایه هشتم برای مقایسه عملکرد آزمون برای دانش‌آموزان ایرانی و آمریکایی انجام گرفته و تاکنون در ایران هیچ‌گونه مطالعه DTF در مورد آزمون‌های کنکور سراسری انجام نگرفته است.

[1] . Barati & Ahmadi
[2] . Amirian, Alavi & Fidalgo

پژوهش‌های خارجی بسیاری نیز در مورد DIF مرتبط با جنسیت در آزمون‌های بزرگ‌مقیاس، اجرا شده اما برخلاف DIF مطالعات کمی در مورد DTF مرتبط با جنسیت انجام گرفته است. این امر می‌تواند به‌واسطه نو بودن نسبی DTF و روش‌شناسی آن نسبت به روش‌شناسی DIF باشد که به‌تازگی توسعه و تکامل پیدا کرده و کاستی‌های پیشین آن تا حدودی برطرف شده است (چالمررز، کانسل و فلورا، ۲۰۱۵). برای مثال، پای (۲۰۱۱) در مطالعه DTF مرتبط با جنسیت آزمون مهارت‌های تحصیلی زبان انگلیسی پیرسون[1] نشان داد که به‌طور کلی و فارغ از DIF موجود در پرسش‌های آزمون، این آزمون بدون DTF است. دولیتل و کلییری[2] (۱۹۸۷) در مطالعه DIF آزمون کالج آمریکایی (ACT[3]) نشان دادند که پرسش‌های هندسه و استدلال ریاضی برای جنس مؤنث، دشوارتر و پرسش‌هایی که بیشتر جنبه الگوریتمی و محاسباتی دارند، آسان‌تر می‌نمایند. کارلتون و هریس[4] (۱۹۹۲) در مطالعه DIF آزمون استعداد تحصیلی (SAT[5]) نشان دادند که پرسش‌های بخش شباهت‌ها و تفاوت‌های آزمون که دارای محتوای روابط انسانی، زیبایی‌شناسی یا فلسفی هستند برای زنان در مقایسه با مردان آسان‌تر هستند. لی[6] (۲۰۰۶) در مطالعه DIF برنامه سنجش بین‌المللی دانش‌آموزان (PISA[7]) نشان داد که به‌طور کلی پرسش‌های چندگزینه‌ای به سود پسران است. وی همچنین به این نتیجه رسید که DIF در PISA به حوزه‌های محتوایی آن وابسته است. دودین و انابی[8] (۲۰۰۸) در مطالعه آزمون بین‌المللی علوم و ریاضیات (TIMSS[9]) نشان دادند که به‌طور متوسط حدود ۲۰ درصد پرسش‌های این آزمون در هر دفترچه، دارای DIF مرتبط با جنسیت هستند. آنها همچنین نشان دادند که پرسش‌های DIF در میان دانش‌آموزان دختر و پسر تقریباً به‌طور مساوی تقسیم شده است. آریادوست، گو و کیم[10] (۲۰۱۱) در مطالعه DIF مجموعه آزمون شنیدن زبان

---

1 . Pearson Test of English Academic
2 . Doolittle & Cleary
3 . American College Testing
4 . Carlton & Harris
5 . Scholastic Aptitude Test
6 . Le
7 . Programme for International Student Assessment
8 . Doudeen & Annabi
9 . Trends in International Mathematics and Science Study
10 . Aryadoust, Goh & Kim

انگلیسی میشیگان[1] دریافتند که بسیاری از پرسش‌های آزمون در سطوح پایین توانایی، به نفع آزمودنی‌های مذکر هستند. بریلند، لی، نجاریان و موراکی[2] (۲۰۰۴) در مطالعه DIF آزمون انگلیسی به‌عنوان زبان دوم (TOEFL[3]) نشان دادند که پرسش‌های تکلیف نوشتن آزمون TOEFL به نفع آزمودنی‌های مؤنث عمل می‌کنند. پارک[4] (۲۰۰۸) در پژوهشی DIF پرسش‌های آزمون شنیدن زبان انگلیسی را بررسی کرد. او به این نتیجه رسید که چهار متغیر در ساخت آزمون شامل محتوای متن شنیداری، ارائه عکس در پرسش ارائه شده، نوع زبان به‌کار گرفته شده و فرمت پرسش با DIF جنسیتی در ارتباط هستند. برای مثال، او نشان داد که پرسش‌هایی که محتوای خرید دارند از زنان و پرسش‌هایی که محتوای ورزشی دارند از مردان حمایت می‌کنند. وانگ و لین[5] (۱۹۹۶) DIF مرتبط با جنسیت را در سنجش عملکرد ریاضی بررسی کردند. آنها نشان دادند تکالیفی که در آن آزمودنی با یک مسئله کاربردی مواجه می‌شود، به نفع جنس مؤنث عمل می‌کنند. براون و کانیونگو[6] (۲۰۰۷) وجود DIF را در آزمون‌های ریاضی بزرگ‌مقیاس کشور ترینیداد و توباگو بررسی کردند. آنها نشان دادند که حدود ۱۷ درصد از پرسش‌های آزمون به‌صورت معنی‌داری دارای DIF مرتبط با جنسیت هستند، اما شدت DIF ناچیز بود. بربراگلو[7] (۱۹۹۵) در مطالعه‌ای تقریباً شبیه به مطالعه حاضر در کشور ترکیه، DIF مرتبط با جنسیت را در برخی خرده آزمون‌های آزمون ورودی دانشگاه شامل حساب، لغات و هندسه مطالعه کرد. او نشان داد که تقریباً در همه مقایسه‌ها، پرسش‌های حساب به نفع مردان و پرسش‌های لغات و هندسه به نفع زنان است. این نتیجه به‌وضوح نشان می‌دهد که جنس مؤنث در توانایی کلامی و فضایی و جنس مذکر در مهارت‌های حساب برتری دارند. از سوی دیگر، این نتیجه با یافته‌های

1 . Michigan English Language Assessment Battery listening test
2 . Breland, Lee, Najarian & Muraki
3 . Test of English as a Foreign Language
4 . Park
5 . Wang & Lane
6 . Brown & Kanyongo
7 . Berberoglu

بسیاری از پژوهشگران تناقض دارد (فنما[1]، ۱۹۸۰؛ فنما و کارپنتر[2]، ۱۹۸۱؛ پاتیسون[3] و گریو[4]،۱۹۸٤؛ وود[5]، ۱۹۷٦) که در آنها جنس مذکر در توانایی فضایی برتری داشته، در حالی که جنس مؤنث در مهارت‌های حساب مثل مجموعه‌های جبری برتری خود را نشان می‌دهد. برخی پژوهشگران مثل هوسن[6] (۱۹٦۷)، هانا[7]، ۱۹۸۹ و اتینگتون[8] (۱۹۹۰) معتقدند که یافته‌های متناقض DIF در مورد تفاوت‌های محتوای سوگیرانه نسبت به جنس مؤنث و مذکر ممکن است به‌واسطه اثر کشوری باشد که مطالعه در آن انجام می‌شود. بنابراین با توجه به نتایج متفاوت و در برخی موارد متناقض مطالعات DIF و DTF مرتبط با جنسیت و ضرورت بررسی بیشتر در این حوزه و نقصان چنین مطالعه‌ای در آزمون‌های کنکور سراسری سازمان سنجش آموزش کشور، نتایج این پژوهش می‌تواند برای متخصصان این حوزه بسیار راهگشا بوده و برای تصمیم‌گیران و آزمون سازان داخل کشور بسیار مفید به فایده باشد. با توجه به این موارد و نقش حیاتی که کنکور سراسری در زندگی دانش‌آموزان و ادامه تحصیل آنها در دانشگاه دارد، کمترین مقدار DTF/DIF در آزمون‌های سرنوشت‌ساز می‌تواند پیامدهای ناخواسته گسترده‌ای در جامعه به دنبال داشته باشد (پای و پارک، ۲۰۰٦). بنابراین پرسش اصلی مطالعه حاضر این است که آیا آزمون‌های کنکور سراسری سازمان سنجش آموزش کشور نسبت به گروه‌های جنسیتی دارای کنش افتراقی سؤال و آزمون (DTF/DIF) و به دنبال آن دارای سوگیری هستند؟

روش

پژوهش حاضر از نظر هدف، کاربردی و از حیث روش، از نوع ارزشیابی است. جامعه و نمونه تحقیق شامل همه آزمودنی‌های یک دفترچه آزمون و درس‌های امتحانی تخصصی تمام گروه‌های آزمایشی کنکور سراسری از سال ۱۳۸۷ تا ۱۳۹۰ بود. از میان درس‌ها برای هر گروه آزمایشی درس‌های تخصصی انتخاب شدند. درس‌های انتخاب

1 . Fennema
2 . Carpenter
3 . Pattison
4 . Grieve
5 . Wood
6 . Husen
7 . Hanna
8 . Ethington

شده شامل: درس ریاضی (۵۰ پرسش)، فیزیک (۴۵ پرسش) و شیمی (۳۵ پرسش) از گروه ریاضی و فنی؛ زیست‌شناسی (۵۰ پرسش) و شیمی (۳۵ پرسش) از گروه علوم تجربی؛ ادبیات فارسی (۳۰ پرسش) از گروه علوم انسانی؛ درک عمومی هنر (۳۰ پرسش)، خلاقیت تصویری و تجسمی (۲۰ پرسش) و خلاقیت نمایشی (۲۰ پرسش) از گروه هنر و زبان انگلیسی تخصصی (۷۰ پرسش) از گروه زبان بودند. خلاصه اطلاعات حجم نمونه آزمودنی‌ها بر اساس جنسیت به تفکیک سال‌های برگزاری کنکور سراسری در جدول (۱) ملاحظه می‌شود.

