فصلنامه مطالعات اندازه‌گیری و ارزشیابی آموزشی
سال اول، شمارهٔ ۲، تابستان ۱۳۹۱، صص ۷۶- ۵۳

Journal of Measurement & Educational Evaluation Studies
Vol. 1, No. 2, Summer 2012

ارزیابی کارکرد افتراقی جنسیتی سئوال‌های آزمون ریاضی با استفاده از دو روش مانتل– هنزل و نظریه سئوال- پاسخ

نگار شریفی یگانه *

تاریخ دریافت مقاله: ۹۱/۰۲/۰۵

تاریخ پذیرش مقاله: ۹۱/۰۶/۲۲

چکیده

آزمون‌ها می‌بایست برای تمام افراد جامعه از هر جنس، نژاد، سن، موقعیت اجتماعی و اقتصادی عادلانه باشد. بر این اساس ارزیابی وجود سوگیری و کارکرد افتراقی در سئوال‌های آزمون‌ها بسیار حائز اهمیت است. در مطالعه حاضر کارکرد افتراقی جنسیتی با استفاده از روش مانتل – هنزل و روش مبتنی بر نظریه سئوال- پاسخ مورد بررسی قرار گرفته است. در ابتدا مبانی نظری و روش‌های ارزیابی کارکرد افتراقی معرفی شده است و در ادامه به منظور ارائه نمونه عملی، پاسخ‌های یک نمونه تصافی طبقه‌ای۴۰۰۰ نفری مشتمل بر ۲۲۰۰ آزمودنی مرد و ۱۸۰۰ زن به ۵۵ سئوال آزمون ریاضی کنکور سراسری مورد تحلیل قرار گرفته است. نتایج تحلیل بیانگر آن است که سئوال‌های آزمون دارای کارکرد افتراقی است. در بررسی کارکرد افتراقی با استفاده از روش مانتل هنزل ۲۳ سئوال دارای شاخص مانتل هنزل معنادار بودند. بر اساس رویکرد سئوال – پاسخ نیز ۹ سئوال دارای کارکرد افتراقی جنسیتی بودند که همگی به نفع دختران بودند. سئوال‌های دارای کارکرد افتراقی به نفع دختران بیشتر در حوزه محتوایی توابع و معادلات و سئوال‌های دارای کارکرد افتراقی به نفع پسران بیشتر در حوزه محتوایی مثلثات، هندسه و احتمال هستند.

واژگان کلیدی: نظریه سئوال – پاسخ، سوگیری سئوال، کارکرد افتراقی سئوال، روش مانتل – هنزل

* دانشجوی دکتری سنجش و اندازه‌گیری دانشگاه علامه طباطبایی و کارشناس سازمان سنجش آموزش کشور (مسئول مکاتبات: nsh-yeganeh@yahoo.com)

مقدمه

روش‌های مختلف سنجش به ویژه آزمون‌ها نقش مهمی در سرنوشت افراد ایفا می‌کنند، از این رو، عادلانه[1] و منصفانه بودن شیوه‌های سنجش اهمیت بسیاری دارد. در واقع روش‌های سنجش از جمله آزمون‌ها می‌بایست برای تمام افراد جامعه از هرجنس، نژاد، سن، موقعیت اجتماعی و اقتصادی منصفانه باشد و آزمودنی‌های متعلق به گروه‌های مختلف با توانایی‌های مشابه، امکان توفیق یکسانی در آزمون داشته باشند. هدف از اجرای آزمون‌ها، تفکیک و تمایز افراد است، اما آزمون‌ها نبایستی منجر به تبعیض میان آزمودنی‌ها شود (ناسستروم[2]، ۲۰۰۳).

در سال‌های اخیر، با توجه به افزایش کاربرد آزمون‌ها در موقعیت‌های مختلف، نگرانی فزاینده‌ای در باره امکان عملکرد متفاوت سئوال‌های آزمون و به عبارت دیگر سوگیری سئوال‌های[3] آزمون در زیرگروه‌های آزمودنی (به طور مثال زنان و مردان) ایجاد شده است. به دلیل تأثیر سوگیری سئوال‌ها در انحراف نتایج آزمون و بالتبع تصمیم‌های اتخاذ شده بر مبنای این آزمون‌ها، دامنه این نگرانی حتی به سیاست‌گذاران آموزشی هم منتقل شده است. به اعتقاد منتقدان آزمون‌ها، سوگیری سئوال‌های آزمون ممکن است منجر به از دست رفتن فرصت‌های آموزشی و شغلی شود (کونولی[4]، ۲۰۰۳).

سوگیری زمانی مطرح می‌شود که احتمال پاسخ‌گویی درست گروهی از آزمودنی‌ها به دلیل مشخصات سئوال‌ها و شرایط آزمون، از گروه دیگری از آزمودنی‌ها کم‌تر یا بیشتر باشد (زومبو[5]، ۱۹۹۹). در واقع سئوال زمانی سوگیری دارد که ویژگی‌های غیرمرتبط با سازه مورد سنجش را بسنجد. سوگیری سئوال ممکن است به علت ابهام در متن، گزینه‌ها و راهنمای آزمون باشد (همبلتون[6] و راجرز[7]، ۱۹۹۵). کارکرد افتراقی سئوال‌ها[8] (DIF) روش آماری تعیین وجود سوگیری در سئوال‌های آزمون است. این اصطلاح توسط هالند[9] و واینر[10] (۱۹۹۳) ارائه شده است (اینگ[11] و هوی[12]، ۲۰۰۵).

---

1. Fair
2. Näsström
3. Item Bias
4. Conoley
5. Zumbo
6. Hambleton
7. Rodgers
8. Differential Item Function
9. Holland
10. Wainer
11. Eng
12. Hoe

کارکرد افتراقی سئوال، زمانی مطرح می‌شود که آزمودنی‌ها با توانایی یکسان، متعلق به گروه‌های مختلف، احتمال پاسخ‌گویی درست اما متفاوتی داشته باشند. به عبارت دیگر سئوال در صورتی دارای کارکرد افتراقی است که در زیر گروه‌های مختلف جامعه عملکرد متفاوت داشته باشد. وجود کارکرد افتراقی بیانگر آن است که عوامل مربوط به عضویت در گروه، احتمال پاسخ‌گویی درست را تحت تأثیر قرار می‌دهد. در صورتی که سئوال، کارکرد افتراقی نداشته باشد، فاقد سوگیری نیز هست، ولی با وجود کارکرد افتراقی، وجود سوگیری سئوال نیز، باید به روش‌های مختلفی چون تحلیل محتوا و ارزشیابی تجربی مورد بررسی قرار گیرد (دانکن، ۲۰۰۶).

تاریخچه و رویکردهای ارزیابی کارکرد افتراقی سئوال

چالش اساسی در خصوص آزمون عادلانه، نگرانی در خصوص سوگیری سئوال‌های آزمون و ضرورت بررسی آن به عنوان بخشی از فرایند آزمون‌سازی، در اواخر دهه ۱۹۶۰ و اوایل دهه ۱۹۷۰ در راستای انتظارات مبنی بر یکسانی نتایج آزمون‌ها و هم‌زمان با آغاز عصر حقوق مدنی مطرح شد (دانکن، ۲۰۰۶). در چهل سال اخیر، نگرانی‌های مربوط به عادلانه بودن آزمون‌ها، منجر به ایجاد روش‌های مختلفی برای ارزیابی کارکرد افتراقی سئوال‌ها شده است. به عنوان مثال هیلز[1] بیش از چهل روش را را مشخص کرده است. روش‌های مختلف ارزیابی کارکرد افتراقی از لحاظ مفروضه‌ها، شرایط کاربرد و نتایج حاصل با یکدیگر تفاوت دارند و همین مسئله زمینه‌ای را برای پژوهشگران، به منظور مقایسه روش‌های مختلف ارزیابی کارکرد افتراقی سئوال‌ها، با هدف دستیابی به مناسب‌ترین روش برای هر موقعیت ایجاد کرده است (زومبو، ۱۹۹۹). پژوهشگران روش‌های ارزیابی کارکرد افتراقی را بر اساس الگوها و آماره‌های به کار رفته طبقه‌بندی کرده‌اند که این موجب درک بهتری از روش‌های مختلف می‌شود.