جدول (۱) حجم نمونه مورد مطالعه برای DTF/DIF به تفکیک سال‌های برگزاری کنکور سراسری و جنسیت

سال و رشته		جنسیت	
		مذکر	مؤنث
سال ۸۷	ریاضی	۱۲۶۳۴	۱۰۳۴۵
	تجربی	۱۱۶۷۵	۲۱۶۶۷
	انسانی	۱۳۴۴۵	۲۶۷۷۵
	هنر	۱۱۱۲	۲۹۷۷
	زبان	۳۱۱۲	۷۳۲۴
سال ۸۸	ریاضی	۱۳۴۶۶	۱۰۵۵۳
	تجربی	۹۹۶۳	۲۵۴۴۲
	انسانی	۱۳۶۷۵	۲۴۵۵۷
	هنر	۱۰۲۸	۳۱۱۶
	زبان	۳۱۰۴	۶۲۴۶
سال ۸۹	ریاضی	۱۳۴۶۹	۱۱۲۹۹
	تجربی	۱۱۲۷۴	۲۴۷۴۰
	انسانی	۱۳۸۹۹	۲۳۴۴۶
	هنر	۱۲۸۶	۲۳۸۹
	زبان	۳۸۷۷	۷۸۹۹
سال ۹۰	ریاضی	۱۳۱۶۷۷	۱۰۱۳۵۶
	تجربی	۸۹۸۴۶	۳۱۲۳۵۳
	انسانی	۱۱۴۹۰۰	۱۹۹۵۶۶
	هنر	۱۰۹۹۱	۲۴۷۲۴
	زبان	۲۹۸۵۵	۶۷۵۴۴

برای تشخیص کنش افتراقی پرسش‌های آزمون (DIF) از رویکرد رگرسیون لجستیک دو وجهی[1] و بسته‌های نرم‌افزاری difR در نرم‌افزار R و SPSS استفاده شد. رویه‌های محاسباتی در این رویکرد شامل سه گام در قالب سه معادله بود که در آن معادله‌ها، پاسخ (درست یا غلط) به سؤال متغیر تابع است: ۱- در معادله اول متغیر پیش‌بین (نمره کل آزمون) وارد معادله شد؛ ۲- در معادله دوم نشانگر جنسیت وارد معادله شد و ۳- در معادله سوم، متغیر تعامل دو متغیر وارد شده قبلی وارد معادله رگرسیون لجستیک شد. سپس با استفاده از لگاریتم نسبت درست‌نمایی در آزمونی با df=2، مدل مرحله ۱ و ۲ و مدل مرحله ۲ و ۳ مقایسه شد. این آزمون، دو آزمون دارای df=1 را ترکیب می‌کند. آزمون اول، مدل مرحله ۱ را با مدل مرحله ۲ مقایسه می‌کند که در این آزمون DIF هماهنگ بررسی شد. رد فرض صفر به معنی وجود DIF هماهنگ است. آزمون دوم نیز با استفاده از لگاریتم نسبت درست‌نمایی، مدل مرحله ۲ را با مدل مرحله ۳ مقایسه می‌کند. این مقایسه، DIF ناهماهنگ را ارزیابی می‌کند (گرامی‌پور، ۱۳۹۳). در نرم‌افزار R برای بررسی معنی‌داری DIF، صرفاً مقادیر اختلاف لگاریتم‌های درست‌نمایی، سطح معنی‌داری و اندازه‌های اثر DIF فراهم شد. همچنین برای تشخیص کنش افتراقی آزمون (DTF) از رویکرد مبتنی بر نظریه IRT چالمرز، کانسل و فلورا (۲۰۱۵) و بسته نرم‌افزاری mirt در نرم‌افزار R استفاده شد. به‌علاوه، یکی از پیش‌فرض‌های تحلیل داده‌ها پیش از آشکارسازی DTF/DIF، تک‌بعدی بودن داده‌ها است (فریزر و مک‌دونالد[2]، ۱۹۸۸). برای آزمون تک‌بعدی بودن داده‌ها از نرم‌افزار NOHARM استفاده شد و برای تأیید آن، سه شاخص مجموع مجذورات باقی‌مانده‌ها، ریشه دوم مجذور میانگین باقی‌مانده‌ها و شاخص تاناکا و تاناکا[3] بنا بر پیشنهاد متخصصان (فینچ و هابینگ[4]، ۲۰۰۷) مورد استناد قرار گرفت.

یافته‌ها

نخست، نتایج سه شاخص مورد استناد در NOHARM نشان‌دهنده تک‌بعدی بودن همه آزمون‌های مورد مطالعه بود، بدین ترتیب که: ۱- مجموع مجذورات باقی‌مانده برای همه آزمون‌های مورد مطالعه کوچک‌تر یا مساوی با ۰/۰۰۵ بودند؛ ۲- مجذور

---

1. Binary
2. Fraser & McDonald
3. Tanaka & Tanaka
4. Finch & Habing

میانگین ریشه باقی‌مانده برای همه آزمون‌ها کوچک‌تر مساوی با ۰/۰۱ بودند و ۳- مقدار شاخص تاناکا و تاناکا که یک شاخص بیشینه درست‌نمایی است، برای همه آزمون‌ها بزرگ‌تر مساوی با ۰/۹۵ بود.

در ادامه، یافته‌های رویکرد رگرسیون لجستیک برای آشکارسازی DIF جنسیتی برای یک نمونه پرسش آزمون درس زیست‌شناسی در گروه علوم تجربی سال ۱۳۹۰ ارائه می‌شود.

پرسش ۳، درس زیست‌شناسی از گروه علوم تجربی سال ۱۳۹۰

به‌طور معمول در باکتری‌هایی که کروموزوم‌های کمکی دارند، به تعداد مولکول‌های DNA، ..... وجود دارد.

۱- جایگاه شروع همانندسازی ۲- ژن مقاومت نسبت به آنتی‌بیوتیک ۳-دوراهی همانندسازی ٤- جایگاه تشخیص آنزیم محدودکننده

همان‌طور که در جدول (۲) ملاحظه می‌شود، نتایج تحلیل رگرسیون لجستیک پرسش ۳ آزمون درس زیست‌شناسی در گروه تجربی سال ۹۰، نشان می‌دهد که جنس مؤنث احتمال بیشتری برای دادن پاسخ درست به پرسش ۳ آزمون را دارد (b= ۰/۱۷٦،P<۰/۰۱). همچنین مقدار ثابت معادله رگرسیون لجستیک (α= -۳/۱۳۵،P<۰/۰۱) و نمره کل آزمون زیست‌شناسی (b=۰/۰۸۳،P<۰/۰۱) با بیش از ۹۹ درصد اطمینان معنی‌دار است. نمره کل آزمون از جمع نمره‌های ۵۰ سؤال آزمون، تشکیل شده است و در معادله رگرسیون لجستیک تأثیر آن بر احتمال پاسخ به پرسش کنترل می‌شود. مقدار ثابت نیز در معادله رگرسیون لجستیک نشان‌دهنده برآورد خام احتمال پاسخ به پرسش است.

جدول (۲) مدل مرحله ۲ رگرسیون لجستیک

آماره / منبع	مقدار برآورد	خطای استاندارد برآورد	مقدار والد	درجه آزادی	سطح معنی‌داری	توان نمایی برآورد
جنسیت	۰/۱۷٦	۰/۰۱۱	۲۷۵/۹۳۰	۱	۰/۰۰۰	۱/۱۹۲
نمره کل زیست	۰/۰۸۳	۰/۰۰۱	۲٤۹٦۱/۵۲۸	۱	۰/۰۰۰	۱/۰۸۷
مقدار ثابت	-۳/۱۳۵	۰/۰۱۷	۳۵٦٦۸/٦۰۸	۱	۰/۰۰۰	۰/۰٤٤

معادله مدل مرحله دوم رگرسیون لجستیک به شرح زیر خواهد بود:

مدل کامل لوجیت:)]P(پرسش ۳ آزمون= پاسخ درست=[(۳/۱۳۵- +) + جنسیت ۰/۱۷٦ نمره کل آزمون ۰/۰۸۳

جدول (۳) نشان‌دهنده درصد پیش‌بینی پاسخ‌ها با مدل رگرسیون لجستیک است. هر چقدر درصد پیش‌بینی پاسخ‌ها برای پرسش ۳ آزمون بیشتر باشد مدل رگرسیون لجستیک قدرت بیشتری دارد. جدول (۳) نشان می‌دهد که مدل رگرسیون لجستیک توانسته است حدود ۸۷ درصد پاسخ‌ها برای پرسش ۳ آزمون را پیش‌بینی کند.