در طبقه‌بندی کلی می‌توان روش‌های تحلیل کارکرد افتراقی را در چارچوب نظریه کلاسیک و یا نظریه سئوال - پاسخ[2] طبقه‌بندی کرد. در نظریه کلاسیک، تفاوت مشخصه‌های کلاسیک سئوال، در بین گروه‌های مختلف آزمودنی با نمره‌های مشابه مورد مقایسه قرار می‌گیرد. سادگی اجرا و کاربرد نمونه‌های کوچک از مزیت‌های روش‌های مبتنی بر رویکرد کلاسیک است (امبرتسون[3] و رایس[4]، ترجمه شریفی و همکاران، ۱۳۸۸). برخی از متداول‌ترین روش‌های بررسی کارکرد افتراقی عبارتند از:

1. Hills
2. Item response theory
3. Embreston
4. Reise

۱- روش‌های دشواری سئوال تبدیل شده[۱]: این روش توسط انگاف[۲] و فورد[۳] پیشنهاد شده است و به آن طرح دلتا نیز گفته می‌شود. در این روش پارامتر دشواری سئوال در هر یک از زیر گروه‌های جامعه به طور جداگانه محاسبه می‌شود، سپس مقادیر دشواری سئوال به مقیاس دلتا (میانگین ۱۳ و انحراف استاندارد ٤) تبدیل می‌شوند (باقی[٤] و فرارا[٥]، ۱۹۸۹). همبستگی مقادیر دلتا مربوط به دو گروه محاسبه می‌شود، به علاوه برای تمام سئوال‌ها نمودار مقادیر دلتا مربوط به هر زوج رسم می‌شود و با داده‌ها خط مستقیمی برازش داده می‌شود. سئوال‌هایی که مقادیر دلتا در آنها با فاصله زیادی از این خط قرار گرفته‌اند دارای سوگیری هستند (اُنیل، ۱۹۹۱). این روش به علت سادگی مورد توجه بسیار واقع شده است، اما نسبت به عدم یکسانی قدرت تشخیص سئوال‌ها حساس است.

۲- روش مانتل- هنزل[٦]: این روش توسط مانتل و هنزل (۱۹۵۹) به عنوان روشی برای مطالعه گروه‌های جور شده معرفی شده است. هالند و تایر (۱۹۸۸) این روش را در بررسی کارکرد افتراقی سئوال مورد استفاده قرار دادند (دورانس[۷] و هالند، ۱۹۹۲). روش مانتل- هنزل، روشی ناپارامتریک برای ارزیابی کارکرد افتراقی است. سئوال‌ها به صورت دوارزشی نمره‌دهی می‌شوند و برای تعیین آزمودنی‌ها با سطوح توانایی یکسان در دو گروه مرجع[۸] و کانونی[۹] نمره کل مورد استفاده قرار می‌گیرد (گروه مرجع گروهی از آزمودنی‌ها است که عملکرد آنها به عنوان نقطه مرجع محسوب می‌شود. گروه کانونی گروهی از آزمودنی‌ها است که انتظار می‌رود سئوال‌های آزمون به نفع آنان نباشد). این روش به دلیل ارائه آزمون آماری، برآورد اندازه اثر و کارآمدی آن در نمونه‌های با حجم کم، به عنوان یکی از رایج‌ترین روش‌های ارزیابی کارکرد افتراقی سئوال‌ها مطرح است. البته این روش فقط در تشخیص کارکرد افتراقی یکنواخت[۱۰] مناسب است و همین موضوع یکی از محدودیت‌های روش مانتل هنزل است. کارکرد افتراقی به دو صورت یکنواخت و غیریکنواخت[۱۱] است. در کارکرد

1. Transformed Item Difficulty
2. Angoff
3. Ford
4. Baghi
5. ferrara
6. Mantel- Haenszel
7. Dorans
8. Reference
9. Focal
10. Uniform
11. Non -Uniform

افتراقی یکنواخت تفاوت احتمال پاسخ‌گویی درست گروه کانونی و گروه مرجع در تمام سطوح توانایی یکسان است یا به عبارت دیگر در کارکرد افتراقی یکنواخت تعامل بین سطح توانایی و عضویت در گروه وجود ندارد (دریانا[1]، ۲۰۰۷). در کارکرد افتراقی یکنواخت سئوال‌ها از لحاظ سطوح دشواری در دو گروه متفاوت هستند، اما از لحاظ قدرت تشخیص تفاوتی ندارند (شولتز[2] و ویتنی[3]، ۲۰۰۵). در کارکرد افتراقی افتراقی غیریکنواخت احتمال پاسخ‌گویی درست گروه کانونی و گروه مرجع در تمام سطوح توانایی یکسان نیست و در واقع بین سطح توانایی و عضویت در گروه تعامل وجود دارد (دریانا، ۲۰۰۷) و سئوال‌ها از لحاظ سطح دشواری و قدرت تشخیص در دو گروه متفاوت قرار می‌گیرند (شولتز و ویتنی، ۲۰۰۵).

۳- روش رگرسیون لجستیک: این روش توسط سوامی‌ناتان[4] و راجرز[5] معرفی شده است. در این روش پاسخ سئوال‌ها، به عنوان متغیری وابسته در نظر گرفته می‌شود. رگرسیون لجستیک مبتنی بر مدل‌سازی آماری احتمال پاسخ صحیح به سئوال، بر اساس عضویت در گروه و ملاک است که در آن ملاک معمولاً نمره کل آزمون است. وجود کارکرد افتراقی با بررسی بهبود ایجاد شده در برازش مدل رگرسیون، پس از اضافه کردن عضویت در گروه و تعامل بین نمره آزمون و عضویت در گروه در مدل رگرسیون تعیین می‌شود (روسسو[6]، برتراند[7] و بیتو[8]، ۲۰۰٤). روش رگرسیون لجستیک لجستیک مانند روش مانتل هنزل، آزمون معناداری آماری و اندازه اثر ارائه می‌کند. از جمله مزایای این روش می‌توان به توانایی بررسی کارکرد افتراقی یکنواخت و غیریکنواخت اشاره کرد.

بررسی کارکرد افتراقی بر اساس نظریه سئوال - پاسخ به ابتدای دهه شصت میلادی و کاربرد مدل راش بر می‌گردد. لرد[9] و نوایک[10] ظرفیت بالای رویکرد سئوال - پاسخ را در ارزیابی کارکرد افتراقی سئوال مورد تاکید قرار دادند (دانکن، ۲۰۰٦). از لحاظ نظری رویکرد سئوال- پاسخ روش مناسبی برای ارزیابی کارکرد افتراقی سئوال‌ها

1. Driana
2. Shultz
3. Whitney
4. Swaminathan
5. Rogers
6. Rousseau
7. Bertrand
8. Boiteau
9. Lord
10. Novick

است. ویژگی نامتغیر بودن پارامترها در این رویکرد، چارچوبی نظری برای تعریف و تعیین کارکرد افتراقی سئوال‌های آزمون فراهم می‌سازد. اگر سئوال، در زیرگروه‌های جامعه یکسان عمل کند، در این حالت احتمال پاسخ‌گویی درست آزمودنی‌ها با سطح توانایی یکسان می‌بایست مشابه باشد. به این ترتیب بر اساس نظریه سئوال – پاسخ، سئوال زمانی دارای کارکرد افتراقی است که منحنی ویژگی سئوال یا به عبارت دیگر احتمال شرطی پاسخ درست به سئوال در سطوح توانایی یکسان در گروه مرجع و کانونی متفاوت باشند (امبرتسون و رایس، ترجمه شریفی و همکاران، ۱۳۸۸). رویکرد سئوال – پاسخ قادر به تشخیص کارکرد افتراقی یکنواخت و غیریکنواخت است. یکی از محدودیت‌های روش‌های مبتنی بر رویکرد سئوال – پاسخ ضرورت وجود گروه‌های نمونه بزرگ است. زیکی[1] (۱۹۹۳)، روسو و ستوات[2] کاربرد گروه نمونه با حجم بالای بالای صد نفر را برای گروه کانونی و بین دویست تا هزار نفر را برای گروه مرجع در ارزیابی کارکرد افتراقی توصیه کردند. از سویی کاربرد نمونه‌های بزرگ منجر به افزایش موارد مثبت کاذب می‌شود (دانکن، ۲۰۰۶).