جدول (۳) طبقه‌بندی پیش‌بینی‌های درست رگرسیون لجستیک در مدل مرحله ۲

مشاهده شده		پیش‌بینی شده		
		پرسش ۳		درصد درست
		پاسخ غلط	پاسخ درست	
پرسش ۳	پاسخ غلط	۳۳۹۵٦۰	۲۷۱۱	۹۹/۲
	پاسخ درست	٤٦۷۸۳	٤۱۲۱	۸/۱
درصد کل پیش‌بینی درست		۸۷/٤		

همچنین جدول (٤) خلاصه مدل مقدار واریانس پیش‌بینی شده مدل توسط متغیر جنسیت و نمره کل آزمون را نشان می‌دهد.

جدول (٤) خلاصه مدل رگرسیون لجستیک برای پیش‌بینی پاسخ پرسش ۳ آزمون در مرحله ۲

مقدار ۲- برابر لگاریتم بخت	ضریب تبیین کاکس و اسنل	ضریب تبیین ناگلکرک
۲۷۸۰۰۲/۵۵٤	۰/۰٦۲	۰/۱۱۵

نتایج جدول (٤) نشان می‌دهد که حدود ۱۱ درصد احتمال پاسخ برای پرسش ۳ آزمون درس زیست‌شناسی توسط متغیر جنسیت و نمره کل آزمون زیست‌شناسی پیش‌بینی می‌شود. تفاوت ضرایب تبیین ملاحظه شده در جدول (٤) در این است که ضریب تبیین کاکس و اسنل ضریب تبیین نمونه و ضریب ناگلکرک مقدار تبیین شده آن برای جامعه است.

معادله محاسبه شده برای مدل کاهش یافته (مدل مرحله اول) در مثال حاضر عبارت است از:

مدل کاهش یافته: لوجیت(P)] پرسش ۳ آزمون= پاسخ درست=[(۲/۹۰٤- +

نمره کل آزمون ۰/۰۸۳

همان‌طور که ملاحظه می‌شود در مدل کاهش یافته (مدل مرحله ۱) متغیر جنسیت از معادله رگرسیون لجستیک حذف شده و محاسبه دوباره انجام شده است. مقدار (۲-) برابر لگاریتم بخت برای معادله کاهش یافته معادل با ۷۸۲۷٤/۹۷۷ است که اگر مقدار (۲-) برابر لگاریتم بخت مدل کامل از آن کم شود مقدار ۲۷۲/٤۲۳ به دست می‌آید، این مقدار با ۲ درجه آزادی نشان‌دهنده DIF هماهنگ پرسش بررسی شده در سطح معنی‌داری کوچک‌تر از یک‌صدم معنی‌دار است. بنابراین پرسش ۳ درس زیست‌شناسی از گروه علوم تجربی سال ۹۰ نسبت به جنسیت مرد سوگیری آماری دارد.

جدول (۵) مدل مرحله ۳ رگرسیون لجستیک

آماره / منبع	مقدار برآورد	خطای استاندارد برآورد	مقدار والد	درجه آزادی	سطح معنی‌داری	توان نمایی برآورد
جنسیت	۰/۲۱۳	۰/۰۱۸	۱۳۶/۲۵۵	۱	۰/۰۰۰	۱/۲۳۸
نمره کل آزمون زیست	۰/۰۸۷	۰/۰۰۲	۳۰۲۷/۴۷۹	۱	۰/۰۰۰	۱/۰۹۱
جنسیت ×	-۰/۰۰۳	۰/۰۰۱	۶/۴۰۴	۱	۰/۰۱۱	۰/۹۹۷
نمره کل آزمون مقدار ثابت	-۳/۱۸۶	۰/۰۲۶	۱۴۷۷۱/۷۲	۱	۰/۰۰۰	۰/۰۴۱

از سوی دیگر، همان‌طور که در جدول (۵) ملاحظه می‌شود، مدل مرحله سوم با در نظر گرفتن متغیر تعامل، نتایج تحلیل رگرسیون لجستیک نشان می‌دهد که جنسیت احتمال بیشتری برای دادن پاسخ درست به پرسش ۳ آزمون را دارد (b= ۰/۲۱۳،p<۰/۰۱). مقدار ثابت معادله رگرسیون لجستیک (α= -۳/۱۸۶ ،P<۰/۰۱) و نمره کل آزمون (b= ۰/۰۸۷،P<۰/۰۱) با بیش از ۹۹ درصد اطمینان، معنی‌دار است. معادله رگرسیون لجستیک به شرح زیر خواهد بود:

کامل مدل لوجیت:)P(پرسش ۳ آزمون= پاسخ درست =[(۳/۱۸۶- +جنسیت ۰/۲۱۳+نمره کل آزمون ۰/۰۸۷-جنسیت۰/۰۰۳×آزمون کل نمره

برای DIF ناهماهنگ معادله فوق به این معنی است که لوجیت احتمال پاسخ درست به پرسش ۳ آزمون عبارت است از تأثیر تعامل جنسیت و نمره کل آزمون بر لگاریتم بخت پاسخ درست به پرسش ۳ آزمون در حالی که تأثیر جنسیت و نمره کل آزمون بر احتمال پاسخ درست به پرسش کنترل شده است.

جدول (۶) نشان‌دهنده درصد پیش‌بینی پاسخ‌ها با مدل رگرسیون لجستیک است. هر چقدر درصد پیش‌بینی پاسخ‌ها برای پرسش ۳ آزمون بیشتر باشد، مدل رگرسیون لجستیک قدرت بیشتری دارد. جدول (۶) نشان می‌دهد که مدل رگرسیون لجستیک توانسته است حدود ۸۷ درصد پاسخ‌ها برای پرسش ۳ درس زیست‌شناسی را پیش‌بینی کند.

جدول (۶) طبقه‌بندی پیش‌بینی‌های درست مدل رگرسیون لجستیک مرحله سوم

مشاهده شده		پیش‌بینی شده		
		پرسش ۳		درصد درست
		پاسخ غلط	پاسخ درست	
پرسش ۳	پاسخ غلط	۳۳۹۵۶۰	۲۷۱۱	۹۹/۲
	پاسخ درست	۴۶۷۸۳	۴۱۲۱	۸/۱
درصد کل پیش‌بینی درست		۸۷/۴		

همچنین جدول (۷) خلاصه مدل مقدار واریانس پیش‌بینی شده مدل توسط متغیر جنسیت و نمره کل آزمون درس ریاضی و تعامل جنسیت با نمره کل آزمون درس زیست‌شناسی را نشان می‌دهد.

جدول (۷) خلاصه مدل رگرسیون لجستیک مرحله سوم برای پیش‌بینی پاسخ پرسش ۳ درس زیست‌شناسی

مقدار ۲- برابر لگاریتم بخت	ضریب تبیین کاکس و اسنل	ضریب تبیین ناگلکرک
۲۷۷۹۹۹/۱۵۶	۰/۰۶۲	۰/۱۱۵

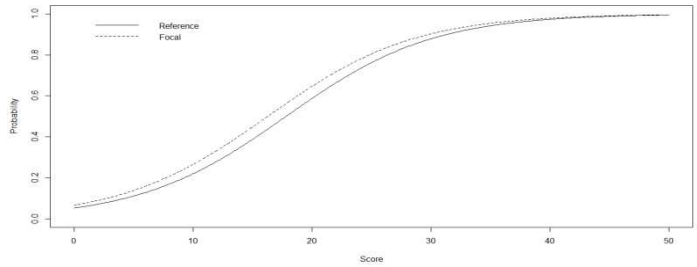
نتایج جدول (۷) نشان می‌دهد که در مدل مرحله سوم نیز حدود ۱۱ درصد احتمال پاسخ برای پرسش ۳ آزمون درس زیست‌شناسی توسط متغیر جنسیت و نمره کل درس زیست‌شناسی و تعامل این دو پیش‌بینی می‌شود.

معادله محاسبه شده برای مدل کاهش یافته DIF ناهماهنگ در پرسش ۳ درس زیست‌شناسی کنکور تجربی سال ۱۳۹۰ حاضر عبارت است از:

مدل کاهش یافته (مرحله ۲): لوجیت)]Pپرسش ۳ آزمون= پاسخ درست=[(۳/۱۳۵-
+جنسیت۰/۱۷۶+نمره کل آزمون۰/۰۸۳

همان‌طور که ملاحظه می‌شود در مدل کاهش یافته DIF ناهماهنگ متغیر تعامل جنسیت با نمره کل آزمون از معادله رگرسیون لجستیک حذف شده و محاسبه دوباره انجام شده است. مقدار (۲-) برابر لگاریتم بخت برای معادله کاهش یافته معادل با ۲۷۸۰۰۲/۵۵٤ است که اگر مقدار (۲-) برابر لگاریتم بخت مدل کامل از آن کم شود مقدار ۳/۳۹۸ به دست می‌آید. این مقدار با ۲ درجه آزادی، مقدار معنی‌داری برای نشان دادن DIF ناهماهنگ پرسش ۳ آزمون نیست.