در چارچوب رویکرد سئوال – پاسخ چندین روش برای ارزیابی کارکرد افتراقی مطرح شده است که این روش‌ها در دو گروه کلی مساحت میان منحنی ویژگی سئوال[3] گروه مرجع و کانونی و روش آزمون‌های آماری یکسانی پارامترهای سئوال طبقه‌بندی می‌شوند (شولتز و ویتنی، ۲۰۰۵). روش آزمون‌های آماری، شامل آزمون چند متغیره و آزمون t بر روی مقادیر پارامتر دشواری می‌شود. در رویکرد مساحت میان دو منحنی ویژگی سئوال، منحنی ویژگی و نه پارامتر سئوال‌ها مورد توجه قرار می‌گیرد. هر چند که منحنی ویژگی سئوال‌ها بر اساس پارامترهای سئوال‌ها ترسیم می‌شوند، اما روش بررسی مساحت بین دو منحنی ویژگی سئوال، مستلزم نگاه دقیق‌تری به تفاوت بین پارامتر سئوال‌ها است. در این روش منطقه بین دو منحنی ویژگی سئوال مورد توجه قرار می‌گیرد. پس از برآورد پارامتر سئوال‌ها و قرار دادن آن‌ها روی مقیاس مشترک، منحنی سئوال در دو گروه ترسیم می‌شود. در صورتی که فاصله بین دو منحنی ویژگی سئوال صفر باشد و در واقع دو منحنی بر هم منطبق باشند، کارکرد افتراقی وجود ندارد. برعکس هنگامی که مساحت بین دو منحنی سئوال صفر نیست کارکرد افتراقی با درجاتی وجود دارد. هر چه

1. Zieky
2. Stout
3. Item Characteristic Curve

این مساحت بیشتر باشد، کارکرد افتراقی نیز بیشتر است. همبلتون و راجو[1] فرمول‌هایی برای محاسبه مساحت ارائه کرده‌اند (همبلتون، سوامیناتان و راجرز، ترجمه فلسفی‌نژاد، ۱۳۸۹). راجو نارسایی رویکرد محاسبه مساحت بین دو منحنی ویژگی سئوال را درک کرده و روش‌های مبتنی بر چارچوب کارکرد افتراقی سئوال‌ها و آزمون[2] (DFIT) را برای ارزیابی کارکرد افتراقی ارائه کرد. شاخص DFIT ویژگی‌هایی دارد که آن را توانمند و منعطف می‌سازد. این ویژگی‌ها عبارتند از :۱- امکان کاربرد این شاخص برای سئوال‌های دو ارزشی و چندارزشی ۲- امکان کاربرد این شاخص در مدل‌های تک بعدی و چند بعدی ۳- توانایی ارزیابی کارکرد افتراقی سئوال و آزمون (اُشیما و موریس، ۲۰۰۸).

مقایسه روش‌های مختلف تعیین کارکرد افتراقی سئوال‌ها

به منظور مقایسه روش‌های مختلف ارزیابی کارکرد افتراقی، مطالعات متعددی با استفاده از داده‌های حقیقی و یا شبیه‌سازی شده انجام شده است که در برخی موارد نتایج متناقضی به همراه داشته‌اند. دو روش رگرسیون لجستیک و روش مانتل هنزل به عنوان رایج‌ترین روش‌های ارزیابی کارکرد افتراقی در بسیاری از مطالعات مورد مقایسه قرار گرفته‌اند. به عنوان مثال هیدالگو[3] و لوپز[4] در مطالعه‌ای به بررسی این دو رویکرد در ارزیابی کارکرد افتراقی سئوال‌های آزمون پرداختند. نتایج مطالعه آنان بیانگر آن بود که روش رگرسیون لجستیک سئوال‌های بیشتری را با کارکرد افتراقی مشخص می‌سازد. سوامی‌ناتان و راجرز (۱۹۹۰) نیز با استفاده از مطالعات شبیه‌سازی شده به نتیجه مشابه دست یافتند (دانکن، ۲۰۰۶). گیرل[5] و همکاران (۱۹۹۹) در مطالعه‌ای کارکرد افتراقی جنسیتی سئوال‌های آزمون ریاضی و علوم را با استفاده از دو روش مانتل هنزل و رگرسیون لجستیک مورد بررسی قرار دادند. در حالی که نتایج این دو روش در بررسی کارکرد افتراقی یکنواخت در آزمون ریاضی مشابه بود، نتایج آزمون علوم ثبات کم‌تری داشت. روش مانتل هنزل در مقایسه با روش رگرسیون لجستیک تعداد کم‌تری از سئوال‌ها را دارای کارکرد افتراقی تشخیص داد و بنابر این روشی محافظه‌کارانه‌تر می‌باشد. با توجه به مطالعات ارائه شده می‌توان نتیجه گرفت

1. Raju
2. Differential Functioning of Items and Tests
3. Hidalgo
4. Lopez
5. Gierl

که تعداد سئوال‌های دارای کارکرد افتراقی، به روش به‌کار رفته وابسته است و از این رو سازندگان آزمون‌ها و سیاست‌گذاران آموزشی می‌بایست تفاوت‌های روش‌های مختلف ارزیابی کارکرد افتراقی سئوال‌ها را همواره در نظر بگیرند (احمدی، ۱۳۸۷).

باقی و فرارا (۱۹۸۹) در مطالعه‌ای سه روش نمودار دلتا، روش مانتل هنزل و رویکرد مبتنی بر نظریه سئوال – پاسخ را با استفاده از نمونه‌هایی با حجم متفاوت مورد مقایسه قرار دادند. روش‌ها از لحاظ ثبات (مطابقت نتایج هر روش در نمونه‌ها با حجم‌های متفاوت)، توافق (مطابقت نتایج روش‌های مختلف ارزیابی کارکرد افتراقی با نتایج روش مبتنی بر رویکرد سئوال – پاسخ) و عملی بودن (نرم‌افزارهای مورد نیاز، حجم گروه نمونه مورد نیاز) با یکدیگر مقایسه شدند. ثبات روش نمودار دلتا و رویکرد سئوال – پاسخ کم تا متوسط بود. در نمونه‌های بزرگ ثبات روش سئوال – پاسخ، نمودار دلتا و روش مانتل هنزل در حد متوسط بود. توافق روش‌ها با محاسبه همبستگی رتبه‌ای اسپیرمن میان شاخص‌های کارکرد افتراقی روش‌ها و مشخص کردن تعداد سئوال‌ها دارای کارکرد افتراقی و فاقد کارکرد افتراقی در هر یک از روش‌ها تعیین شد. بر اساس ضرایب همبستگی رتبه‌ای اسپیرمن، توافق میان روش راش و نمودار دلتا در تمام نمونه‌ها در حد بسیار بالا بود، توافق روش راش با مدل سه پارامتری در حد متوسط بود و توافق سایر روش‌ها با روش سه پارامتری کم بود. هریس[1] و کولن[2] (۱۹۸۶)، هریس و هوور[3] (۱۹۸۶) و اسکاگز[4] و لیسیتز[5] (۱۹۸۶) اظهار داشتند که میزان توافق بین روش‌های مختلف ارزیابی کارکرد افتراقی خیلی زیاد نیست (باقی و فرارار، ۱۹۸۹).

عبدالعزیز[6] (۲۰۱۰) کارکرد افتراقی جنسیتی سئوال‌ها را در آزمون ریاضی با استفاده از روش مانتل هنزل، تفاوت پارامتر دشواری سئوال‌های آزمون و روش پارامتر دشواری تبدیل شده(TID) را بررسی و میزان توافق این سه روش را تعیین کرد. نتایج تحقیق وی بیانگر آن بود که میزان توافق بین این سه روش پایین است. بیشترین توافق، بین روش مانتل هنزل و روش پارامتر دشواری تبدیل شده (TID) به دست آمد که این مسئله ممکن است به این دلیل باشد که هر دو این روش‌ها در چارچوب

1. Harris
2. Kolen
3. Hoover
4. Skaggs
5. Lissitz
6. Abedalaziz

نظریه کلاسیک هستند. کم‌ترین توافق میان روش پارامتر دشواری تبدیل شده (TID) و تفاوت پارامتر دشواری سئوال‌ها است.

دوراند[1] و پارک[2] (۲۰۰۶) کارکرد افتراقی را در سئوال‌های آزمون زبان با استفاده از سه روش SIBTEST، مانتل هنزل و رویکرد سئوال - پاسخ بررسی کردند. نتایج نشان داد که دو روش SIBTEST و مانتل هنزل تعداد زیادی سئوال دارای کارکرد افتراقی را مشخص ساختند، در حالی‌که روش سئوال - پاسخ تعداد سئوال‌های کم‌تری را به عنوان سئوال‌های دارای کارکرد افتراقی مشخص کرده است. بر اساس این نتایج آنان مطالعات بیشتر و دقیق‌تری را در باره روش‌های مختلف ارزیابی کارکرد افتراقی توصیه کردند.

یکی از مسایل مهم در مطالعات کارکرد افتراقی تعیین عوامل به وجود آورنده آن است. در باره سئوال‌های دو ارزشی مطالعاتی صورت گرفته است که از آن جمله می‌توان به مطالعه اونیل و مکپیک[3] اشاره کرد که تأثیر عواملی چون محتوا و شکل سئوال را در آزمون استعداد تحصیلی[4] (SAT)، آزمون‌های ورودی تحصیلات تکمیلی[5] (GRE) و سایر آزمون‌های پذیرش مورد بررسی قرار دادند. نتایج مطالعه آنان بیانگر آن بود که در آزمون ریاضی آزمودنی‌های دختر در مقایسه با آزمودنی‌های پسر در سئوال‌های جبر عملکرد بهتری دارند، در حالی‌که آزمودنی‌های پسر در حل مسایل، عملکرد بهتری دارند (وانگ و لان، ۱۹۹٤). پژوهشگران هم‌چنین نشان دادند که کارکرد افتراقی جنسیتی به میزان شباهت سئوال با سئوال‌های کتاب‌های درسی نیز بستگی دارد. آزمودنی‌های دختر در مقایسه با پسران در سئوال‌هایی که مشابه سئوال‌های کتاب‌های درسی هستند، عملکرد بهتری دارند (ویلینگهام و کول، ۱۹۹۷). در مورد عوامل به وجود آورنده کارکرد افتراقی در سئوال‌های چند ارزشی مطالعات اندکی انجام شده است (وانگ و لان، ۱۹۹٤). دودین[6] و الدرابی[7] (۲۰۰۳) نیز کارکرد افتراقی سئوال‌های آزمون ریاضی را با استفاده از روش مانتل هنزل مورد بررسی قرار دادند. آزمون متشکل از پنجاه سئوال چندگزینه‌ای بود که با نمونه چهارصد نفری از دانش‌آموزان اجرا شد (۲۰۰ دانش‌آموزان دختر و ۲۰۰ دانش‌آموزان پسر). از این پنجاه

1. Durand
2. Park
3. McPeek
4. Scholastic Aptitude Test
5. Graduate Record Examinations (GRE)
6. Doudin
7. Al-Darabee

سئوال ۱۷ سئوال (۳٤ درصد) کارکرد افتراقی داشتند که هشت سئوال به نفع دانش‌آموزان پسر و نُه سئوال به نفع دانش‌آموزان دختر بود. نتایج مطالعه آنان بیانگر برتری دانش‌آموزان پسر در سئوال‌های حساب و حل مسئله و هم‌چنین برتری دانش‌آموزان دختر در سئوال‌ها جبر و هندسه بود. بربراوغلو[1] (۱۹۹٦) وجود کارکرد افتراقی جنسیتی در سئوال‌های ریاضی آزمون سراسری ورود به دانشگاه را با استفاده از مدل دو پارامتری مورد بررسی قرار داد. نتایج تحلیل مبین آن بود که از ۳۲ سئوال چندگزینه‌ای آزمون، ۵۳ درصد سئوال‌ها به نفع مردان و ٤۷ درصد به نفع زنان بوده است (دریانا، ۲۰۰۷).

مندز[2]، بارنت[3] و ارکیکان[4] (۲۰۰٦) منابع کارکرد افتراقی سئوال‌ها با توجه به عامل عامل جنسیت را مورد بررسی قرار دادند. بر اساس فرضیه آنان، محتوا، سطوح شناختی و متن سئوال، منابع احتمالی کارکرد افتراقی سئوال محسوب می‌شوند. آنان دریافتند که در حوزه محتوایی، مجموعه سئوال‌های حل مسئله و مجموعه سئوال‌های مربوط به حوزه‌های لگاریتم و توان به نفع دانش‌آموزان پسر و مجموعه سئوال‌های چندجمله‌ای‌ها و روابط درجه دوم به نفع دانش‌آموزان دختر است. بر اساس سطوح شناختی سئوال‌ها، دانش‌آموزان پسر در مقایسه با دختران در سئوال‌های سطوح بالای شناختی عملکرد بهتری داشتند. در سئوال‌های مربوط به سطوح پایین شناختی تفاوتی مشاهده نشد. مطالعه آنان در باره متن سئوال‌ها نشان داد سئوال‌هایی که شامل شکل و نمودار هستند به نفع پسران است. نتایج این پژوهش موافق با نتایج سایر مطالعات انجام شده در این زمینه است (دریانا، ۲۰۰۷).

اهداف مطالعه

اتخاذ تصمیمات درست و منطقی برپایه نتایج آزمون‌ها مستلزم کسب اطمینان از عادلانه بودن آزمون‌ها است. بنابراین ارزیابی کارکرد افتراقی سئوال‌های آزمون‌ها اهمیت فراوانی دارد و می‌بایست به عنوان بخشی از فرایند تحلیل آزمون در نظر

1. Berberoglu
2. Mendes
3. Barnett
4. Ercikan

گرفته شود. با توجه به اهمیت آزمون‌های ورودی دانشگاه‌ها و نقش آنها در سرنوشت تحصیلی و شغلی داوطلبان، لازم است سئوال‌های آزمون‌های ورودی از نظر وجود کارکرد مورد بررسی دقیق و موشکافانه قرار گیرد. در این راستا مطالعه حاضر به ارزیابی کارکرد افتراقی جنسیتی در سئوال‌های آزمون ریاضی گروه آزمایشی ریاضی و فنی با استفاده از روش مانتل – هنزل و روش مبتنی بر نظریه سئوال– پاسخ پرداخته است. به علاوه میزان توافق بین این دو روش در تعیین کارکرد افتراقی نیز مورد بررسی قرار گرفته است. در مطالعه حاضر آزمودنی‌های دختر به عنوان گروه کانونی و آزمودنی‌های پسر به عنوان گروه مرجع در نظر گرفته شده‌اند.

سئوال تحقیق

آیا عملکرد آزمودنی‌های دختر و پسر با سطوح توانایی مشابه در آزمون ریاضی کنکور سراسری گروه آزمایشی ریاضی و فنی متفاوت است؟

داده‌های تحلیل

تفاوت عملکرد در آزمون‌های ریاضی با توجه به ویژگی‌ها نظیر وضعیت اقتصادی – اجتماعی، جنسیت و نژاد از جمله مسایلی است که مورد توجه و پژوهش فراوان قرار گرفته است. در مورد آزمون‌های ریاضی، کارکرد افتراقی دغدغه‌ای اساسی است. در مطالعه حاضر کارکرد افتراقی جنسیتی سئوال‌های آزمون ریاضی مورد بررسی قرار گرفته است. در تحلیل حاضر از پاسخ‌های آزمودنی‌های شرکت کننده در کنکور سراسری گروه آزمایشی علوم ریاضی و فنی سال ۱۳۸۲ استفاده شده است. ۳۹۹۱۹۹ آزمودنی در این آزمون شرکت کرده‌اند که ۷۸۷۵ آزمودنی به تمام سئوال‌های آزمون ریاضی پاسخ غلط داده‌اند و در نتیجه از روند تحلیل کنار گذاشته شدند. به این ترتیب تحلیل براساس پاسخ‌های ۳۹۱۳۲٤ آزمودنی انجام شده است. ۲۱۳٦۲۵ آزمودنی پسر و ۱۷۷٦۹۹ نفر دختر بودند. با استفاده از روش نمونه‌گیری تصادفی طبقه‌ای گروه نمونه چهارهزار نفری از آزمودنی‌ها انتخاب شد که شامل ۲۲۰۰ آزمودنی مرد و ۱۸۰۰ آزمودنی بود. آزمون شامل ۵۵ سئوال چهارگزینه‌ای بود که برای پاسخ‌های نادرست نمره منفی لحاظ شده است.

روش تحلیل

در مطالعه حاضر بررسی کارکرد افتراقی سئوال‌ها براساس دو روش سئوال - پاسخ و مانتل - هنزل صورت گرفته است. اولین گام در تحلیل مبتنی بر نظریه سئوال - پاسخ بررسی برقراری مفروضه‌های مدل است، چون کاربرد این نظریه و کسب نتایج دقیق از آن مستلزم برقراری مفروضه بنیادی آن است. هر چند برخی شواهد بیانگر مقاوم بودن نظریه سئوال - پاسخ به تخطی از مفروضه‌های بنیادی‌اش است (همبلتون، ۱۹۸۹). یکی از مفروضه‌های مدل، تک‌بعدی بودن[1] است، به این معنا که یک خصیصه مکنون با مجموعه سئوال‌های آزمون اندازه‌گیری شود. برقراری مطلق این مفروضه به دلیل دخالت عوامل گوناگون (اضطراب، انگیزش، مهارت‌ها و فرایندهای شناختی و وضعیت جسمانی و روانی آزمودنی‌ها) بر عملکرد آزمودنی‌ها غیرممکن است، اما آن چه اهمیت دارد وجود عاملی بارز در آزمون است که معرف عملکرد آزمودنی‌ها در آزمون باشد. روش‌های مختلفی برای بررسی برقراری مفروضه تک‌بعدی هست که یکی از مناسب‌ترین آنها تحلیل عاملی[2] است. در مطالعه حاضر تحلیل عاملی سئوال‌ها با استفاده از نرم‌افزار TESTFACT و بر اساس کل داده‌ها انجام شده است. ارزش ویژه نخستین عامل استخراج شده در تحلیل عاملی برابر با ۱۵/۶۳ است و ۲۸/۳۶ درصد واریانس را تبیین می‌کند. دومین ارزش ویژه برابر با ۲/۰۱ است که ۳/۰۵ درصد واریانس را تبیین می‌کند. طبق اظهار ریکازی[3] هنگامی که عامل نخست حداقل ۲۰ درصد واریانس کل را تبیین کند، می‌توان نتیجه گرفت آزمون تک‌بعدی است (احمدی، ۲۰۰۸). بر این اساس در مطالعه حاضر آزمون تک‌بعدی است و تحلیل آزمون بر اساس نظریه سئوال- پاسخ تک‌بعدی امکان پذیر است.