نمودار (۱) نشان‌دهنده منحنی‌های ویژگی پرسش (ICC[1]) مورد مطالعه به تفکیک جنسیت است.



نمودار (۱) منحنی‌های ویژگی پرسش (ICC) مورد مطالعه به تفکیک جنسیت برای پرسش ۳ آزمون زیست‌شناسی سال ۱۳۹۰

[1] . Item Characteristic Curve

همان‌طور که در نمودار (۱) ملاحظه می‌شود گروه کانونی[1] (جنسیت مؤنث) نسبت به گروه مرجع[2] تا حدود زیادی در سراسر نمره‌های توانایی (به‌جز سطوح بالای توانایی) احتمال بیشتری برای پاسخ درست به پرسش مورد بررسی را دارند. همچنین، اندازه اثر DIF مبتنی بر R^2 ناگلکرک برای این پرسش کمتر از ۰/۰۰۰۱ بود که بسیار ناچیز و قابل اغماض تلقی می‌شود.

سایر نتایج تحلیل DIF مبتنی بر رویکرد تحلیل رگرسیون هیچ‌گونه DIF ناهماهنگ معنی‌داری را بر اساس آماره آزمون نسبت درست‌نمایی در پرسش‌های آزمون‌های کنکور سراسری نشان نداد ($LRT^3<1، P>0/05$)، اما برخی از پرسش‌های آزمون‌های سراسری، بر اساس آماره آزمون نسبت درست‌نمایی، دارای DIF هماهنگ معنی‌داری بودند ($LRT >4، P<0/05$) که نرخ آن در آزمون‌های درس‌های مورد مطالعه برحسب سال‌های برگزاری آزمون (در سطح معنی‌داری ۰/۰۵) در جدول (۸) ملاحظه می‌شود.

جدول (۸) تعداد و درصد پرسش‌های DIF در آزمون‌های درس‌های مورد مطالعه برحسب سال‌های برگزاری آزمون

گروه داوطلبی	درس	سال‌های برگزاری آزمون: ۱۳۸۷ تعداد	۱۳۸۷ درصد	۱۳۸۸ تعداد	۱۳۸۸ درصد	۱۳۸۹ تعداد	۱۳۸۹ درصد	۱۳۹۰ تعداد	۱۳۹۰ درصد	متوسط آزمون تعداد	متوسط آزمون درصد	درصد متوسط در گروه
ریاضی و فیزیک	ریاضی	۹	۱۸/۰۰	۶	۱۲/۰۰	۵	۱۰/۰۰	۸	۱۶/۰۰	۷	۱۴/۰۰	۱۴/۶۹
	فیزیک	۵	۱۱/۱۱	۹	۲۰	۸	۱۷/۷۸	۱۰	۲۲/۲۲	۸	۱۷/۷۸	
	شیمی	۴	۱۱/۴۳	۵	۱۴/۲۹	۴	۱۱/۴۳	۶	۱۷/۱۴	۴/۷۵	۱۳/۵۷	
علوم تجربی	زیست‌شناسی	۶	۱۲	۷	۱۴	۸	۱۶	۸	۱۶	۷/۲۵	۱۴/۵	۱۶/۱۸
	شیمی	۷	۲۰	۶	۱۷/۱۴	۷	۲۰	۵	۱۴/۲۹	۶/۲۵	۱۷/۸۶	
علوم انسانی	ادبیات فارسی	۴	۱۳/۳۳	۴	۱۳/۳۳	۵	۱۶/۶۷	۶	۲۰	۴/۷۵	۱۵/۸۳	۱۵/۸۳
هنر	درک عمومی هنر	۳	۱۰	۰	۰	۰	۰	۶	۲۰	۳/۲۵	۷/۵	۱۲/۹۲

[1] . Focal
[2] . References
[3] . Item Characteristic

سال‌های برگزاری آزمون / گروه داوطلبی/درس		۱۳۸۷		۱۳۸۸		۱۳۸۹		۱۳۹۰		متوسط آزمون		درصد متوسط در گروه
		تعداد	درصد	تعداد	درصد	تعداد	درصد	تعداد	درصد	تعداد	درصد	
	خلاقیت تصویری و تجسمی	۲	۱۰	۰	۰	٤	۲۰	۰	۰	۱/۵	۷/۵	
	خلاقیت نمایشی	٤	۲۰	۳	۱۵	۵	۲۵	۷	۳۵	٤/۷۵	۲۳/۷۵	
زبان	زبان انگلیسی تخصصی	۹	۱۲/۸٦	۱۰	۱٤/۲۹	۸	۱۱/٤۳	۷	۱۰	۸/۵	۱۲/۱٤	۱۲/۱٤
درصد متوسط برحسب سال		۱۲/۷۱		۱۲/۹۰		۱۲/۷٤		۱۸/۸۲		۱٤/۳۲		

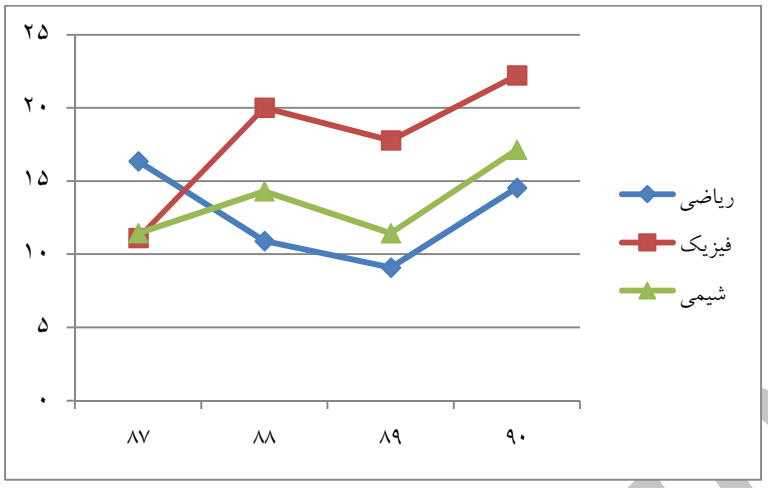
جدول (۹) نیز تعداد و درصد سال‌های DIF علیه جنس مذکر یا مؤنث را نشان می‌دهد.

جدول (۹) تعداد و درصد سال‌های DIF آزمون علیه جنسیت مذکر و مؤنث برحسب سال‌های برگزاری آزمون

سال‌های برگزاری آزمون / گروه داوطلبی/درس/جنسیت			۱۳۸۷		۱۳۸۸		۱۳۸۹		۱۳۹۰		متوسط آزمون	
			تعداد	درصد	تعداد	درصد	تعداد	درصد	تعداد	درصد	تعداد	درصد
ریاضی و فیزیک	ریاضی	مذکر	۰	۰	۲	٤	۰	۰	۳	٦	۱/۲۵	۲/۵۰
		مؤنث	۹	۱۸	٤	۸	۵	۱۰	۵	۱۰	۵/۷۵	۱۱/۵۰
	فیزیک	مذکر	۵	۱۱/۱۱	۰	۰	۰	۰	٦	۱۳/۳۳	۲/۷۵	٦/۱۱
		مؤنث	۰	۰	۹	۲۰	۸	۱۷/۷۸	٤	۸/۸۹	۵/۲۵	۱۱/٦۷
	شیمی	مذکر	٤	۱۱/٤۳	۲	۵/۷۱	٤	۱۱/٤۳	۱	۲/۸٦	۲/۷۵	۷/۸٦
		مؤنث	۰	۰	۳	۸/۵۷	۰	۰	۵	۱٤/۲۹	۲	۵/۷۲
علوم تجربی	زیست‌شناسی	مذکر	۲	٤	٤	۸	٤	۸	۵	۱۰	۳/۷۵	۷/۵۰
		مؤنث	٤	۸	۳	٦	٤	۸	۳	٦	۳/۵۰	۷
	شیمی	مذکر	۷	۲۰	٦	۱۷/۱٤	۵	۱٤/۲۹	۱	۲/۸٦	٤/۷۵	۱۳/۵۷
		مؤنث	۰	۰	۰	۰	۲	۵/۷۱	٤	۱۱/٤۳	۱/۵۰	٤/۲۹
علوم انسانی	ادبیات فارسی	مذکر	٤	۱۳/۳۳	٤	۱۳/۳۳	٤	۱۳/۳۳	۳	۱۰	۳/۷۵	۱۲/۵۰
		مؤنث	۰	۰	۰	۰	۱	۳/۳۳	۳	۱۰	۱	۳/۳۳
هنر	درک عمومی هنر	مذکر	۲	٦/٦۷	۰	۰	۰	۰	۲	٦/٦۷	۱	۳/۳۸
		مؤنث	۱	۳/۳۳	۰	۰	۰	۰	٤	۱۳/۳۳	۱/۲۵	٤/۱۷

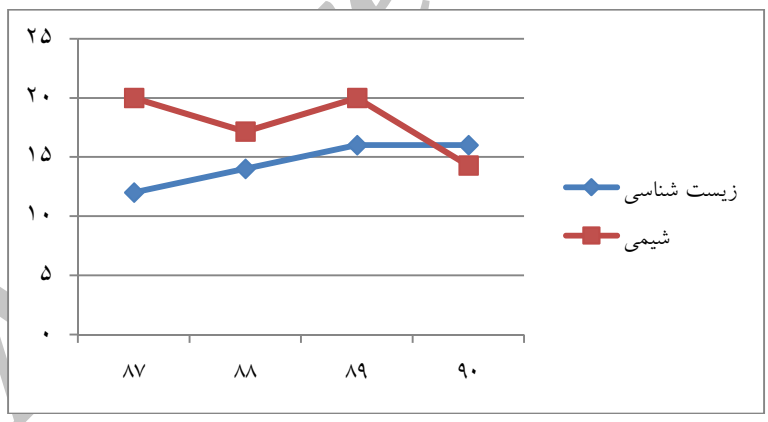
گروه داوطلبی/درس/جنسیت		سال‌های برگزاری آزمون	۱۳۸۷		۱۳۸۸		۱۳۸۹		۱۳۹۰		متوسط آزمون	
			تعداد	درصد	تعداد	درصد	تعداد	درصد	تعداد	درصد	تعداد	درصد
	خلاقیت تصویری و تجسمی	مذکر	۰	۰	۰	۰	۲	۱۰	۰	۰	۰/۵۰	۲/۵۰
		مؤنث	۲	۱۰	۰	۰	۲	۱۰	۰	۰	۱	۵
	خلاقیت نمایشی	مذکر	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰
		مؤنث	٤	۲۰	۳	۱۵	۵	۲۵	۷	۳۵	٤/۷۵	۲۳/۷۵
زبان	زبان انگلیسی تخصصی	مذکر	۹	۱۲/۸٦	٦	۸/۵۷	۳	٤/۲۹	٤	۵/۷۱	۵/۵۰	۷/۸٦
		مؤنث	۰	۰	٤	۵/۷۱	۵	۷/۱٤	۳	٤/۲۹	۳	٤/۲۹

همان‌طور که در جدول‌های (۸ و ۹) ملاحظه می‌شود در گروه ریاضی و فیزیک، نرخ پرسش‌های DIF علیه جنس مؤنث در سال‌های ۸۷ و ۹۰ نسبت به سال‌های ۸۸ و ۸۹ بیشتر است. این در حالی است که نرخ DIF علیه جنس مؤنث در درس ریاضی مشهود است. نرخ پرسش‌های DIF در آزمون فیزیک سال ۸۷ نسبت به سایر سال‌های برگزاری آزمون کمتر است، در حالی که پرسش‌های DIF علیه جنس مذکر کمی بیشتر است. به‌علاوه، آزمون فیزیک سال ۹۰ به نسبت دارای بیشترین درصد پرسش‌های DIF علیه هر دو جنس مؤنث و مذکر است، اما به‌طور کلی نرخ پرسش‌های DIF علیه جنس مؤنث کمی بیشتر است. اما در آزمون شیمی گروه ریاضی و فیزیک، پرسش‌های DIF کمی بیشتر علیه جنس مذکر است. به‌طور متوسط می‌توان مشاهده کرد که آزمون فیزیک در این گروه در مقایسه با سایر درس‌های مورد مطالعه دارای پرسش‌های DIF بیشتری است. روند پرسش‌های DIF در آزمون‌های مورد مطالعه گروه ریاضی و فیزیک در نمودار (۲) ملاحظه می‌شود.



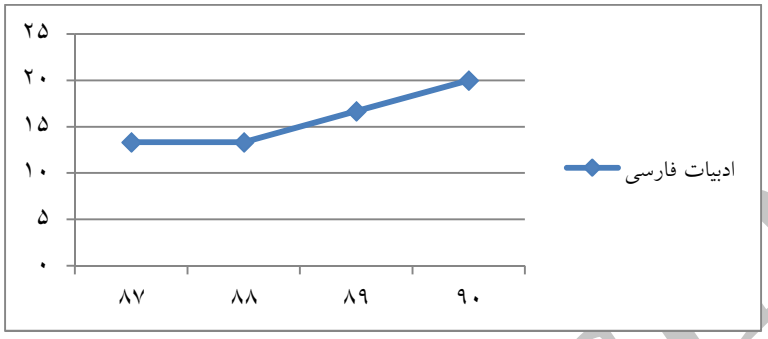
نمودار (۲) درصد پرسش‌های دارای DIF در آزمون‌های مورد مطالعه گروه ریاضی و فیزیک

در گروه علوم تجربی، برخی پرسش‌های آزمون زیست‌شناسی نسبت جنس مؤنث و برخی نسبت به جنس مذکر دارای DIF است. درس شیمی در این گروه نیز شبیه به گروه ریاضی و فیزیک دارد، اما نرخ پرسش‌های DIF در آزمون شیمی به‌طور متوسط از درس زیست‌شناسی بیشتر است. نمودار (۳) نشان‌دهنده نرخ پرسش‌های DIF در درس‌های مورد مطالعه گروه علوم تجربی است.



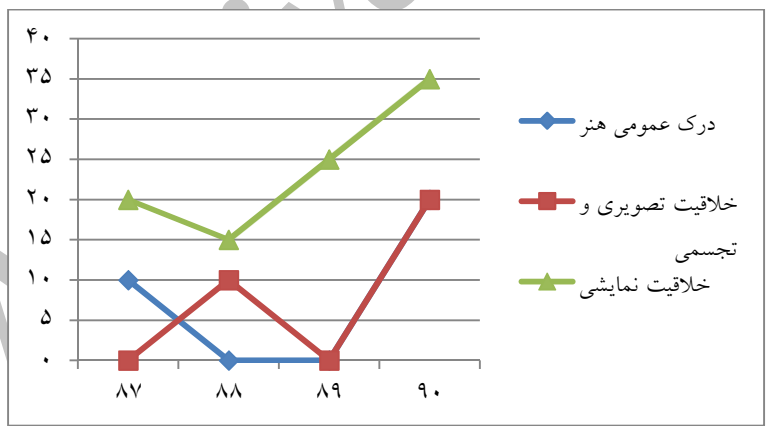
نمودار (۳) درصد پرسش‌های DIF در درس‌های مورد مطالعه گروه علوم تجربی

تحلیل آزمون ادبیات فارسی گروه علوم انسانی نشان می‌دهد که پرسش‌های DIF به نفع جنس مؤنث عمل می‌کنند، در حالی که نرخ آنها از سال ۸۷ تا ۹۰ رو به افزایش است. چنین روندی در نمودار (٤) ملاحظه می‌شود.



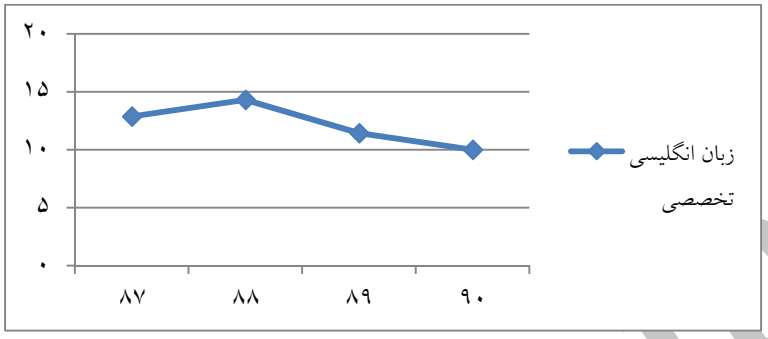
نمودار (٤) درصد پرسش‌های DIF در آزمون ادبیات فارسی گروه علوم انسانی

نمودار (٤) نشان می‌دهد که بیشترین نوسان در گروه هنر است، بدین ترتیب که کمترین نرخ پرسش‌های DIF در درس‌های درک عمومی هنر و خلاقیت تصویری و تجسمی (۷/۵۰ درصد) وجود دارد و در عین حال بیشترین نرخ پرسش‌های DIF (۲۳/۷۵ درصد) نیز در همین گروه و در آزمون درس خلاقیت نمایشی است. در این درس به‌طور مشهودی پرسش‌های DIF علیه جنس مؤنث عمل می‌کنند.



نمودار (۵) نرخ سؤال‌های دارای DIF در آزمون‌های تخصصی گروه هنر

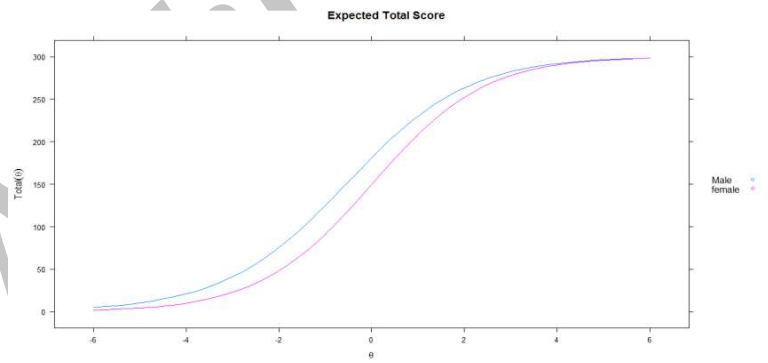
سرانجام نمودار (٦) نشان می‌دهد که آزمون زبان تخصصی گروه زبان دارای کمترین نرخ پرسش‌های DIF است، در حالی که نرخ پرسش‌های DIF از سال ۸۸ تا ۹۰ سیر نزولی داشته است.