جدول (۱) عامل‌های استخراج شده آزمون ریاضی

عامل	ارزش ویژه	درصد واریانس تبیین شده
۱	۱۵/۶۳	۲۸/۳۶
۲	۲/۰۱	۳/۰۵
۳	۱/۳۵	۲/٤٤
٤	۱/۱۷	۲/۱۱
٥	۱/۰۹	۱/۹۸

1. Unidimensionality
2. Factor analysis
3. Reakase

بررسی وجود کارکرد افتراقی سئوال‌ها با استفاده از نرم‌افزارBILOG- MG صورت گرفته است. این نرم‌افزار توانایی مدل‌سازی چندگروهی نظریه سئوال – پاسخ را دارد. بنابراین می‌توان از آن برای ارزیابی کارکرد افتراقی سئوال استفاده کرد. نرم‌افزارBILOG- MG، کارکرد افتراقی سئوال را از نظر پارامترهای دشواری سئوال بررسی می‌کند و قادر به بررسی تفاوت‌های گروهی در پارامتر تشخیص نیست (امبرتسون و رایس، ترجمه شریفی و همکاران، ۱۳۸۸). نکته قابل تأمل در کاربرد روش سئوال– پاسخ، برازش مدل با داده‌ها است. در واقع، برتری‌های رویکرد سئوال– پاسخ تنها در صورت برازش داده‌ها با مدل برقرار است. برازش مدل انتخابی با استفاده از شاخص لگاریتم درستنمایی صورت گرفت. تفاوت مقدار این شاخص در مدل‌های مختلف دارای توزیع مجذور کای با درجه آزادی برابر با تعداد پارامترهای افزوده برای مدل پیچیده‌تر است (امبرتسون و رایس، ترجمه شریفی و همکاران، ۱۳۸۸). مقادیر این شاخص در جدول (۲) ارائه شده است. با توجه به جدول، مدل دو پارامتری با داده‌ها برازش دارد.

جدول (۲) مقادیرشاخص برازش لگاریتم درستنمایی مدل‌های مختلف

مدل سئوال – پاسخ	لگاریتم درستنمایی
تک پارامتری	٤٧٥٢٥/٤٨٩
دو پارامتری	٤٦٨٠٦/٨٧٧٩
سه پارامتری	٤٦٨١١/١٤٣٩

پس از کسب اطمینان از برقراری مفروضه‌های مدل و برازش مدل با داده‌ها، در گام اول به منظور تعیین وجود کارکرد افتراقی در سئوال‌ها، داده‌ها دوبار تحلیل می‌شوند. در نوبت اول داده‌ها به طور کلی تحلیل می‌شوند به گونه‌ای که از جامعه یکسانی انتخاب شده‌اند. در نوبت دوم داده‌ها در زیرگروه‌ها با استفاده از تحلیل کارکرد افتراقی با فرض صفر عدم وجود کارکرد افتراقی، تحلیل می‌شوند. تفاوت لگاریتم درستنمایی دو مرحله دارای توزیع مجذور کای با درجه آزادی (n-1)(m-1) است که n معرف تعداد سئوال‌ها و m معرف تعداد گروه‌ها است. هنگامی که مجذور کای معنادار است، شواهدی از کارکرد افتراقی وجود دارد. بر این اساس در مطالعه حاضر داده‌ها یک‌بار به طور کلی و یک‌بار در دو گروه داوطلبان دختر و پسر با استفاده از تحلیل کارکرد افتراقی تحت فرضیه صفر مبنی بر عدم وجود کارکرد افتراقی

تحلیل شدند. مقدار لگاریتم درستنمایی در مرحله اول برابر با ۱۸۸۰۳٤/۸٦ و در مرحله دوم ۱۸۷۷۸٤/۰۲ است که تفاوت آن ۲۵۰/۸٤ است و در سطح ۰/۰۵ با درجه آزادی ۵٤ معنادار است که این نتیجه بیانگر وجود اثرات کارکرد افتراقی در آزمون ریاضی است.

پس از تأیید وجود کارکرد افتراقی در آزمون، در گام بعدی تحلیل کارکرد افتراقی سئوال‌ها انجام شد. نرم‌افزار کارکرد افتراقی را فقط بر اساس پارامتر دشواری بررسی می‌کند. کارکرد افتراقی در سه مرحله مورد بررسی قرار می‌گیرد. در مرحله نخست پارامتر سئوال‌ها برای هر یک از دو گروه آزمودنی به طور جداگانه برآورد می‌شود. البته پارامتر تشخیص سئوال‌ها برای دو گروه یکسان در نظر گرفته می‌شود. برنامه BILOG- MG میانگین و انحراف استاندارد گروه‌ها را یکسان فرض نمی‌کند. میانگین و انحراف استاندارد گروه مرجع به ترتیب صفر و یک است، میانگین و انحراف استاندارد گروه کانونی پارامترهایی هستند که برآورد می‌شوند. در مرحله دوم میانگین دشواری سئوال‌ها برای گروه مرجع و کانونی محاسبه می‌شود و تفاوت میانگین دشواری سئوال‌های دو گروه کانونی و مرجع محاسبه می‌شود. پارامتر دشواری سئوال‌های گروه کانونی با کم کردن تفاوت میانگین دشواری دو گروه از پارامتر دشواری سئوال‌های گروه کانونی، تعدیل می‌شوند. در مرحله سوم تفاوت میان پارامترهای دشواری سئوال‌های مربوط به گروه مرجع و کانونی و خطای استاندارد محاسبه می‌شود، تفاوت زیاد نشان‌دهنده کارکرد افتراقی سئوال می‌باشد. در واقع سئوال‌هایی که تفاوت میان پارامتر دشواری آنها در دو گروه تقریباً بیش از دو برابر خطای استاندارد (S. E×۱/۹٦) است به عنوان سئوال‌هایی با کارکرد افتراقی در نظر گرفته می‌شوند. در تحلیل حاضر با توجه به این‌که داوطلبان دختر گروه کانونی و داوطلبان پسر، گروه مرجع را تشکیل می‌دهند، مقادیر منفی تفاوت میان پارامترهای دشواری سئوال‌ها در گروه مرجع و کانونی بیانگر آن است که سئوال به نفع داوطلبان دختر است و برعکس. نتایج تحلیل سئوال‌ها آزمون در جدول (۳) ارائه شده است. همان گونه که در جدول مشخص شده است نُه سئوال آزمون (۱٦/٤ درصد سئوال‌ها) کارکرد افتراقی دارند که تمام این سئوال‌ها به نفع آزمودنی‌های دختر سوگیری دارند (سئوال‌های شماره ۵، ۷، ۹، ۱۸، ۲۷، ۳۲، ۳٦، ٤۱ و ٤۹).

جدول (۳) نتایج تحلیل مبتنی بر رویکرد سئوال - پاسخ

سئوال	تفاوت دشواری گروه‌ها داوطلبان پسر - داوطلبان دختر	DIF	سئوال	تفاوت دشواری گروه‌ها داوطلبان پسر - داوطلبان دختر	DIF
۱	-۰/۰۳ *۰/۰۵٤		۲۹	۰/۰۵٤ *۰/۱۵۹	
۲	۰/۰۵٤ *۰/۲۰۸		۳۰	۰/۰۷۱ *۰/۲۹۷	
۳	-۰/۰٤۲ *۰/۰٦۵		۳۱	-۰/۰۲٤ *۰/۱۹٤	
٤	۰/۰۱۲ *۰/۳٦۷		۳۲	-۰/۳٤٦ *۰/۰٦٤	+
۵	-۰/۲۱۳ *۰/۰٦	+	۳۳	۰/۳۱ *۰/٤۲۱	
٦	-۰/۰۳۲ *۰/۰٦۹		۳٤	۰/۲۵۱ *۰/۲۵۸	
۷	-۰/۲۸۵ *۰/۱۱۱	+	۳۵	۰/۱۵۹ *۰/۰۹۸	
۸	۰/۱۵۲ *۰/۰۹۱		۳٦	-۰/۱۸۸ *۰/۰۹٤	+
۹	-۰/۱۸ *۰/۰۵۲	+	۳۷	-۰/۲۰۱ *۰/۱۱٦	
۱۰	۰/٦۲۲ *۰/٤۲		۳۸	۰/۰۱ *۰/۲۰۵	
۱۱	-۰/۱٤۵ *۰/۱۰۷		۳۹	-۰/۰۱۸ *۰/۲۷۲	
۱۲	۰/۲۰۸ *۰/۵۸۸		٤۰	-۰/۸ *۰/۱۸٤	
۱۳	۰/۲۸٤ *۰/۳۳۳		٤۱	-۰/۲۰۲ *۰/۰۷۷	+
۱٤	۰/۳۱۱ *۰/۳٤۸		٤۲	-۰/۰۲۹ *۰/۰۹۹	
۱۵	-۰/۲٤٤ *۰/۲۸۹		٤۳	۰/۲۸۱ *۰/۳۲۲	
۱٦	۰/۳۷٤ *۰/۵۲		٤٤	۰/۳۹٤ *۰/۳۰٦	