نمودار (٦) نرخ پرسش‌های DIF در آزمون زبان تخصصی گروه زبان

همچنین در جدول (۹) می‌توان دید که در درس زبان تخصصی گروه زبان، به‌جز سال ۸۷ که پرسش‌ها علیه جنس مذکر عمل می‌کنند، در سایر سال‌ها برخی پرسش‌های DIF علیه جنس مؤنث و برخی علیه جنس مذکر عمل می‌کنند.

سرانجام، نتایج تحلیل DTF نشان داد که به غیر از آزمون درس خلاقیت نمایشی گروه هنر، سایر آزمون‌های مورد مطالعه فاقد DTF جنسیتی هستند، به‌طوری که مقادیر قدر مطلق sDTF و مقادیر uDTF محاسبه شده برای آنها کمتر از ۰/۰٦ و غیر معنی‌دار بود (P>۰/۰۵). نمودار (۷) نشان‌دهنده DTF جنسیتی برای آزمون خلاقیت نمایشی است.



نمودار (۷) DTF مرتبط با جنسیت برای آزمون درس خلاقیت نمایشی گروه هنر

مقادیر محاسبه شده DTF آزمون خلاقیت نمایشی شامل sDTF=-۳/۱۲ و ۳/۳٤= uDTF است. نمودار (۷) نشان می‌دهد که این آزمون حدود ۳ نمره خام از گروه مرجع (جنس مذکر) طرفداری می‌کند. بنابراین فرض صفر نبود DTF مرتبط با جنسیت در آزمون خلاقیت نمایشی رد شد (P<۰/۰۱). همچنین همان‌طور که در نمودار (۷) دیده می‌شود، در سطوح بسیار بالای توانایی سوگیری آزمون خلاقیت نمایشی نسبت به جنس مؤنث از بین می‌رود و این سوگیری در توانایی پایین‌تر از آن سطح مشاهده می‌شود. البته مقدار اندازه اثر استاندارد شده برای این مقدار DTF معادل با ١٦ درصد محاسبه شده و همچنان در حد «کم» ارزیابی می‌شود.

بحث و نتیجه‌گیری

وجود پرسش‌های DIF در آزمون‌های بزرگ‌مقیاس، اجتناب‌ناپذیر است، اما وجود پرسش‌های DIF لزوماً به معنای سوگیری پرسش‌های آزمون نیست (لی، ٢٠٠٦). با این وجود، تحلیل DIF پرسش‌های آزمون سرنوشت‌سازی مانند کنکور سراسری بر اساس جنسیت (به‌عنوان اساسی‌ترین متغیر گروهی) بسیار ضروری است؛ زیرا وجود DIF مرتبط با جنسیت به‌صورت منفی بر منصفانه بودن آزمون برای دختران و پسران تأثیر می‌گذارد و دقت نمره‌ها و کاربرد آزمون برای گزینش مبتنی بر جنسیت را کاهش می‌دهد.

یافته‌های این پژوهش نشان می‌دهد که به‌طور کلی حدود ١٤ درصد پرسش‌های آزمون‌های تخصصی مورد مطالعه در کنکور سراسری سال‌های ۸۷ تا ۹۰ دارای DIF هستند. البته شدت DIF محاسبه شده برای این پرسش‌ها بسیار ناچیز هستند. نرخ کلی آشکارسازی پرسش‌های DIF در آزمون تقریباً با یافته‌های دودین و انابی (۲۰۰۸) و براون و کانیونگو (۲۰۰۷) همخوانی دارد. همچنین به غیر از آزمون درس خلاقیت نمایشی گروه هنر، سایر آزمون‌های مورد مطالعه بدون DTF جنسیتی بودند، در حالی که شدت DTF برای آزمون این درس نیز «کم» ارزیابی شد. به‌طور کلی، این نرخ و به‌ویژه شدت DIF و DTF می‌تواند تا حدود زیادی نشانگر نبود سوگیری جنسیتی در آزمون‌های کنکور سراسری باشد.

همان‌طور که اشاره شد DTF/ DIF مرتبط با جنسیت لزوماً به معنای سوگیری جنسیتی نیست، زیرا برای سوگیری پرسش و آزمون وجود DTF/ DIF لازم است

ولی کافی نیست (زومبو[1]، ۱۹۹۹). بنابراین برای قضاوت نهایی در مورد بود یا نبود سوگیری مرتبط با جنسیت، کمیته‌های تخصصی موضوعی تشکیل شد. این کمیته‌ها شامل ٦ تا ۱۲ تن از متخصصان و استادان درس‌های کنکور (بسته به آزمون درس تخصصی) شامل متخصصان زن و مرد بودند که با رهبری یک متخصص سنجش و اندازه‌گیری هدایت می‌شدند. معیارهای اصلی قضاوت در مورد سوگیری جنسیتی پرسش‌های آزمون در درجه اول یافته‌های مربوط به DIF و اندازه اثر DIF پرسش‌های آزمون و سپس محتوای پرسش‌های آزمون بود که از نظر میزان تطابق با تجربیات زیسته دختران و پسران یا زنان و مردان از سوی متخصصان موضوع درسی مورد بررسی و مداقه قرار گرفت و سرانجام پیشینه پژوهشی به‌عنوان یکی از ملاک‌های اصلی برای قضاوت در مورد سوگیری پرسش‌های آزمون مد نظر گروه متخصصان قرار گرفت. به‌طور مثال در بررسی DIF و DTF آزمون درس خلاقیت نمایشی گروه هنر در کنکور سراسری سال ۱۳۹۰، پرسش ۲۰۳ دفترچه گروه هنر که دارای DIF آماری در درس مذکور بود، به‌صورت زیر مطرح شده است:

۲۰۳- این فیلم مستند داستان ندارد، گزارش روزانه زندگی یک پسر بچه است.

۱) سیاه و سفید ۲) خاک ۳) یک اتفاق ساده ٤) طبیعت بی‌جان

که بر اساس قضاوت‌های کیفی متخصصان موضوع درسی مذکور (اعم از متخصصان زن یا مرد) دارای سوگیری علیه جنس مؤنث بود. نتایج مرور و تشخیص نرخ سوگیری پرسش‌های DIF در درس‌های تخصصی بنا به قضاوت کمیته‌های موضوع درسی در جدول (۱۰) ملاحظه می‌شوند.

جدول (۱۰) نرخ سوگیری پرسش‌های DIF در درس‌های تخصصی بنا به قضاوت کمیته‌های متخصص موضوع درسی

سال‌های برگزاری آزمون / گروه داوطلبی/درس		۱۳۸۷		۱۳۸۸		۱۳۸۹		۱۳۹۰		متوسط آزمون		درصد متوسط در گروه
		تعداد	درصد	تعداد	درصد	تعداد	درصد	تعداد	درصد	تعداد	درصد	
ریاضی و فیزیک	ریاضی	۳	۵/٤۵	۱	۱/۸۲	۰	۰	۲	۳/٦٤	۱/۵	۲/۷۳	۲/٦۸
	فیزیک	۱	۲/۲۲	۲	٤/٤٤	۱	۲/۲۲	۳	٦/٦۷	۱/۷۵	۳/۸۹	
	شیمی	۰	۰	۱	۲/۸٦	۰	۰	۱	۲/۸٦	۰/۵	۱/٤۳	
	زیست‌شناسی	۱	۲	۲	٤	۱	۲	۱	۲	۱/۲۵	۲/۵۰	۲/٦۸

[1]. Zumbo

سال‌های برگزاری آزمون / گروه داوطلبی/درس		۱۳۸۷		۱۳۸۸		۱۳۸۹		۱۳۹۰		متوسط آزمون		درصد متوسط در گروه
		تعداد	درصد	تعداد	درصد	تعداد	درصد	تعداد	درصد	تعداد	درصد	
علوم تجربی	شیمی	۱	۲/۸۶	۱	۲/۸۶	۲	۵/۷۱	۰	۰	۱	۲/۸۶	
علوم انسانی	ادبیات فارسی	۰	۰	۰	۰	۱	۳/۳۳	۱	۳/۳۳	۰/۵	۱/۶۷	۱/۶۷
هنر	درک عمومی هنر	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	
	خلاقیت تصویری و تجسمی	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۱/۶۷
	خلاقیت نمایشی	۱	۵	۰	۰	۱	۵	۲	۱۰	۱	۵	
زبان	زبان انگلیسی تخصصی	۲	۲/۸۶	۲	۲/۸۶	۱	۱/۴۳	۱	۱/۴۳	۱/۵	۲/۱۴	۲/۱۴
درصد متوسط برحسب سال		۲/۰۴		۱/۸۸		۱/۹۷		۲/۹۹		۲/۲۲		

همان‌طور که در جدول (۱۰) ملاحظه می‌شود، بر اساس قضاوت کمیته‌های تخصصی به‌طور متوسط حدود ۲ درصد پرسش‌های DIF در آزمون‌های مورد مطالعه کنکور سراسری دارای سوگیری که با توجه به اندازه‌های اثر بسیار کوچک و ناچیزی که به دست آمد، قابل توجیه است. در گروه ریاضی و فیزیک، پرسش‌های سودار بیشتر به نفع جنس مذکر عمل می‌کنند. به نظر می‌رسد پرسش‌هایی که برای محاسبه پاسخ نهایی به مراحل بیشتری نیاز دارند، به نفع جنس مذکر عمل می‌کنند. این نتیجه با یافته‌های دولیتل و کلییری (۱۹۸۷)، انابی و دودین (۲۰۰۶) و کارلتون و هریس (۱۹۹۲) همخوانی دارد، آنها نیز به این نتیجه رسیدند که پرسش‌هایی که در پاسخ به آن راهبردهای متعددی مورد نیاز است، به ضرر دانش آموزان دختر عمل می‌کند. آنها همچنین به این نتیجه رسیدند که پرسش‌های دارای نمودار یا جدول نیز به نفع دانش آموزان مذکر عمل می‌کنند. همچنین نرخ پرسش‌های سودار تشخیص داده شده در آزمون درس شیمی نسبت به سایر درس‌های انتخاب شده از گروه ریاضی و فیزیک کمتر است، به‌طوری که تنها ۲ پرسش در همه آزمون‌های برگزار شده، به دلایل محتوایی

فوق‌الذکر، نسبت به جنس مؤنث سوگیری داشتند. این در حالی بود که نرخ پرسش‌های DIF آماری علیه جنس مذکر در درس شیمی بیشتر بود.

همچنین نرخ پرسش‌های سودار تشخیص داده شده در جدول (۱۰) نمایانگر این است که پرسش‌های سودار درس‌های منتخب گروه‌های علوم تجربی و علوم انسانی بیشتر به نفع جنس مؤنث عمل می‌کنند. به نظر می‌رسد که پاسخگویی به این پرسش‌ها وابستگی بیشتری به حافظه و یادآوری دارند. این نتیجه تا حدودی با یافته‌های انابی (۲۰۰۸) همخوانی دارد. همچنین این یافته با پژوهش‌هایی همخوانی دارد که نشان می‌دهند دانش‌آموزان مؤنث در پرسش‌هایی که یادآوری مفاهیم تدریس شده، اجتناب از تفکر چالشی و ترجمه مفاهیم را می‌سنجند، به‌صورت سوگیرانه‌ای عملکرد بهتری دارند (گالاکر[1]، ۱۹۹۸؛ گالاکر و دلیسی[2]، ۱۹۹٤ و هریس و کارلتون[3]، ۱۹۹۳).

همان‌طور که در جدول (۱۰) ملاحظه می‌شود، بر اساس قضاوت کمیته متخصصان، در آزمون‌های درس‌های درک عمومی هنر و خلاقیت تصویری و تجسمی گروه هنر هیچ پرسش سوداری یافت نشد. اما در آزمون درس خلاقیت نمایشی، پرسش‌های سودار بر علیه جنس مؤنث عمل می‌کردند. بر اساس نظرات کیفی کمیته خبرگان این گروه، دلیل احتمالی DIF و DTF در این آزمون می‌تواند این باشد که محتوای پرسش‌های سودار این درس و به‌طور کلی این آزمون، بیشتر با تجارب جنس مذکر متناسب است، بدین معنی که بافت پرسش برای جنس مذکر، ملموس‌تر بوده و با تجربیات واقعی زندگی آنها بیشتر قابل پیوند است.

سرانجام، یافته‌های تحلیلی کمیته متخصصان درس زبان انگلیسی، نشان‌دهنده آن بود که پرسش‌های سوداری که محتوای علوم اجتماعی دارند بیشتر به نفع جنس مؤنث و پرسش‌های سوداری که محتوای علوم سخت دارند بیشتر به نفع جنس مذکر عمل می‌کنند. این نتیجه با یافته‌های امیریان، علوی و فیدالگو[4] (۲۰۱٤) همخوانی دارد. همچنین این نتیجه با یافته‌های کارلتون و هریس (۱۹۹۲) و اونیل و مک‌پیک[5] (۱۹۹۳) از این جهت همخوانی دارد که آنها نیز به این نتیجه رسیدند که پرسش‌های دارای مفاهیم انتزاعی به نفع جنس مؤنث عمل می‌کنند.

---

1. Gallagher
2. Gallagher & DeLisi
3. Harries & Carlton
4. Amirian, Alavi & Fidalgo
5. O'Neill & McPeek

یافته‌های مطالعه حاضر دلالت‌های کاربردی معینی در مورد DIF و DTF مرتبط با جنسیت در کنکور سراسری سازمان سنجش آموزش کشور دارد. از این مطالعه درباره تأثیر جنسیت بر عملکرد آزمودنی‌ها در کنکور سراسری اطلاعات سودمندی به دست آمد و ویژگی‌های محتوایی پرسش‌هایی که ممکن است نسبت به جنس مؤنث یا مذکر غیرمنصفانه (حتی به میزان خیلی کم و ناچیز) عمل کنند، مشخص شد. بنابراین تعیین نرخ پرسش‌های DIF و آزمون‌های DTF و میزان سوگیری آنها و تعیین ویژگی‌های محتوایی پرسش‌های سودار، ازجمله هدف‌هایی بوده است که در پژوهش حاضر، محقق شد.

برای پژوهش‌های آینده پیشنهاد می‌شود که بدون در نظر گرفتن اطلاعات آماری مربوط به DIF مرتبط با جنسیت، پرسش‌های کنکور صرفاً به لحاظ محتوایی و به‌صورت کیفی توسط گروه خبرگان بررسی و نتایج حاصل با یافته‌های این پژوهش، مقایسه شوند. انگلهارد، هانش و روتلج[1] (۱۹۹۰) چنین رویکردی را برای مطالعات DIF پیشنهاد کرده است. همچنین یافته‌های پژوهش نشان داد که تنها آزمون خلاقیت نمایشی که دارای بیشترین نرخ پرسش‌های DIF است، دارای DTF معنی‌داری بود. اگرچه چندین مطالعه تجربی در زمینه رابطه DIF وDTF آزمون انجام گرفته است (تاکالا و کافتاندیجوا[2]، ۲۰۰۰؛ زومبو، ۲۰۰۳)، اما هنوز مشخص نیست که ۱- وجود DTF در آزمون ناشی از تأثیر تراکمی پرسش‌های DIF آزمون است، کما اینکه چنین پدیده‌ای در مطالعه حاضر به‌صورت موردی مشاهده شد، ولی هنوز نمی‌توان آن را تعمیم داد و یا ۲- آزمون دارای پرسش‌های DIF به دلیل لغو DIF در سطح تحلیل آزمون، DTF را نشان نمی‌دهند و یا اینکه ۳- آزمونی با پرسش‌های DIF، DTF را نشان نمی‌دهد، زیرا DIF از DTF مستقل است. بنابراین پیشنهاد می‌شود که رابطه میان وجود پرسش‌های DIF در آزمون و کنش افتراقی آزمون (DTF) داده‌های شبیه‌سازی شده و داده‌های تجربی کنکور سراسری با حجم نمونه‌های بیشتر، به‌منظور کاربرد در تحلیل داده‌های کنکور سراسری، بررسی و مطالعه شود.

---

1. Engelhard, Hansche & Rutledge
2. Takala & Kaftandjieva

منابع

آبکار، کبری (۱۳۹۱). *بررسی ویژگی‌های روان‌سنجی سؤالات کنکور سراسری در رشته علوم تجربی سال ۱۳۸۹ از نظر تئوری سؤال و پاسخ (IRT)*. پایان‌نامه کارشناسی ارشد، دانشگاه آزاد اسلامی واحد تهران مرکز.

آسیابی، مینا (۱۳۹۱). *ارزشیابی آزمون کارشناسی ارشد رشته جغرافیای سیاسی با استفاده از مدل‌های جدید اندازه‌گیری و تعیین سوگیری جنسیتی در آن*. پایان‌نامه کارشناسی ارشد، دانشگاه علامه طباطبایی.

ترکاشوند، علی (۱۳۹٤). *بررسی ویژگی‌های روان‌سنجی آزمون سراسری درس زیست‌شناسی بر اساس مدل چندگزینه‌ای IRT*. پایان‌نامه کارشناسی ارشد، دانشگاه خوارزمی.