سئوال	تفاوت دشواری گروه‌ها داوطلبان پسر – داوطلبان دختر	DIF	سئوال	تفاوت دشواری گروه‌ها داوطلبان پسر – داوطلبان دختر	DIF
۱۷	۰/۰۳۸ *۰/۱۲٤		٤٥	−۰/۳۰۹ *۰/۱٦٥	
۱۸	−۰/۳۸ *۰/۰۷٦	+	٤٦	۰/۰٦۳ *۰/۱٥۹	
۱۹	۰/۱۱٦ *۰/۲٤۸		٤۷	۰/۰٥ *۰/۱۰٥	
۲۰	۰/۱۲۷ *۰/٤٥۳		٤۸	−۰/۲۲٥ *۰/۱٤٦	
۲۱	۰/۰۰۷ *۰/۱۱۲		٤۹	−۲/۱۲٦ *۰/٦۸٤	+
۲۲	−۰/۰۰٤ *۰/۱۱		٥۰	۰/۰۸٥ *۰/۳۳۹	
۲۳	−۰/۱۸۷ *۰/۱۹۷		٥۱	۰/۰۲۳ *۱/۱۰۸	
۲٤	−۰/۱۳۸ *۰/۰۷۲		٥۲	۰/۲۰٦ *۰/۳۱۸	
۲٥	۰/۰٥۲ *۰/۱٦		٥۳	۰/۲٥٤ *۰/۱۷۱	
۲٦	−۰/۱۱۷ *۰/۰۸۲		٥٤	−۰/۰۲۹ *۰/۱۷۱	
۲۷	−۰/۱۹۷ *۰/۰۷٥	+	٥٥	۱/۳۰٦ *۱/۰۳۹	
۲۸	۰/۱۸۸ *۰/۱۳۱				

(*)خطای استاندارد، (+) وجود کارکرد افتراقی

در ادامه تحلیل کارکرد افتراقی سئوال‌ها با استفاده از روش مانتل هنزل مورد ارزیابی قرار گرفت. در مطالعه حاضر تحلیل مانتل هنزل با استفاده از نرم‌افزار Leartap انجام شده است. برای هر سئوال در هر سطح نمره، تعداد و نسبت افراد هریک از گروه‌های مرجع و کانونی که به سئوال، پاسخ درست داده‌اند، شاخص مجذور کای مانتل هنزل،

آلفای مانتل هنزل[1]، دلتای مانتل هنزل[2]، نسبت بخت‌ها، معناداری و اندازه اثر محاسبه می‌شود. نسبت بخت، اندازه‌ای نسبی است که احتمال این‌که فردی از گروه مرجع در مقایسه با گروه کانونی به سئوال پاسخ درست بدهد را نشان می‌دهد. مقادیر بزرگ‌تر از یک نشان‌دهنده این است که احتمال پاسخ‌گویی درست بیشتر به نفع گروه مرجع است و مقادیر کم‌تر از یک بیانگر آن است که احتمال پاسخ‌گویی درست بیشتر به نفع گروه کانونی است. آلفای مانتل هنزل عبارت است از میانگین نسبت بخت‌ها در تمام سطوح نمره که براساس تعداد افراد هر یک از گروه‌ها در هر سطح نمره وزن‌دهی شده‌اند. بخت، به عنوان شاخصی از احتمال وقوع یک حادثه، با این محدودیت همراه است که دامنه آن همواره بین صفر تا مثبت بی‌نهایت است یا به عبارتی ناقرینه[3] است. به همین علت از آن لگاریتم گرفته و به این ترتیب لگاریتم آلفای مانتل هنزل به دست می‌آید که دامنه آن در بازه منفی بی‌نهایت تا مثبت بی‌نهایت است. سئوال‌ها با مقادیر مثبت به سود گروه مرجع و سئوال‌ها با مقادیر منفی به نفع گروه کانونی سوگیری دارند. هالند و تایر برای نشان دادن مقدار کارکرد افتراقی شاخص دیگری را پیشنهاد کردند که در واقع تبدیل آلفای مانتل هنزل به مقیاسی دیگر است که دلتای مانتل هنزل نامیده می‌شود و از رابطه زیر به دست می‌آید:

$$\text{MH D-DIF} = -2.35^{*} \text{ Ln MH alpha}$$

سئوال‌ها با مقادیر دلتای مانتل هنزل مثبت به نفع گروه کانونی و سئوال‌ها با مقادیر منفی به نفع گروه مرجع هستند. خدمات سنجش آموزشی[4] طبقه‌بندی سه سطحی را برای کارکرد افتراقی سئوال ارائه کرده است (زویک[5]، ۲۰۱۲) که عبارتند از:

۱- سطح A : قدر مطلق دلتای مانتل هنزل سئوال‌های این طبقه کم‌تر از ۱ است. سئوال‌های این طبقه فاقد کارکرد افتراقی یا کارکرد افتراقی ناچیزی دارند.

۲- سطح B : سئوال‌ها با قدر مطلق دلتای مانتل هنزل بین ۱ تا ۱/۵در این طبقه قرار می‌گیرند. این سئوال‌ها کارکرد افتراقی متوسطی دارند.

1. MH alpha
2. MH D-DIF
3. Asymmetry
4. Educational Testing Service (ETS)
5. Zwick

۳- سطح C : سئوال‌های این طبقه با شاخص مجذور کای مانتل هنزل معنادار و قدر مطلق دلتای مانتل هنزل بیش از ۱/۵ مشخص می‌شوند. کارکرد افتراقی سئوال‌ها زیاد است و چنین سئوال‌هایی از آزمون حذف می‌شوند (زویک، ۲۰۱۲).

نتایج تحلیل مانتل هنزل سئوال‌های آزمون در جدول (۵) ارائه شده است.

جدول (۵) نتایج تحلیل مانتل هنزل

سئوال	شاخص مانتل هنزل	معناداری	آلفای مانتل هنزل	دلتای مانتل هنزل	طبقه‌بندی ETS
۱	۰/٦۱	۰/٤٤	۱/۰٦	-۰/۱۳	A
۲	۰/٦٥	۰/٤۲	۱/۱۰	-۰/۲۲	A
۳	۰/۲	۰/٦٦	۱/۰۳	-۰/۰۸	A
٤	۰/۱٤	۰/۷۱	۱/۰٤	-۰/۱۰	A
٥	۷/۸۱	*۰/۰۱	۰/۸۰	۰/٥۲	A
٦	۰/۲۹	۰/٥۹	۱/۰٤	-۰/۱۰	A
۷	۷/۷۲	*۰/۰۱	۰/۸۱	۰/٥۰	A
۸	۹/۲٤	*۰/۰۰	۱/۳۱	-۰/٦۳	A
۹	٤/۲۲	*۰/۰٤	۰/۸٦	۰/۳٦	A
۱۰	۱٤/۳٦	*۰/۰۰	۱/٥۹	-۱/۰۹	B
۱۱	۰/۷۷	۰/۳۸	۰/۹٤	۰/۱٥	A
۱۲	۲/۱٦	۰/۱٤	۱/۱۳	-۰/۲۸	A
۱۳	٤/٥۹	*۰/۰۳	۱/۲٤	-۰/٥۰	A
۱٤	٥/۲٦	*۰/۰۲	۱/۳٦	-۰/۷۲	A
۱٥	۱/٤۹	۰/۲۲	۰/۹۱	۰/۲۱	A
۱٦	٤/٥۷	*۰/۰۳	۱/۳۲	-۰/٦٥	A
۱۷	۰/۰۱	۰/۹۳	۰/۹۹	۰/۰۲	A
۱۸	۲۳/٥٥	*۰/۰۰	۰/۷۰	۰/۸٥	A
۱۹	۱/۰٤	۰/۳۱	۱/۱۳	-۰/۲۹	A
۲۰	۱/۰۷	۰/۳۰	۱/۱۱	-۰/۲٥	A
۲۱	۰/٤٤	۰/٥۱	۱/۰٦	-۰/۱٤	A
۲۲	۰/۲۳	۰/٦۳	۱/۰٤	-۰/۰۹	A
۲۳	۱/٥٦	۰/۲۱	۰/۸۸	۰/۳۱	A
۲٤	۱/۳۷	۰/۲٤	۰/۹۱	۰/۲۳	A
۲٥	۱/٤۹	۰/۲۲	۱/۱٤	-۰/۳۰	A
۲٦	۱/۱٤	۰/۲۹	۰/۹۰	۰/۲٥	A