حبیبی، مجتبی (۱۳۹۲). *بررسی عوامل مؤثر بر پیشرفت تحصیلی دانشجویان مقطع کارشناسی و پیش‌بینی آن بر اساس نمرات تراز کنکور: اعتباریابی بیرونی نمرات تراز کنکور با مطالعه موردی دانشگاه شهید بهشتی*. طرح پژوهشی، وزارت علوم، تحقیقات و فناوری.

فلاحی‌سرشت، شیوا (۱۳۹٤). *بررسی کارکرد افتراقی سؤالات (DIF) استعداد تحصیلی آزمون نیمه‌متمرکز دکتری سال ۹۳ با کاربرد نظریه سؤال-پاسخ (IRT) و رگرسیون لجستیک*. پایان‌نامه کارشناسی ارشد، دانشگاه علامه طباطبایی.

گرامی‌پور، مسعود (۱۳۹۳). *مبانی نظری و کاربرد نظریه‌های اندازه‌گیری در علوم رفتاری*. تهران: انتشارات تمدن علمی.

گرامی‌پور، مسعود و فلسفی‌نژاد، محمدرضا (۱۳۹۲). *روش‌های آماری بررسی کنش افتراقی سؤال (DIF) در آزمون‌های سرنوشت‌ساز*. تهران: انتشارات جهاد دانشگاهی، واحد تربیت معلم.

معلمی اوره، مهرناز (۱۳۸۷). *مقایسه دقت برآورد توانایی در سؤالات چندگزینه‌ای با به‌کارگیری مدل‌های سؤال- پاسخ دو و چندارزشی*. پایان‌نامه کارشناسی ارشد، دانشگاه علامه طباطبایی.

میری، محمد (۱۳۹٤). *بررسی و مقایسه ویژگی‌های روان‌سنجی بخش فیزیک آزمون سراسری ورود به دانشگاه بر اساس مدل‌های دو ارزشی IRT*. پایان‌نامه کارشناسی ارشد، دانشگاه خوارزمی.

مینایی، 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. *فصلنامه اندازه‌گیری تربیتی*، ٤ (۱۱)، ۱۰۹-۱٤٦.

نژادنجف، فیروز (۱۳۹۳). نقد و بررسی سؤالات کنکور سراسری درس دین و زندگی. *رشد آموزش معارف اسلامی*، ٢٦، ٤٨-٥٣.

Amirian, S. M.R.; Alavi, S. M. & Fidalgo, A. M. (2014). Analyzing Gender Differences with an English Proficiency Test in EFL Context. *Iranian Journal of Language Testing.*

Aryadoust, V.; Goh, C. C. M. & Kim, L. O. (2011). An investigation of differential item functioning in the MELAB listening test. *Language Assessment Quarterly*, 8 (4), 361– 385.

Barati, H. & Ahmadi, A. R. (2010). Gender-based DIF across the subject area: A study of the Iranian National University Entrance Exam. *The Journal of Teaching Language Skills (JTLS)*, 2 (3), 1-22.

Berberoglu, G. (1995). Differential item functioning (DIF) analysis of computation, word problem and geometry questions across gender and SES groups. *Studies in Educational Evaluation,* 21 (4), 439-456.

Breland, H.; Lee, Y. W.; Najarian, M. & Muraki, E. (2004). *An analysis of the TOEFL CBT writing prompt difficulty and comparability of different gender groups* (TOEFL Research Report No. 76). Princeton, NJ: Educational Testing Service.

Brown, I. & Kanyongo, Y. (2007). Differential Item Functioning and male-female differences in a large-scale mathematics assessment in Trinidad and Tobago. *Caribbean Curriculum*, 14, 49–71.

Carlton, S. T. & Harris, A. M. (1992). *Characteristics associated with differential item functioning on the Scholastic Aptitude Test: Gender and majority/minority group comparisons*. Princeton, NJ: Educational Testing Service.

Chalmers, R. P.; Counsell, A. & Flora, D. B. (2015). It might not make a big DIF: Improved.

Differential Test Functioning statistics that account for sampling variability. *Educational and Psychological Measurement*, 1-27.

Doolittle, A. E. & Cleary, T. A. (1987). Gender-based differential item performance in mathematics achievement items. *Journal of Educational Measurement*, 24, 157-166.

Doudeen, Hamzah M. & Annabi, Hanan A. (2008). Sex-Related Differential Item Functioning (DIF) Analysis of TIMSS. *Dirasat, Educational Sciences*, Vol. 35.

Drasgow, F. (1984). Scrutinizing psychological tests: Measurement equivalence and equivalent relations with external variables are central issues. *Psychological Bulletin*, 95, 134-135.
Drasgow, F. (1987). Study of the measurement bias of two standardized psychological tests. *Journal of Applied Psychology*, 72, 19-29.
Embretson, S. E. & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates.
Engelhard, G.; Hansche, L. & Rutledge, K. (1990). Accuracy of bias review judges in identifying differential item functioning on teacher certification tests. *Applied Measurement in Education*, 3, 347–360.
Ethington. A. (1990). Gender differences in mathematics: An international perspective. *Journal for Research in Mathematics Education*. 21 (1), 74-80.
Fennema. E (1980). Sex-related differences in mathematics achievement: Where and why. In Fox, L. H.; Brody, L. & Tobin, D. (Eds.). *Women and the mathematic mystique*, (pp. 76-93). Baltimore: Johns Hopkins University Press.
Fennema. E. & Carpenter. T. P. (1981). Sex-related differences in mathematics: Results from national assessment. *Mathematics Teacher*. 74, 554-559.
Finch, H. & Habing, B. (2007). Performance of DIMTEST- and NOHARM based statistics for testing unidimensionality. *Applied Psychological Measurement*, 31, 292–307.
Flora, D., Curran, P., Hussong, A., & Edwards, M. (2008). Incorporating measurement Nonequivalence in a cross-study latent growth curve analysis. *Structural Equation Modeling*, 15, 676-704.
Fraser, C., & McDonald, R. P. (1988). NOHARM: Least squares item factor analysis. *Multivariate Behavioral Research*, 23, 267-269.
Gallagher, A. (1998). Gender and antecedents of performance in mathematics testing. *Teachers College Record*, 100 (2), 297-314.
Gallagher, A. M., & DeLisi, R. (1994). Gender differences in scholastic aptitude tests mathematics problem solving among high-ability students. *Journal of Educational Psychology*, 86, 204-211.
Hanna. G. (1989). Mathematics achievement of girls and boys in grade eight: Results from twenty countries. *Educational Studies in Mathematics*, 20, 225-232.
Harries, A. & Carlton, S. (1993). Patterns of gender difference on mathematics items on the scholastic aptitude test. *Applied Measurement in Education*, 6 (2), 151- 173.
Husen, T. (1967). *International study of achievement in mathematics: A comparison of twelve countries*. Vol. 11. Stockholm: Almqvist & Wiksell.

Innabi, H., & Dodeen, H. (2006). Content Analysis of Gender-related Differential Item Functioning of TIMSS Items in Mathematics in Jordan. *School Science and Mathematics,* 106 (8), 328-337.

Le, Luc T. (2006). Investigating gender differential item functioning across Countries and Test Languages for PISA science items. *International Journal of Testing*, 9, 2, 122-133.

O'Neill, K. A. & McPeek, W. M. (1993). Item and test characteristics that are associated with differential item functioning. In Holland, P. W. & Wainer, H. (Eds.), *Differential item functioning*, (pp. 255-276). Hillsdale, N J: Lawrence Earlbaum.

Pae, H. K. (2011). Differential item functioning and unidimensionality in the Pearson Test of English Academic. http://pearsonpte.com/research/Documents/Pae.pdf.

Pae, T. & Park, G. P. (2006). Examining the relationship between differential item functioning and differential test functioning. *Language testing*, 23 (4), 475-496.

Park, G. P. (2008). Differential item functioning on an English listening test across gender. *TESOL Quarterly,* 42 (1), 115-123.

Pattison. P. & Grieve, N. (1984). Do spatial skills contribute to sex differences in different types of mathematical problems? *Journal of Educational Psychology,* 76 (4). 677-689.

Raju, N. S.; van der Linden, W. J. & Fleer, P. F. (1995). IRT-based internal measures of differential functioning of items and tests. *Applied Psychological Measurement,* 19, 353–368.

Rudner, L.; Getson, P. & Knight, D. (1980). Biased item detection techniques. *Journal of Educational Statistics*, 5, 213-233.

Russell, S. S. (2005). *Estimates of Type I error and power for indices of differential bundle and test functioning*. Ph.D. dissertation, Bowling Green State University, United States -- Ohio.

Takala, S. & Kaftandjieva, F. (2000). Test fairness: A DIF analysis of an L2 vocabulary test. *Language Testing,* 17, 323–340.

Wang, N. & Lane, S. (1996). Detection of gender-related differential item functioning in a mathematics performance assessment. *Applied Measurement in Education,* 9 (2), 175–199.

Wood, R. (1976). Sex differences in mathematics attainment at GCE ordinary level. *Educational Studies,* 2. 141- 160.

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.

Zumbo, B. (2003). Does item-level DIF manifest itself in scale-level analysis? Implications for translating language tests. *Language Testing*, 20, 136–147.