سئوال	شاخص مانتل هنزل	معناداری	آلفای مانتل هنزل	دلتای مانتل هنزل	طبقه‌بندی ETS
۲۷	٦/۱۲	*۰/۰۱	۰/۷۵	۰/٦۷	A
۲۸	۵/٦۲	*۰/۰۲	۱/۲۰	-۰/٤۳	A
۲۹	۰/۵٦	۰/٤۵	۱/۰۸	-۰/۱۸	A
۳۰	۰/۳۹	۰/۵۳	۱/۰۸	-۰/۱۷	A
۳۱	۰/۰۲	۰/۹۰	۱/۰۱	-۰/۰۳	A
۳۲	۲۳/۰۹	*۰/۰۰	۰/٦۹	۰/۸٦	A
۳۳	٤/۳۲	*۰/۰٤	۱/۳۷	-۰/۷٤	A
۳٤	٤/۰٦	*۰/۰٤	۱/۳۰	-۰/٦۲	A
۳۵	٦/۲۸	*۰/۰۱	۱/۱۹	-۰/٤۲	A
۳٦	۳/٤۱	۰/۰٦	۰/۸٦	۰/۳٤	A
۳۷	۲/۵۳	۰/۱۱	۰/۸۸	۰/۳۰	A
۳۸	۰/۰٦	۰/۸۱	۱/۰۳	-۰/۰۸	A
۳۹	۰/۰۲	۰/۹۰	۰/۹۸	۰/۰۵	A
٤۰	۰/۰	۰/۹۵	۰/۹۹	۰/۰۲	A
٤۱	۵/۰٦	*۰/۰۲	۰/۸۱	۰/۵۱	A
٤۲	۰/۱۱	۰/۷٤	۱/۰۳	۰/۰۷	A
٤۳	۵/۰	*۰/۰۳	۱/۳۸	-۰/۷۵	A
٤٤	۹/۰٦	*۰/۰۰	۱/۳٦	-۰/۷۳	A
٤۵	٤/۷۹	*۰/۰۳	۰/۸٦	۰/۳٦	A
٤٦	۱/۲۷	۰/۲۳	۱/۱۰	-۰/۲۳	A
٤۷	۲/۰۹	۰/۱۵	۱/۱٤	-۰/۳۰	A
٤۸	٤/٤۲	*۰/۰٤	۰/۸٤	۰/٤۰	A
٤۹	۲۱/٦٤	*۰/۰۰	۰/۷٤	۰/۷۱	A
۵۰	۰/۳۰	۰/۵۸	۱/۰۵	-۰/۱۳	A
۵۱	۰/۰۵	۰/۸۲	۱/۰۳	-۰/۰۸	A
۵۲	۲/٤۷	۰/۱۲	۱/۲۷	-۰/۵۷	A
۵۳	٦/۸٤	*۰/۰۱	۱/۳۲	-۰/٦٦	A
۵٤	۰/۰٤	۰/۸٤	۱/۰۲	-۰/۰٤	A
۵۵	۱۸/۸۸	*۰/۰۰	۱/٤۰	-۰/۷۸	A

همان‌گونه که در جدول مشخص شده است، شاخص مانتل هنزل ۲۳ سئوال (٤۱/۸ درصد سئوال‌ها) در سطح ۰/۰۵ معنادار است. از این ۲۳ سئوال با توجه به مقادیر دلتای مانتل هنزل ۲۲ سئوال در سطح A هستند، به این معنا که مقدار دلتای مانتل هنزل آنها

کم‌تر از یک است و در واقع با وجود معنادار بودن شاخص مانتل هنزل، کارکرد افتراقی ناچیزی دارند و می‌توان این سئوال‌ها را در آزمون حفظ کرد. سئوال شماره ۱۰ در سطح B است و دارای کارکرد افتراقی متوسط است. هیچ یک از سئوال‌های آزمون کارکرد افتراقی زیاد از خود نشان نداده‌اند (سطح c). از این ۲۳ سئوال که شاخص مانتل هنزل معنادار دارند، ۱۰سئوال به نفع داوطلبان دختر (سئوال‌های شماره ۵، ۷، ۹، ۱۸، ۲۷، ۳۲، ٤۱، ٤۵، ٤۸ و ٤۹) و ۱۳ سئوال (سئوال‌های شماره ۸، ۱۰، ۱۳، ۱٤، ۱٦، ۲۸، ۳۳، ۳٤، ۳۵، ٤۲، ٤٤، ۵۳ و۵۵) به نفع داوطلبان پسر است.

همان‌گونه که مشاهده می‌شود روش مبتنی بر نظریه سئوال – پاسخ تعداد سئوال‌های کم‌تری را به عنوان سئوال‌های دارای کارکرد افتراقی مشخص کرده است (۹ سئوال در مقابل ۲۳ سئوال) که این یافته با نتایج مطالعه دوراند و پارک همسو است. یک دلیل تفاوت حساسیت روش‌های مختلف تعیین کارکرد افتراقی ممکن است مربوط به تعدیلات ریاضی خطای نوع اول باشد که در هر برنامه‌ای به گونه متفاوتی اجرا می‌شود. بررسی سئوال‌هایی که دارای کارکرد افتراقی هستند بیانگر آن است که سئوال‌هایی که به نفع دختران کارکرد افتراقی دارند در حوزه محتوایی توابع و معادلات هستند و سئوال‌هایی که به نفع پسران کارکرد افتراقی دارند بیشتر در حوزه محتوایی مثلثات، هندسه و احتمال می‌باشند. البته بررسی دقیق‌تر سئوال‌های آزمون توسط کارشناسان موضوعی به منظور روشن ساختن منبع و دلیل کارکرد افتراقی ضروری می‌باشد.

به منظور بررسی توافق و هم‌خوانی میان دو روش مانتل هنزل و روش مبتنی بر رویکرد سئوال – پاسخ در تشخیص کارکرد افتراقی، ضریب کاپا[1] محاسبه شد و به این ترتیب میزان هم‌خوانی روش‌ها در تعیین سئوال‌های دارای کارکرد افتراقی مشخص شد. در جدول (٦) فراوانی سئوال‌های دارای کارکرد افتراقی و فاقد کارکرد افتراقی براساس دو رویکرد ارائه شده است. همان‌گونه که در جدول مشخص شده است، دو روش در تعیین هشت سئوال دارای کارکرد افتراقی و در ۳۱ سئوال فاقد کارکرد افتراقی مطابقت داشتند. مقادیر ضریب کاپا در مطالعه حاضر برابر با ۰/۳٤٦ (p=۰/۰۰۲) و فاصله اطمینان ۹۵ درصد (۰/۵٦۵، ۰/۱۲٦) است. پیشنهاد لاندیس[2] و کخ[3] (۱۹۹۷) این است که کاپای بیش از ۰/۷۵ حاکی از توافق عالی، کاپای کمتراز

1. Kappa coefficient
2. Landis
3. Kokh

۰/٤ نشاندهنده توافق ضعیف و بالاخره کاپای بین ۰/٤ تا ۰/۷۵ نشان از توافق نسبی خوب دارد. در مطالعه حاضر میزان توافق میان دو روش، ضعیف است.

بیشتر مطالعات انجام شده در خصوص توافق روش‌های مختلف ارزیابی کارکرد افتراقی بیانگر توافق کم میان روش‌های مختلف است. همگرایی روش‌های بررسی کارکرد افتراقی به شدت تحت تأثیر ناپایایی شاخص‌های ارزیابی کارکرد افتراقی است (عبدالعزیز، ۲۰۱۰).

جدول (٦) فراوانی سئوال‌های دارای کارکرد افتراقی و فاقد کارکرد افتراقی

		رویکرد سئوال - پاسخ		تعداد کل
		دارای کارکرد افتراقی	فاقد کارکرد افتراقی	
مانتل هنزل	دارای کارکرد افتراقی	۸	۱۵	۲۳
	فاقد کارکرد افتراقی	۱	۳۱	۳۲
تعداد کل		۹	٤٦	۵۵

نتیجه‌گیری

مطالعه حاضر به ارزیابی کارکرد افتراقی سئوال‌های آزمون ریاضی با استفاده از دو روش مانتل هنزل و روش مبتنی بر نظریه سئوال - پاسخ پرداخته است. تحلیل مبتنی بر رویکرد سئوال - پاسخ با استفاده از نرم افزار BILOG- MG انجام شده است. این برنامه کارکرد افتراقی را صرفاً براساس پارامتر دشواری بررسی می‌کند. نظریه سئوال- پاسخ روش مناسبی برای ارزیابی کارکرد افتراقی سئوال‌ها محسوب می‌شود که این به واسطه ویژگی نامتغیر بودن پارامترها در این رویکرد است. نتایج تحلیل، بیانگر وجود کارکرد افتراقی در سئوال‌های آزمون ریاضی است. براساس رویکرد سئوال- پاسخ، نُه سئوال کارکرد افتراقی دارند که تمامی آنها نیز به نفع آزمودنی‌های دختر هستند. در تحلیل آزمون براساس روش مانتل هنزل، شاخص مانتل هنزل ۲۳ سئوال آزمون در سطح ۰/۰۵ معنادار بود که ده سئوال به نفع آزمودنی‌های دختر و ۱۳سئوال به نفع آزمودنی‌های پسر است. با توجه به طبقه‌بندی خدمات سنجش آموزشی ۲۲ سئوال دارای کارکرد افتراقی ناچیز می‌باشند (سطح A) و یک سئوال کارکرد افتراقی متوسط دارد (سطح B). شایان ذکر است که از نُه سئوالی که بر اساس رویکرد سئوال - پاسخ دارای کارکرد افتراقی هستند، هشت مورد آن نیز در روش مانتل هنزل دارای کارکرد

افتراقی می‌باشند. تنها سئوال شماره ۳٦ است که براساس روش مبتنی بر رویکرد سئوال – پاسخ دارای کارکرد افتراقی است، اما در روش مانتل هنزل فاقد کارکرد افتراقی است. میزان توافق و هم‌خوانی میان دو روش مانتل هنزل و روش مبتنی بر رویکرد سئوال – پاسخ در تشخیص کارکرد افتراقی با توجه به ضریب کاپا ۰/۳٤٦ ضعیف است. رویکرد سئوال – پاسخ در مقایسه با روش مانتل هنزل تعداد کمتری از سئوال‌ها را دارای کارکرد افتراقی تشخیص داده است، این یافته مشابه نتایج مطالعه دوراند و پارک (۲۰۰٦) است. بر این اساس سازندگان آزمون‌ها می‌بایست همواره به تفاوت‌های روش‌های مختلف ارزیابی کارکرد افتراقی سئوال‌ها توجه داشته باشند و سئوال‌هایی را که بر اساس یک روش دارای کارکرد افتراقی هستند با سایر روش‌های معتبر نیز مورد بررسی قرار دهند.

بررسی سئوال‌های دارای کارکرد افتراقی در آزمون ریاضی بیانگر آن است که سئوال‌های دارای کارکرد افتراقی به نفع دختران بیشتر در حوزه محتوایی توابع و معادلات و سئوال‌های دارای کارکرد افتراقی به نفع پسران در حوزه محتوایی مثلثات، هندسه و احتمال هستند. پس از تعیین کارکرد افتراقی، کارشناسان موضوعی می‌بایست محتوای سئوال‌ها را به منظور روش‌ساختن منبع و دلیل کارکرد افتراقی مورد بازبینی کامل قرار دهند. از این رو لازم است در مطالعات بعدی سئوال‌های آزمون که دارای کارکرد افتراقی بودند از لحاظ محتوایی توسط کارشناسان موضوعی مورد بررسی قرار گیرند. وجود سئوال‌های دارای کارکرد افتراقی در آزمون، ضرورت بررسی و توجه بیشتر به فرایند طراحی سئوال‌های آزمون و آموزش جامع طراحان سئوال را به منظور تهیه آزمون‌های عادلانه و منصفانه مطرح می‌سازد.

منابع

- امبرتسون، سوزان ای و رایس، استیون پی (۲۰۰۰). *نظریه‌های جدید روانسنجی برای روانشناسان (IRT)*. ترجمه: دکتر حسن پاشا شریفی، دکتر ولی‌الـله فرزاد، مجتبی حبیبی عسگرآباد و بلال ایزانلو (۱۳۸۸). تهران: انتشارات رشد.
- همبلتون، رونالد ک، سوامینانان، اچ و راجرز، جین. (۱۹۹۱). *مبانی نظریه پرسش-پاسخ*. ترجمه: دکتر محمدرضا فلسفی‌نژاد (۱۳۸۹). تهران: انتشارات دانشگاه علامه طباطبایی.

- Abedalaziz, Nabeel. (2010). A gender - related differential item functioning of mathematics test item. *The International Journal of Educational and Psychological Assessment,* 5,101-116 .
- Ahmadi, Alireza. (2008). *Differential Item Functioning in High-stakes Tests: the Effect of Gender and Field of Study.* Doctoral dissertation, University of Isfahan, Faculty of Foreign Languages, Department of English .
- Baghi, Heibatollah & Ferrara, Steven. (1989*). A comparison of IRT, Delta Plot and Mantel- Haenszel techniques for detecting differential item functioning across subpopulations in the Maryland test of citizenship skills.* Paper presented at the annual meeting of the American Educational Research Association. San Francisco, CA: March 27-31 (Eric database, Report: ED324364) .
- Conoley, C. Adele (2003). *Differential item functioning in the Peabody picture vocabulary test - third edition: partial correlation versus expert judgment.* Doctoral dissertation. Texas A&M University .
- Dorans, Neil J. & Holland,Paul W. (1992). *DIF Detection and Description: Maentel - Haenszel and Standardization.* Paper presented at the Educational Testing Service/AFHRL Conference (Princeton. NJ. October)
- Driana, Elin. (2007). *Gender item functioning on a ninth- grade mathematics proficiency test in Appalachian Ohio.* Doctoral dissertation, Ohio University, Ohio .
- Duncan, Cromwell, Susan. (2006). *improving the prediction of differential item functioning: A comparison of the use of an effect size for logistic regression DIF and Mantel- Haenszel DIF methods.* Doctoral dissertation, Texas A&M University .
- Durand, Jeffrey & Park, Siwo. (2006). A Study of Gender and Academic Major - Based Differential Item Functioning (DIF) In KEPT 2006, Mexico .

- Eng, L. S., & Hoe, L. S. (2005). Detecting Differential Item Functioning (DIF) in Standardized Multiple-Choice Test: An Application of Item Response Theory (IRT). [ONLINE] Available at:http://www.ipbl.edu.my/inter/penyelidikan/seminarpapers/2005/linguitm. pdf.
- Hambleton, R. , & Rodgers, J. (1995). Item bias review. Practical Assessment, Research, and valuation, 4(6). Retrieved November 18, 2006, from *http://PAREonline. net/getvn. asp?v=4&n=6*
- Hambleton ،R. K. (1989). Principles and selected applications of item response theory. In R. L. Linn (Ed.) ،Educational measurement (3rd ed ،.pp. 147-200). New York NY: American Council on Education & Macmillan Publishing .
- Landis, J. R. , Koch, G. G. (1977). The measurement of observer agreement for categorical data. Biometrics 33:159-174.
- Nässtrom, Gunilla. (2003). *Differential item functioning for items in the Swedish National test in mathematics, course B.* Paper presented at the Pre-ICME Conference in Växjö .
- O'Neal, Marcia R. (1991). A Comparison of Method for Detecting Item Bias**.** Paper presented at the annual meeting of the Mid-South Educational Research Association (20th, Lexington, KY, November 12-15 .
- Oshima, T. C. and Morris, S. B. (2008), Raju's Differential Functioning of Items and Tests (DFIT). Educational Measurement: Issues and Practice, 27: 43–50. doi: 10. 1111/j. 1745-3992. 2008. 00127 .
- Rousseau, M. , Bertrand, R. , & Boiteau, N. (2004). *Impact of missing data on robustness of DIF IRT-based and Non-IRT-based methods*. Paper presented at the 2004 AERA annual meeting .
- Shultz, S. Kenneth & Whitney, David. (2005). Measurement Theory in Action. Case Studies and Exercises. Sage publication .
- Willingham, W. W. & Cole, N. S; (1997). Gender and fair asssessment. New Jersey, U. S. A: Lawrence Erbaum associate
- 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-like (ordinal) item scores. Ottawa, Canada: Directorate of Human Resources Research and Evaluation .
- Zwick ,Rebecca (2012). A Review of ETS Differential Item Functioning Assessment Procedures: Flagging Rules, Minimum Sample Size Requirements, and Criterion Refinement. [ONLINE] Available at: http://www. ets. org/Media/Research/pdf/RR-12-08. pdf.